Journal of Economic Psychology 75 (2019) 102143

Contents lists available at ScienceDirect

Journal of Economic Psychology

journal homepage: www.elsevier.com/locate/joep

Preregistration and reproducibility☆

Eirik Strømland

Department of Economics, University of Bergen, Hermann Fossgate 6, 5007 Bergen, Norway

A B S T R A C T

Many view preregistration as a promising way to improve research credibility. However, scholars have argued that using pre-analysis plans in Experimental Economics has limited benefits. This paper argues that preregistration of studies is likely to improve research credibility. I show that in a setting with selective reporting and low statistical power, effect sizes are highly inflated, and this translates into low reproducibility. Preregistering the original studies could avoid such inflation of effect sizes—through increasing the share of "frequentist" researchers—and would lead to more credible power analyses for replication studies. Numerical applications of the model indicate that the inflation bias could be very large in practice, and available empirical evidence is in line with the central assumptions of the model.

1. Introduction

Frequentist statistics rely on the assumption that the analysis of an experiment is independent of how the results turn out. If we regard the estimates as a random draw from some fixed population model, the parameter estimates and p-values will, on average, be correct (Neyman, 1937). In contrast, if the data analysis is contingent on the data, the p-values and parameter estimates are hard to interpret (Gelman & Loken, 2014; Simmons, Nelson, & Simonsohn, 2011). A way to ensure that the frequentist assumption holds is to tie our hands in advance — preregister the choices to be made before the data have been seen.

Researchers have claimed that preregistering studies is a major step toward greater research credibility (Munafò et al., 2017; Nosek, Ebersole, DeHaven, & Mellor, 2018). However, the literature offers little justification for the use of preregistration in Experimental Economics. Based on a paper by Brodeur, Lé, Sangnier, and Zylberberg (2016), several prominent researchers have claimed that Experimental Economics seems to suffer less from p-hacking – selective reporting of statistically significant results – than other applied fields in Economics (Camerer et al., 2016; Coffman & Niederle, 2015; Maniadis, Tufano, & List, 2017). If researchers do not selectively report statistically significant findings, there is no reason for advocating for policies that will force them to tie their hands in advance of the data analysis. Moreover, Coffman and Niederle (2015) make a theoretical argument that preregistration of studies offers limited benefits, except for conducting replication studies. They argue that preregistration would have a small effect on the conditional probability that published significant findings are true, and that scientific beliefs typically rapidly converge to the truth even in absence of preregistration.

This paper offers a theoretical justification for preregistration in Experimental Economics and argues that available empirical evidence supports this theoretical account. I present a simple model in which the research community is populated by agents who may report their results unconditionally or conditionally on statistical significance and show that greater reliance on preregistration improves the estimation of the effect sizes through increasing the share of "frequentist" researchers. As replicators are likely to estimate statistical power based on the published effect sizes, preregistration is therefore also expected to improve reproducibility rates. Numerical illustrations show that the bias from effect size inflation may be very large in practice.

The model assumes that researchers access many possible tests of the same hypothesis, that some fraction in the population selectively report their main findings and that the statistical power is low on average. Empirical evidence seems to support these assumptions. First, a

☆ I am grateful to Anna Dreber Almenberg, Rune Jansen Hagen, Amanda Kvarven, Bjørn Sandvik, Nina Serdarevic, Erik Ø. Sørensen, Sigve Tjøtta, Fabio Tufano and two anonymous referees for helpful comments and discussions.

E-mail address: Eirik.Stromland@uib.no.

https://doi.org/10.1016/j.joep.2019.01.006
Received 24 May 2018; Received in revised form 24 January 2019; Accepted 25 January 2019
Available online 06 February 2019

re-analysis of Brodeur et al. (2016) suggests similar patterns of p-hacking in Experimental Economics as in other applied fields. Second, recent attempts to quantify statistical power in Economics (Ioannidis, Stanley, & Doucouliagos, 2017; Zhang & Ortmann, 2013) suggest that studies tend to be underpowered, and attempts to estimate the "inflation bias," the extent to which published effects are overestimated, show that the published effects are inflated by a factor of two or more (Ioannidis et al., 2017). The latter estimate is very close to the "worst-case" theoretical magnitude of the inflation bias suggested in the numerical applications of the model.

This paper contributes to a small but growing number of studies in the intersection between "meta-research" – the field of research that studies the institutions and practices of scientific research (Ioannidis, 2018) – and economic theory (Gall, Ioannidis, & Maniadis, 2017; Gall & Maniadis, 2019; Maniadis et al., 2017). Specifically, this paper is an application of Ioannidis's (2008) and Gelman and Carlin's (2014) ideas to Economics. These papers showed that even if the null hypothesis is false, selective reporting and low statistical power together lead to highly exaggerated published effect sizes that, in turn, lower reproducibility. Specifically, the contribution of this paper is to put these ideas into a simple economic model and show that a move toward preregistration within the context of the model may improve reproducibility through a shift to an equilibrium with a higher share of "frequentist" researchers.

