Identification of and correction for publication bias

By Isaiah Andrews and Maximilian Kasy*

Some empirical results are more likely to be published than others. Selective publication leads to biased estimates and distorted inference. We propose two approaches for identifying the conditional probability of publication as a function of a study's results, the first based on systematic replication studies and the second on meta-studies. For known conditional publication probabilities, we propose bias-corrected estimators and confidence sets. We apply our methods to recent replication studies in experimental economics and psychology, and to a meta-study on the effect of the minimum wage. When replication and meta-study data are available, we find similar results from both.

JEL: Publication bias, replication, meta-studies, identification, experimental economics, minimum wage
Keywords: C18, C12, C13

Despite following the same protocols, replications of published experiments frequently find effects of smaller magnitude or opposite sign than those in the initial studies (cf. Open Science Collaboration, 2015; Camerer et al., 2016). A leading explanation for replication failure is publication bias (cf. Ioannidis, 2005, 2008; McCrary et al., 2016; Christensen and Miguel, 2016). Journal editors and referees may be more likely to publish results that are statistically significant, that confirm some prior belief or, conversely, that are surprising. Researchers in turn face strong incentives to select which findings to write up and submit to journals based on the likelihood of ultimate publication, leading to what is sometimes called the file drawer problem (Rosenthal, 1979). We refer to these behaviors collectively as selective publication or publication bias. Left unaddressed, such selectivity can lead to biased estimates and misleading confidence sets in published studies.

We first show how bias from selective publication can be corrected if the conditional publication probability (i.e. the probability of publication as a function of a study's results) is known. We then show how the conditional publication probability can be nonparametrically identified. Finally, we apply the proposed methods to several empirical literatures.

* Andrews: Department of Economics, Harvard University, iandrews@fas.harvard.edu. Kasy: Department of Economics, Harvard University, maximiliankasy@fas.harvard.edu. We thank Josh Angrist, Ellora Derenoncourt, Gary Chamberlain, Xavier D'Haultfoeuille, Andrew Gelman, Gary King, Jesse Shapiro, Jann Spiess, and numerous seminar participants for helpful comments and suggestions. We also thank Paul Wolfson and Dale Belman as well as Michael Kremer for sharing their data. This research was funded in part by the Silverman (1968) Family Career Development Chair at MIT, and the National Science Foundation (under grant number 1654234), and the Alfred P. Sloan Foundation (under the grant "Publication bias and specification searching. Identification, correction, and reform proposals").

Correcting for publication bias After introducing our setup, Section I discusses the consequences of selective publication for statistical inference. When selectivity is known we propose median unbiased estimators and valid confidence sets for scalar parameters.[1]

Identification of publication bias Section II considers two approaches to identification. The first uses data from systematic replications of a collection of original studies. Following e.g. Camerer et al. (2016), by a replication we mean a study that applies the same experimental protocol to a new sample from the same population as the corresponding original study.[2] When there is no selectivity and the original and replication studies have the same sample size, the joint distribution of initial and replication estimates is symmetric, in the sense that it is unchanged when we reverse the roles of the original and replication results. Under the assumption that publication decisions depend only on the original estimates, asymmetries in this joint distribution nonparametrically identify conditional publication probabilities. While replication sample sizes often differ from those in the initial study, we show that nonparametric identification extends to this case as well.

Our second identification approach uses data from meta-studies, by which we mean studies that collect estimates and standard errors from multiple published studies. Under an independence assumption common in the meta-studies literature, if there is no selectivity then we can write the distribution of estimates for high variance studies as the distribution for low variance studies plus noise. Deviations from this prediction again identify conditional publication probabilities.

In applications where we can apply both approaches, which rely on different sources of identification, we find that they yield very similar conclusions. This finding adds to the credibility of our widely applicable meta-studies based method.

Both approaches identify conditional publication probabilities up to scale. Multiplying publication probabilities by a constant factor does not change the distribution of published results, and so does not affect the behavior of estimators and confidence sets. Hence, identification up to scale is sufficient to apply our bias corrections.

Applications Section III applies the theory developed in this paper to three empirical literatures. Our first two applications use data from the experimental economics and psychology replication studies of Camerer et al. (2016) and Open Science Collaboration (2015), respectively. Estimates based on our replication

[1] While our corrections eliminate bias due to selective publication, they cannot correct for problems with the underlying studies. If a study suffers from omitted variables bias (cf. Bruns and Ioannidis, 2016; Bruns, 2017), for instance, our corrections provide median unbiased estimates for the sum of the parameter of interest and the omitted variables bias. See Section IV.E below.

[2] Clemens (2017) terms such studies "reproductions," to distinguish them from "verifications" (cf. Chang and Li, 2018; Gertler et al., 2018) which try to reproduce the same results as the original paper based on the original sample.

approach suggest that results significant at the 5% level are over 30 times more likely to be published than are insignificant results, providing strong evidence of selectivity. Estimation based on our meta-study approach, which uses only the originally published results, yields similar conclusions.