Although there are several applications based on Ioannidis (2005) in Economics (Coffman & Niederle, 2015; Maniadis et al., 2014, 2017), there are, to my knowledge, no applications of the framework presented in Ioannidis (2008). This distinction is important in practice. Coffman and Niederle (2015) follow Ioannidis (2005) model setup, where hypotheses are modeled as either "true" or "false," and conclude that preregistering studies offers little benefit to experimental economists because it has little effect on the probability that a significant finding is true, and beliefs tend to converge to the truth even in absence of preregistration of studies. In contrast, I adopt the focus on effect sizes taken by Ioannidis (2008) and reach the opposite conclusion. A move toward more preregistration will, in general, improve the credibility and reproducibility of published results through reducing effect size inflation and improve power estimates based on published effect sizes. Importantly, Bayesian updating in this setting rapidly leads to convergence on a true belief that the null hypothesis is false both when effects are inflated and unbiased — showing that correct belief formation is not necessarily an argument against preregistration of studies. Thus, this paper is in line with the recommendation not to always treat research evidence as "true" or "false"(McShane & Gal, 2017): A finding may be "true" but still misleading.

This paper proceeds as follows. Section 2 presents a simple economic model of the reporting process of main experimental findings and shows that given this model setup, greater reliance on preregistration would improve the quality of our estimates of treatment effects and improve reproducibility through more accurate power estimates. Using numerical illustrations for a specific assumed probability distribution, I show that the bias in the absence of preregistration may be very large in practice and illustrate how belief formation is affected in different scenarios. In Section 3 I discuss how the key assumptions of the model fit with available empirical evidence. Finally, Section 4 concludes.

2. Theoretical framework

2.1. Inflation of effect sizes

Suppose researchers want to infer whether an effect exists ($\theta \neq 0$) or not ($\theta = 0$) and aim to estimate the size of the true effect. Each researcher accesses multiple possible choices of test results of the point null hypothesis that $\theta = 0$ and parameter estimates from some set Θ and decides which of these possible choices to report as her main finding. The set Θ can be considered a set of realized draws from some underlying "latent" joint distribution of estimates and test results. We denote by $f(t, \hat{\theta}|\theta^*)$ the latent joint distribution of the test results and estimates with correlation ρ, conditional on the true value θ^*, which is normalized to be positive. The estimate $\hat{\theta}$ may be thought of as measuring some treatment effect of interest in a randomized experiment, which ensures that the unconditional expectation of $\hat{\theta}$ is equal to the true value θ^*. Denote by V the number of available tests for which $|t_k| > \underline{t}$ (the test is statistically significant) and denote by $Pr(|t_k| < \underline{t})$ the type II error probability for test k. For K independent tests, the probability of at least one correct rejection is

$$\Pr(V \geq 1|\ \theta^*) = 1 - \prod_{k=1}^{K} \Pr(|t_k| < \underline{t}|\ \theta^*), \tag{1}$$

where $\prod_{k=1}^{K} \Pr(|t_k| < \underline{t})$ is the joint probability of only making type II errors, which converges to zero as K→ ∞.[1] We will assume K is sufficiently large so that $\Pr(V \geq 1|\ \theta^*) = 1$. Researchers access at least one possible test that rejects the null hypothesis.[2] One interpretation of this assumption is that we simply assume the underlying hypothesis being tested is somewhat "flexible", and there are many ways of rejecting the null hypothesis if so desired. This interpretation captures what Simmons et al. (2011) refer to as "researcher degrees of freedom" and Gelman and Loken (2014) call the "garden of forking paths": Researchers typically have at their disposal at least one possible test that would corroborate their underlying theory. Another way to make sure to access enough choices

[1] Suppose that $\{\Pr(|t_k| < \underline{t})|\theta^*\}_k < a < 1\ \forall\ k$. Then $lim_{K\to\infty}(a^K) = 0$, which implies that $lim_{k\to\infty}[\prod_{k=1}^{K} \Pr(|t_k| < \underline{t}|\ \theta^*)] = 0$. This result does not rely on the independence assumption and could be extended to conditionally dependent tests. For independent tests, convergence is fast; for only 20 independent tests and assumed power at 20%, Eq. (1) yields a 98.8% chance of at least one statistically significant test. For dependent tests, the convergence depends on how correlated the tests are.

[2] In our model setup, the choice set Θ generally contains different choices for each researcher as each possible choice is a random draw from some joint probability distribution. However, this does not matter in practice as long as we assume that each researcher has a sufficient number of choices so that at least one statistically significant result is available.

is to conduct multiple experiments testing the same hypothesis.

I assume that the researcher reports a single result $(\hat{\theta}_k, t_k) \in \Theta$. As a simplification, I consider this choice as a choice between two possible reporting strategies: The researcher can report her parameter estimate either unconditionally and choose a random result or conditionally on statistical significance and report an estimate for which $t_k > \underline{t}$. The expected value of the effect across researchers who report unconditionally is then $E(\hat{\theta}_k) = \theta^*$ whereas the expected value of the effect across researchers who report conditionally on the test passing the threshold of statistical significance is written as θ^T: $=E(\hat{\theta}_k | t_k > \underline{t})$. The assumptions behind the selective reporting strategy conform to standard models of publication selection that assume selective reporting follows a one-sided form of "incidental truncation" where the reported effect is selected subject to the value of a test statistic passing some threshold in a desirable direction (Stanley & Doucouliagos, 2013).

I assume that the expected utility gain from choosing a strategy of selective reporting, relative to choosing an unconditional reporting rule, is $u = b - \omega$. Here, $b \in [\underline{b}, \bar{b}] \subset \mathbb{R}^+$ is a fixed material benefit parameter, and $\omega \in [\underline{\omega}, \bar{\omega}] \subset \mathbb{R}^+$ is the subjective cost of the selective reporting strategy, with corresponding cumulative distribution function $G(\omega)$. The term b may be regarded as a fixed incentive parameter that captures the fact that behaving according to a selective reporting rule may benefit researchers in material terms through the increased chance of a good publication, while ω is the subjective cost component that captures that some researchers may experience a cost from deviating from an unconditional reporting rule.[3] While b in reality will be a function of a vector of policy variables, b: $\boldsymbol{D} \rightarrow [\underline{b}, \bar{b}]$, we will for notational convenience treat it as a parameter that is set exogenously by an agent outside of the model. Thus a higher b will in reality reflect a chosen policy vector $\boldsymbol{D}_1$ such that $b(\boldsymbol{D}_1) > b(\boldsymbol{D}_0)$ where $\boldsymbol{D}_0$ was the initial policy vector before the policy change.

This utility function is adopted from the literature on cheating behavior (Andvig & Moene, 1990; Brocas & Carillo, 2018; Tirole, 1996), which, in our context, would be understood as "p-hacking" (Simmons et al., 2011). However, it is also possible to make data-contingent choices without consciously looking for statistically significant results. The researcher may simply intend to describe the data in the best possible way and ends up conditioning on statistical significance (Gelman & Loken, 2014). For this reason, I choose the neutral term "frequentist" for a researcher who reports unconditionally and I choose not to put a particular label on researchers who report conditionally on statistical significance.

In the setup above, researchers may choose a strategy of selective reporting and gain $b - \omega$ or choose to report their main result unconditionally on the observed data and gain zero. Thus, researchers will choose to be "frequentist" if $\omega > b$. We denote by $\dot{\omega}$ the subjective cost parameter for the marginal agent who is indifferent between the two strategies $(b = \dot{\omega})$ so that in equilibrium the share of frequentist researchers is determined by the complementary cumulative distribution function $\bar{G}(\dot{\omega}) = \Pr(\omega > \dot{\omega})$. A fraction $\bar{G}(\dot{\omega})$ chooses to report unconditionally, and a fraction $1 - \bar{G}(\dot{\omega})$ chooses a strategy of selective reporting. Note that an increase in b—increasing the incentives to choose a strategy of selective reporting —lowers the share of agents choosing to engage in "frequentist" reporting of results $(\frac{\partial \bar{G}(\dot{\omega})}{\partial b} < 0)$.

The mean effect reported in the literature is a weighted sum of the expected estimates reported by the two types of researchers. We can express the mean effect as follows, where $\bar{G}(\dot{\omega})$ is the share of "frequentist" researchers:

$$\bar{\theta} = \bar{G}(\dot{\omega})\theta^* + [1 - \bar{G}(\dot{\omega})]\theta^T.$$

Given the simple model outlined here, we may immediately establish an important result.

Proposition 1. Suppose $\rho > 0$. *Then the inflation bias, $\delta \equiv \bar{\theta} - \theta^*$, is decreasing in $\bar{G}(\dot{\omega})$, the share of frequentist researchers.*

Proof. *Write the inflation bias as*

$$\delta = \bar{\theta} - \theta^* = [1 - \bar{G}(\dot{\omega})](\theta^T - \theta^*)$$

Now consider two arbitrary distributions of the frequentist types in the population, $\bar{G}_0(\dot{\omega})$ and $\bar{G}_1(\dot{\omega})$, with $\bar{G}_1(\dot{\omega}) > \bar{G}_0(\dot{\omega})$. The difference in inflation bias is given by

$$\delta_1 - \delta_0 = [1 - \bar{G}_1(\dot{\omega})](\theta^T - \theta^*) - [1 - \bar{G}_0(\dot{\omega})](\theta^T - \theta^*) = [\bar{G}_0(\dot{\omega}) - \bar{G}_1(\dot{\omega})](\theta^T - \theta^*)$$

As θ^T is an incidentally truncated mean and the correlation between t and $\hat{\theta}$ is $\rho > 0$, truncation from below pushes the sampling distribution to the right; thus, $\theta^T > \theta^*$. As $\bar{G}_0(\dot{\omega}) - \bar{G}_1(\dot{\omega}) < 0$ by assumption, the difference $(\delta_1 - \delta_0)$ is strictly negative. Thus, the inflation rate δ is strictly lower for $\bar{G}_1(\dot{\omega}) > \bar{G}_0(\dot{\omega})$.

Preregistration policies could be seen as inducing an exogenous decrease in b, the expected benefit of selective reporting, and therefore a corresponding increase in $\bar{G}(\dot{\omega})$, improving parameter estimation. One example of such a preregistration policy is for a journal to adopt the format of a "registered replication report" where the journal commits to publishing based on a pre-analysis plan and reviewers make sure that this plan is followed (Simons, Holcombe, & Spellman, 2014).[4] Such a policy would reduce b through lowering the set of journals that publish results based on a strategy of selective reporting, and would increase $\bar{G}(\dot{\omega})$ as the critical $\dot{\omega}$ would now be lower for an agent to prefer unconditional reporting over a strategy of selective reporting.

[3] The model abstracts from strategic interaction among agents involved in the publication process, which would be a more realistic way of modeling the reporting process but is beyond the scope of the present paper.

[4] An alternative model setup would be to assume that ω is endogenous; for instance, the aversion to selective reporting may be driven by social norms in the profession. In that case, one could obtain a move toward more preregistration through means other than changing the material incentives associated with selective reporting.