Our third application considers the literature on the impact of minimum wages on employment, where no replication estimates are available. Estimates based on data from the meta-study Wolfson and Belman (2015) suggest that results corresponding to a negative and significant effect of minimum wages on employment are about 3 times more likely to be published than are insignificant results. Our point estimates suggest that results showing a positive and significant effect of minimum wages on employment are less likely to be published than negative and significant results, consistent with prior work by Card and Krueger (1995) and Wolfson and Belman (2015), but we cannot reject that selection depends only on significance and not on sign. In the supplement we discuss two additional applications of our methods, using data from Croke et al. (2016) and Camerer et al. (2018).

Alternative approaches There is a large prior literature on publication bias. Section IV discusses some of the alternatives from this literature, including meta-regression and approaches based on the distribution of p-values or z-statistics, and relates them to our framework. We further discuss the implications of "p-hacking" as studied by e.g. Simonsohn et al. (2014) and Bruns and Ioannidis (2016) for our results.

Supplement A variety of supporting materials and extensions of our results are provided in the online supplement. Section A contains proofs for all results discussed in the main text. Section B provides additional discussion of the data and methods used in our empirical applications, as well as a range of robustness checks. Section C contains further empirical results, including estimates based on alternative GMM estimation approaches and results for the Croke et al. (2016) and Camerer et al. (2018) applications. Finally, Section D discusses additional theoretical results, including on inference with multidimensional selection and the impact of selection on Bayesian inference.

Notation Throughout the paper, upper case letters denote random variables and lower case letters denote realizations. We observe normally distributed estimates X with mean Θ and standard error Σ, where Θ and Σ may vary across studies.[3] We condition on Θ and Σ whenever frequentist objects are considered, while unconditional expectations, probabilities, and densities integrate over the population distribution of Θ and Σ. Estimates normalized by their standard error Σ are denoted by Z, and parameters Θ normalized by Σ are denoted by Ω.

[3] Note that we use Σ to denote the (scalar) standard error rather than a variance matrix.

Latent studies (published or unpublished) are marked by a superscript $*$, while published studies have no superscript.

I. Setting

Throughout this paper we consider variants of the following data generating process. Within an empirical literature of interest, there is a population of latent studies i. The true effect Θ_i^* in study i is drawn from distribution μ_Θ. Thus, different latent studies may estimate different true parameters.[4] Conditional on the true effect Θ_i^* and the standard error Σ_i^* (which may also vary across studies), the result X_i^* in latent study i is drawn from the normal distribution $N(\Theta_i^*, \Sigma_i^{*2})$. For simplicity of notation we suppress the subscript i when possible.

Studies are published if $D = 1$, which occurs with probability $p(Z^*)$, where $Z^* = X^*/\Sigma^*$. We observe the truncated sample of published studies (that is, we observe draws from the conditional distribution of (X^*, Σ^*) given $D = 1$) and denote observations in this sample by (X, Σ). Publication decisions reflect both researcher and journal decisions; we do not attempt to disentangle the two. We obtain the following model:

DEFINITION 1 (Truncated sampling process): $(\Theta^*, \Sigma^*, X^*, D)$ *are jointly i.i.d. across latent studies, with*

$$\begin{aligned} (\Theta^*, \Sigma^*) &\sim \mu_{\Theta,\Sigma} \\ X^*|\Theta^*, \Sigma^* &\sim N(\Theta^*, \Sigma^{*2}) \\ D|X^*, \Theta^*, \Sigma^* &\sim Ber(p(Z^*)), \end{aligned}$$

where $Z^* = X^*/\Sigma^*$. *We observe i.i.d. draws* (X, Σ) *from the conditional distribution of* (X^*, Σ^*) *given* $D = 1$. *Define* $Z = X/\Sigma$, $\Omega^* = \Theta^*/\Sigma^*$, $\Omega = \Theta/\Sigma$, *and denote the marginal distribution of* Θ^* *by* μ_Θ.

As we discuss in the proofs, many of our results can be extended to the case where X^* is non-normal. Our focus on the normal case is motivated by the fact that that X^* represents the estimate in each study. Such estimates are approximately normal with a consistently estimable variance under mild conditions. Moreover, approximate normality of estimates is widely assumed in practice (for example to justify reporting standard errors), including in all the papers discussed in our applications.