As the inflation bias is zero only for $\bar{G}(\breve{\omega}) = 1$, it also follows directly from the above that replicators who estimate the statistical power based on previously reported effect sizes will, on average, condition on a false assumption that the underlying effect is larger than the true effect θ^*. Therefore, replicators will in general tend to overestimate the statistical power of the replication study.[5]

Proposition 1 ensures that preregistration policies that reduce the incentives towards selective reporting lead to an improvement in estimation of effect sizes. However, this is not in itself informative about the size of the inflation bias. To illustrate how big the inflation bias may be in practice, I impose some specific distributional assumptions on the latent density $f(\hat{\theta}, t|\theta^*)$ and show the inflation bias numerically for different assumptions in the share of frequentist researchers in the population. I assume that each estimate is normally distributed; thus, $\hat{\theta}_k \sim N(\theta^*, \frac{2\sigma^2}{N})$ where σ is a known standard deviation common to an experimental treatment and a control group, and θ^* is the true treatment effect in the study population. I assume a sample size of N = 100 in each treatment group in some initial study; thus, the total sample size is 200, which gives a statistical power of 20.5% for a true effect size of 8 percentage points (θ^*=0.08) and a common standard deviation of $\sigma = 0.5$ assumed known to all researchers. In this setup, the formula for an incidentally truncated random variable yields $\theta^T = \theta^* + \sqrt{\frac{2\sigma^2}{N}}\left[\frac{\phi(\alpha)}{1-\Phi(\alpha)}\right]$ where $\phi(.)$ and $\Phi(.)$ are the probability density function (PDF) and the cumulative distribution function (CDF) of the standard normal distribution, and α is the value of the Z-statistic at which truncation occurs (Greene, 2012, p. 913). Inserting a truncation point of $\alpha = [1.96 - 0.08/\mathrm{SE}(\hat{\theta})]$ for the Z-statistic that would make the estimate statistically significant I obtain $\theta^T = 0.1783$.[6]

I also display the results from hypothetical power calculations carried out by replicators who draw upon previously published results when determining their sample size. As the standard deviation is assumed common to the control group and the treatment group and known, I assume that the replicators estimate their necessary total sample size based on a z-test using the formula $N^* = \left(\frac{2.8}{\hat{\theta}}\right)^2$ (e.g. Gelman & Hill, 2007, p. 443). The true required sample size is, however, calculated by inserting the true effect θ^* in the above formula. Table 1 displays the mean reported effect for different shares of frequentist researchers, the sample size estimated to yield 80% power of a replication and the true power given this estimated sample size.

Table 1 shows that in a low-powered setting, the inflation bias is generally sizeable unless the share of frequentist researchers is very high. In the worst-case scenario, the reported effect will be, on average, 2.2 times as high as the true effect, and researchers who estimate the power of a replication study based on this effect will set a sample size that in reality gives them only 24% power. Thus, only 24% of replications would actually succeed in obtaining a statistically significant effect in the direction of the original study – one of the key measures of reproducibility used in recent replication studies (Camerer et al., 2016, 2018). Even when the share of frequentist reporting is high ($\bar{G}(\breve{\omega}) = 0.6$), the mean effect is still 1.5 times higher than the true effect, and the power is estimated to 80% with a sample size of 552 subjects while the true power is just 46.8%. Thus, in this low-powered setting a move toward full formal preregistration would lead to major improvement in the expected reproducibility rate.

The results shown in Table 1 depend on the assumptions made about statistical power. If the statistical power in the original study is high (80%), then the inflation bias will generally be very small even in the worst-case scenario with only selective reporting. The reason is that for a constant θ^* and conventional test statistics (e.g. the *t*-test), an increase in sample size leads to higher values of the test statistic for lower values of the estimate. The truncated sampling distribution would then approach the "frequentist" sampling distribution as $N \to \infty$, and even replicators who base their power estimates on selectively reported estimates would obtain quite accurate power estimates. Table A.1 in Appendix A illustrates how the conclusions in Table 1 change when we instead consider a scenario with statistical power equal to 80%; in this high-powered setting, there is little to gain from promoting preregistration policies as the scope for improvement of effect size estimation is small. However, it can be shown that as long as the statistical power is below about 50%, the sampling distribution of only statistically significant effects will be different enough from the unconditional sampling distribution that the difference between the two distributions will be large in practice (Gelman & Carlin, 2014).

2.2. Belief formation

A possible argument against preregistration is that Bayesian updating rapidly leads to a true belief that $\theta \neq 0$ even in its absence (Coffman & Niederle, 2015). However, beliefs will gradually converge to a true belief on the hypothesis that $\theta \neq 0$ regardless of whether effect sizes are unbiased or inflated. Thus, it does not follow from the fact that beliefs adjust quickly to the truth that replications can substitute for preregistration. To see this, I outline a simple extension of the model of belief formation considered in Coffman and Niederle (2015). The model is intended to describe a setting where no researchers preregister their studies, but they engage in conceptual replications of an initial study, maintaining their researcher degrees of freedom. Other researchers observe the results from each of these studies and update their beliefs using Bayesian updating. The model may not be accurate as a descriptive model of belief formation, but serves as a useful benchmark for showing that correct belief formation on a binary hypothesis may

[5] This assumes that replicators estimate power based on the assumption that the underlying latent distribution that the data are drawn from belongs to a family of probability distributions satisfying the monotone likelihood ratio property, in which case the statistical power function will be increasing in the effect size. This will be the case for common parametric tests such as the *t*-test or the chi-squared test, but may not hold if replicators plan to use a non-parametric test and specifically try to estimate the power of that test.

[6] The effect size needs to be 1.96 standard errors away from zero to be statistically significant. As the estimate is assumed to be a mean treatment effect and the standard deviation is assumed to be known and common to both treatment groups, the standard error of the estimated effect is equal to $\sqrt{\frac{2\sigma^2}{N}}$, where *N* is the sample size in each group. In our context, the effect size needs to be at least 0.1386 to be statistically significant. Using the cumulative normal distribution, the probability of an estimate higher than 0.1386 is 20.5% for a true effect size $\theta^* = 0.08$.

Table 1
Numerical application of the model assuming a normal distribution for the estimates and 20.5% power in the original study to detect the true effect.

$\bar{G}(\breve{\omega})$	θ^*	$\bar{\theta}$	Estimated total N to achieve 80% power	Actual power
0	0.08	0.1783	248	24.3%
0.2	0.08	0.1586	312	29.3%
0.4	0.08	0.13898	406	36.4%
0.6	0.08	0.11932	552	46.8%
0.8	0.08	0.09966	790	61.4%
1	0.08	0.08	1226	80%

Note: $\bar{G}(\breve{\omega})$ is the share of frequentist researchers, θ^* is the true effect, and $\bar{\theta}$ is the mean reported effect given the share of frequentist researchers.

coincide with highly inflated published effects.

Denote agents' beliefs about the likelihood that $\theta \neq 0$, conditional on the data observed in period p as $\beta_p(\theta \neq 0|t_k; \hat{\theta}_k)$. We assume that belief formation depends only on observed test results and that agents do not distinguish between different tests that are statistically significant; thus, $\beta_p(\theta \neq 0|t_k; \hat{\theta}_k) = \beta_p(\theta \neq 0|t_k)$ and $\beta_p(\theta \neq 0||t_k| \geq \underline{t}) = \beta_p(\theta \neq 0||t_l| \geq \underline{t})\ \forall\ t_k, t_l$. This again captures that the underlying theory or hypothesis is "flexible" in the sense that many different tests may serve as corroborating evidence for the association under testing. From Bayes' rule, the belief updating rule may then be written as

$$\beta_p(\theta \neq 0||t_k| \geq \underline{t}) = \frac{Pr(|t_k| \geq \underline{t}|\ \theta \neq 0) \times \beta_{p-1}(\theta \neq 0||t_k| \geq \underline{t})}{Pr(|t_k| \geq \underline{t})}. \quad (2)$$

The prior belief (when p = 0) is given by β_0. In addition, the probability of the reported test passing the statistical significance threshold (the "perceived power") can be expressed as $Pr(|t_k| \geq \underline{t}|\ \theta^*) = \bar{G}(\breve{\omega}) Pr(|t_i| \geq \underline{t}|\ \theta^*) + [1 - \bar{G}(\breve{\omega})] \Pr(V \geq 1|\ \theta^*)$ where t_k is the reported test result, and t_i is a random test. $Pr(|t_k| \geq \underline{t})$ is the unconditional probability of a statistically significant effect and is equal to $\alpha[1 - \beta_{p-1}(\theta \neq 0||t_k| \geq \underline{t})] + Pr(|t_k| \geq \underline{t}|\ \theta^*) \beta_{p-1}(\theta \neq 0||t_k| \geq \underline{t})$. Note that we assume that agents do not hold correct beliefs about power. In reality, individual tests are underpowered (we assume that power is equal to 20%), but agents assume that power in each period is equal to the probability that the reported finding is statistically significant, conditional on θ^*. This expression is equal to the actual power only when $\bar{G}(\breve{\omega}) = 1$ (all researchers are frequentist).

Fig. 1 illustrates the belief updating process for a range of priors β_0 in the interval [0, 1] and different assumptions about the share of frequentist researchers. The results are displayed in Fig. 1.

Fig. 1 shows that for the scenario with full selective reporting of statistically significant findings ($\bar{G}(\breve{\omega}) = 0$), beliefs converge rapidly to the true belief that $\theta \neq 0$. Only three significant tests for the hypothesis are necessary for a posterior belief of 1 that the association under testing is true. Convergence to this belief is somewhat slower with full frequentist reporting ($\bar{G}(\breve{\omega}) = 1$) as the perceived power in this scenario will be equal to true power. Thus, one would expect belief formation to be slower in a scenario where all researchers preregistered their experiments.

This result shows that beliefs converge quickly on the truth even in a setting where the findings are severely inflated by selective reporting. Actually, beliefs converge faster the more severe selective reporting there is. Thus, the rapid convergence of beliefs does not tell us anything about the merits of preregistration: Beliefs will converge on the belief that $\theta \neq 0$ in the presence and absence of effect inflation, but the reproducibility rates in these two scenarios will be very different.[7]

3. Discussion of assumptions

The model and numerical illustrations make three crucial assumptions for effect size inflation to be large. First, it is assumed that there are sufficient researcher degrees of freedom such that researchers are always able to report their main estimate conditional on it being statistically significant. Second, it is assumed that individual tests are underpowered to detect the underlying effect θ^*. Finally, it is assumed that there is a sizeable fraction of researchers who actually engage in selective reporting of their main findings. These researchers preferentially report results that pass the threshold associated with statistical significance.

The assumption that researchers access enough possible tests so that at least one will be statistically significant may seem strong, but even in a simple economic experiment there is a large number of ways to test the null hypothesis of interest. For only 20 binary choices, there are more than a million (2^{20}) ways to analyze the same data (Ioannidis, 2018). The typical economic experiment involves many subjective choices: which treatment to focus on, whether to perform a parametric or non-parametric test, and if so, which one (e.g., linear or non-linear regression model, Mann-Whitney test, or robust rank-order test). The researcher also chooses what outcome variable(s) to focus on and whether to focus on the main sample or a subgroup (and if so, which subgroup), which control variables to add in a regression, and what to do with the standard errors. There may exist a reasonable justification for each of the possible choices, so that researchers do not need to consciously "manipulate" the data to end up reporting their main estimate conditional on statistical significance (Gelman & Loken, 2014).

[7] In Fig. 1, it is assumed that the power is 20% for individual tests. In a scenario where true power was 80%, convergence to the belief that $\theta \neq 0$ would be very fast even in the scenario with full preregistration.

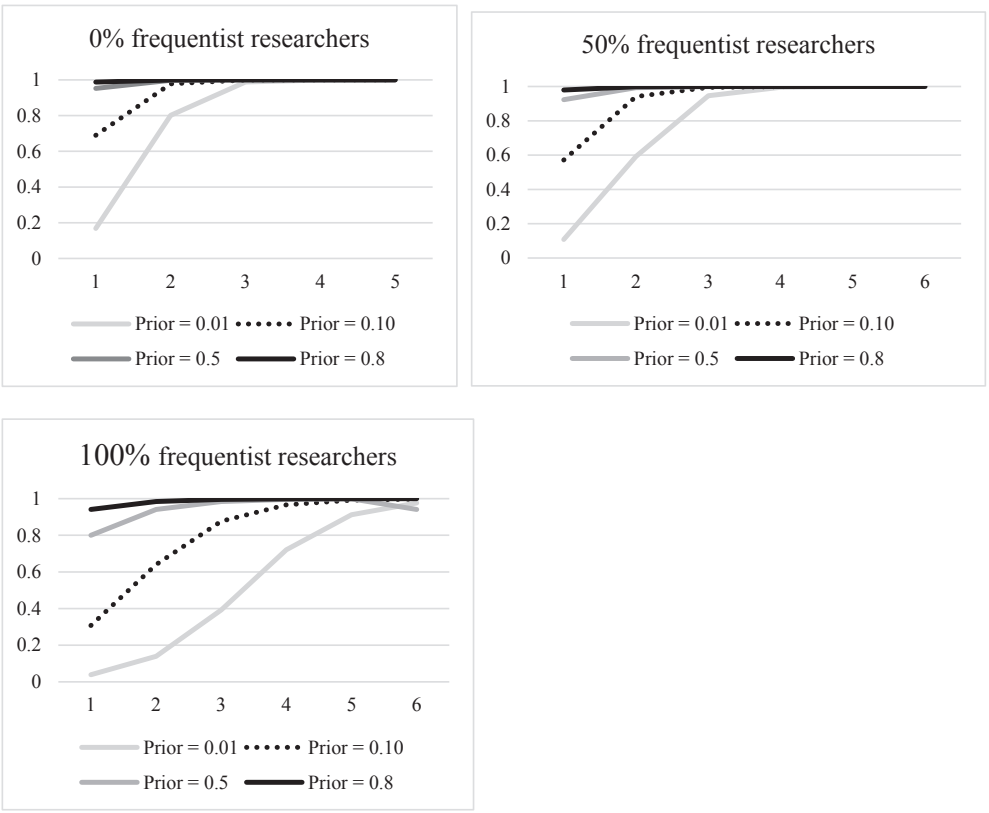

Fig. 1. Belief updating for different possible shares of frequentist researchers. These illustrations are based on Eq. (2) and the assumption that statistical power of the individual tests is equal to 20%. The y-axis measures the posterior belief attached to the hypothesis that $\theta \neq 0$ in period p, $\beta_p(\theta \neq 0 | t_k | \geq \underline{t})$.

For the assumption of statistical power, a recent study by Ioannidis et al. (2017) uses meta-analytic methods to estimate the median power in economic research. They find that the median power in Economics is only 18% and argue that this constitutes an empirical upper bound on true median power as publication bias will lead to overestimation of true effects and therefore, overestimation of power. Comparing high-powered and low-powered studies, Ioannidis et al. also suggest that estimates in Economics tend to be overestimated by a factor of two or more, which closely corresponds to the degree of overestimation in Table 1 in this paper for the scenario with full selective reporting. In addition, an unpublished study by Zhang and Ortmann (2013) estimates median power only for Dictator games and finds the median power is less than 25%. These power estimates are very close to the assumption made in the numerical applications of the model illustrated in Fig. 2.[8]

The assumption that researchers in Experimental Economics preferentially report statistically significant results is potentially more problematic. Several papers (Camerer et al., 2016; Coffman & Niederle, 2015; Maniadis et al., 2017) cite Brodeur et al. (2016)—a study of 50,000 hypothesis tests published in the *American Economic Review*, the *Journal of Political Economy*, and the *Quarterly Journal of Econo*mics in the time period 2005–2011—as evidence that Experimental Economics does not suffer from "p-hacking" because there is little evidence for jumps in the distribution of p-values for randomized trials around the cutoff associated with statistical significance. Thus, in terms of the model considered in this paper, we may already be in an equilibrium with close to all researchers being "frequentist" even in the absence of formal preregistration. However, the lack of p-hacking for all randomized trials does not in itself constitute evidence against the hypothesis of p-hacking in Experimental Economics as the Brodeur et al. (2016) analysis does not distinguish between lab experiments and field experiments. Therefore, I re-analyze the data splitting the Brodeur et al. data into lab experiments and field experiments. Fig. 2 displays the results.

The formal accounting method proposed by Brodeur et al. (2016) cannot be applied separately to the experimental data. Thus, we cannot formally test for p-hacking for these data. However, visual inspection suggests that p-hacking is prevalent in lab experiments, whereas this "eyeballing" test suggests that field experiments drive the overall smoothness of the experimental data reported in the final published version.[9] This evidence points to the need for systematic data collection of additional experimental papers in

[8] Similar estimates are reported in other fields, e.g. Neuroscience (Button et al., 2013).

[9] A previous working paper version of the paper (Brodeur, Lé, Sangnier, & Zylberberg, 2013) was also divided into lab experiments and RCTs, but this division seems to have been missed in the subsequent discussion of the paper, perhaps because this split is not included in the final version (Brodeur et al., 2016).

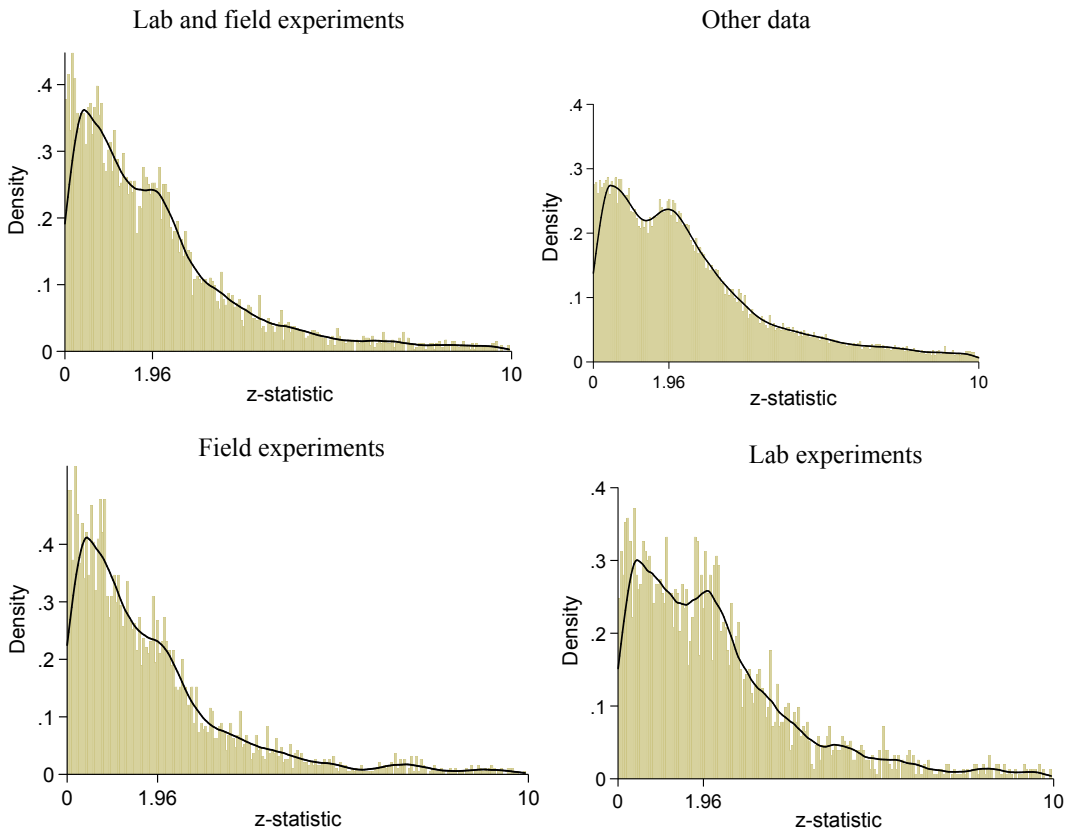

Fig. 2. Distribution of (de-rounded) z-statistics reproduced from Brodeur et al. (2016), by data type. The lines are kernel density estimates.

Economics to test more formally for excess statistical significance. However, we should at least be skeptical of claims that Experimental Economics suffers less from selective reporting than other fields in Economics.

4. Concluding remarks

Scholars have argued in the literature that preregistration is likely to have limited value in Experimental Economics (Coffman & Niederle, 2015). However, this paper has shown that under a plausible set of assumptions about the reporting process in experimental research, correct beliefs that the association is true may coincide with low reproducibility due to inflated effect sizes in the published literature. Preregistration policies, e.g. allowing for a publication format where the journal commits to publishing a paper based on an ex-ante analysis plan (e.g. Simons et al., 2014), will reduce the incentives to selectively report significant findings through restricting the set of possible target journals for such a reporting strategy. The reduction in incentives will lead to a larger fraction of "frequentist" researchers and improved effect sizes, leading replication researchers to more accurately estimate statistical power.

One interpretation of the model and available empirical evidence is that experimental economists should aim to put in place incentives for formally preregistering studies. However, the numerical illustrations of the model also show that in a research environment with only high-powered studies, preregistration would have a small influence on the inflation bias as even findings reported conditional on statistical significance would be close to the true effect. Thus, an alternative solution to recommending preregistration is for experimental researchers to aim to increase statistical power. A problem with this alternative solution is that we cannot test whether such an attempt will succeed. As power depends on the true effect, which is always unknown, there will always be a risk of overoptimistic power calculations. However, both policies that aim to increase statistical power and policies promoting preregistration are expected to have a positive effect on the reproducibility of experimental research.

A clear drawback of the simple model considered in the present paper is that it only allows for two possible choices – unconditional reporting, or reporting results conditionally on obtaining statistical significance. In reality, it is possible that a preregistration policy that reduces the incentives to engage in selective reporting will just lead agents towards other and more serious forms of undesirable behavior, such as fabricating data. A more sophisticated, game-theoretic model of the publication process by Gall and Maniadis (2019) studies these kinds of possible strategic spillovers between choices following a transparency policy in the context of false positive findings. They show that across several model specifications, provided the psychological cost associated with severe misconduct is sufficiently high, policies that aim to reduce mild misconduct will lead to a reduction in overall misconduct. Moreover, removing "mild misconduct" from agents' choice sets tends to shift agents' behavior towards no misconduct rather than

severe misconduct. Extending the model of effect size inflation considered in the present paper to a richer strategic setting such as the one explored in Gall and Maniadis (2019) is a promising venue for future work.

Another drawback of the model is that it does not explicitly distinguish between preregistration policies per se and policies that require authors to be more transparent in their reporting of the analyses they have done along with the published paper (Simmons et al., 2011). Like a preregistration policy, such a "transparency policy" could also be viewed as reducing the expected benefits of selective reporting and thereby increase the share of frequentist researchers. In the model, these two policies could be interpreted as two possible means of obtaining the same end. A preregistration policy could be seen as an "ex ante" transparency policy, aiming to ensure that the analysis plan is actually adhered to ex post. Future work – theoretical and empirical – should aim to discuss the relative costs and benefits of ex ante versus ex post transparency policies in improving effect size estimation.

Finally, it could be argued that the numerical applications of the model in Table 1 implicitly assume that replicators are somewhat "naïve," in the sense that they systematically base their power estimates on previously published effect sizes without taking into account the possibility for inflated effects. This is apparently at odds with evidence from expert surveys and prediction markets showing that researchers are surprisingly accurate in judging which results will replicate or not (Camerer et al., 2016, 2018; Dreber et al., 2015; Forsell et al., 2018). However, systematic overestimation of power may be entirely consistent with perfectly informed replicators who behave rationally according to incentives. To see this, suppose that to get a replication study published a replication team needs to demonstrate that the replication study has at least 80% power to detect the original effect and that beyond this point, increasing the sample size has no effect on the probability of getting the replication study published. Then the optimal sample size rule for replicators will be the minimal sample size that is sufficient to achieve 80% power to detect the original effect, as further increases in power are likely to be very costly.

Although empirical evidence is in line with the suggestion that preregistration potentially offers a large reduction in the inflation bias, there is not enough direct evidence to draw strong conclusions. This suggests two important implications for future empirical research. First, researchers should aim to inquire further into the extent of selective reporting in Experimental Economics, perhaps by collecting more data from strong field journals that publish lab experiments and testing for excess significance in these journals. Second, researchers should aim to estimate statistical power specifically for Experimental Economics, as existing estimates may not be representative of power for laboratory experiments and the three journals reported in Brodeur et al. (2016) constitute only a small fraction of all lab experiments. More evidence along these lines would help us make a better informed decision about whether to recommend preregistration of all studies.

Appendix A

See Table A1.

Table A1
Numerical application of the model assuming a normal distribution for the estimates and 80% power in the original study to detect the true effect.

$\bar{G}(\dot{\omega})$	θ^*	$\bar{\theta}$	Estimated total N to achieve 80% power	Actual power
0	0.08	0.09	968	70.16%
0.2	0.08	0.088	1014	72.16%
0.4	0.08	0.086	1060	74.04%
0.6	0.08	0.084	1112	76.05%
0.8	0.08	0.082	1166	78%
1	0.08	0.08	1226	80%

Note: $\bar{G}(\dot{\omega})$ is the share of frequentist researchers, θ^* is the true effect and $\bar{\theta}$ is the mean reported effect.

Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.joep.2019.01.006.

References

Andvig, J. C., & Moene, K. O. (1990). How corruption may corrupt. *Journal of Economic Behavior & Organization, 13*(1), 63–76.

Brocas, I., & Carrillo, J. D. (2018). A neuroeconomic theory of (dis)honesty. *Journal of Economic Psychology*.

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2013). Star wars: The empirics strike back. Working paper.

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

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... Altmejd, A. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour, 2*(9), 637.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., ... Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences, 112*(50), 15343–15347.

Coffman, L. C., & Niederle, M. (2015). Pre-analysis plans have limited upside, especially where replications are feasible. *Journal of Economic Perspectives, 29*(3), 81–98.

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., ... Dreber, A. (2018). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*.

Gall, T., Ioannidis, J. P., & Maniadis, Z. (2017). The credibility crisis in research: Can economics tools help? *PLoS Biology, 15*(4) e2001846.

Gall, T., & Maniadis, Z. (2019). Evaluating solutions to the problem of false positives. *Research Policy, 48*(2), 506–515.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641–651.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460.

Greene, W. (2012). *Econometric analysis* (7th ed.). Upper Saddle River, NJ: Pearson Education.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640–648.

Ioannidis, J. P. (2018). Meta-research: Why research on research matters. *PLoS Biology, 16*(3) e2005468.

Ioannidis, J., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal, 127*(605), F236–F265.

Maniadis, Z., Tufano, F., & List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review, 104*(1), 277–290.

Maniadis, Z., Tufano, F., & List, J. A. (2017). To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study. *The Economic Journal, 127*(605), F209–F235.

McShane, B. B., & Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association, 112*(519), 885–895.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021.

Neyman, J. (1937). X—outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London A, 236*(767), 333–380.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science, 9*(5), 552–555.

Stanley, T. D., & Doucouliagos, H. (2013). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*(1), 60–78.

Tirole, J. (1996). A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *The Review of Economic Studies, 63*(1), 1–22.

Zhang, L., & Ortmann, A. (2013). Exploring the meaning of significance in experimental economics. Retrieved from papers.ssrn.com/sol3/papers.cfm?abstract_id=2356018.