The truncated sampling process of Definition 1 implies the likelihood.

$$f_{Z|\Omega,\Sigma}(z|\omega,\sigma) = f_{Z^*|\Omega^*,\Sigma^*,D}(z|\omega,\sigma,1) = \frac{p(z)}{E\left[p(Z^*)\,|\Omega^* = \omega\right]}\varphi(z-\omega), \tag{1}$$

[4] The case where all latent studies estimate the same parameter is nested by taking the distribution μ_Θ to be degenerate.

for $\varphi(\cdot)$ the standard normal density. Note that $f_{Z|\Omega,\Sigma}(z|\omega,\sigma) = f_{Z|\Omega}(z|\omega)$. Moreover, the scale of the publication probability does not affect the distribution of published results, since for $c > 0$, $p(\cdot)$ and $c \cdot p(\cdot)$ imply the same $f_{Z|\Omega}(z|\omega)$.

A. *Illustrative example: Selection on statistical significance*

To illustrate our setting we consider a simple example to which we will return throughout the paper. A journal receives a stream of studies reporting experimental estimates $X^* \sim N(\Theta^*, \Sigma^{*2})$ of treatment effects Θ^*, where each experiment examines a different treatment. The journal publishes studies with Z^* in the interval $[-1.96, 1.96]$ with probability $p(Z^*) = .1$, while results outside this interval are published with probability $p(Z^*) = 1$. This publication policy reflects a preference for "significant results," where a two-sided z-test rejects the null hypothesis $\Theta^* = 0$ (or equivalently, $\Omega^* = 0$) at the 5% level. This journal is ten times more likely to publish significant results than insignificant ones. Consequently, published results tend to over-estimate the magnitude of the treatment effect.[5] Published confidence intervals also under-cover the true parameter value for small values of Ω and over-cover for somewhat larger values. This is demonstrated by Figure 1, which plots the median bias, $med(\hat{\Omega}|\Omega = \omega) - \omega$, of the usual estimator $\hat{\Omega} = Z$, as well as the coverage of the conventional 95% confidence interval $[Z - 1.96, Z + 1.96]$.[6] While we have described this example in terms of selection by the journal, it could equivalently be interpreted as reflecting selection by researchers, or by both researchers and journals.

B. *Corrected inference*

If we know the form of selectivity we can correct the bias from selective publication. This section derives median unbiased estimators and valid confidence sets for Ω, which can immediately be turned into estimators and confidence sets for Θ via multiplication by Σ. These results ensure unbiasedness and correct coverage conditional on (Θ, Σ) for all (Θ, Σ), rather than just on average across the distribution of (Θ, Σ). For now we assume $p(\cdot)$ is known up to scale; corrections accounting for estimation error in $p(\cdot)$ are discussed in Section B.1 of the supplement.

Selective publication reweights the distribution of Z by $p(\cdot)$. To obtain valid estimators and confidence sets, we need to correct for this reweighting. To define these corrections, denote the distribution function for published results Z given true effect Ω by $F_{Z|\Omega} = \int_{-\infty}^{z} f_{Z|\Omega}(\tilde{z}|\omega)\, d\tilde{z}$, for $f_{Z|\Omega}(z|\omega)$ as in Equation (1). Recall that $f_{Z|\Omega}$ is the same for $p(\cdot)$ and $c \cdot p(\cdot)$, so we only need to know $p(\cdot)$ up to scale to calculate $F_{Z|\Omega}$. We adapt an approach previously applied by, among others, D. Andrews (1993) and Stock and Watson (1998), and invert the distribution

[5] See Ioannidis (2008) and Gelman (2018) for more discussion of this point.

[6] Note that $med(\hat{\Omega}|\Omega = \omega) - \omega = (med(\hat{\Theta}|\Theta = \theta, \Sigma = \sigma) - \theta)/\sigma$ so the median bias of $\hat{\Omega}$ can be interpreted as the median bias of X for θ, scaled by the standard error.

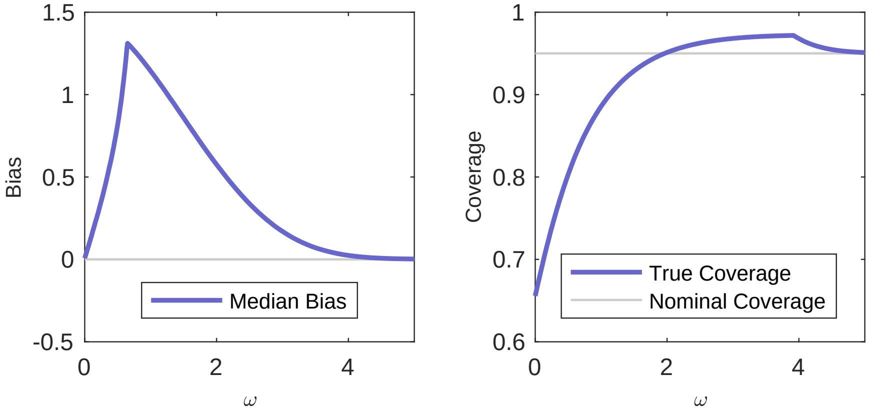

Figure 1. Bias and coverage conditional on publication

Note: The left panel plots the median bias of the conventional estimator $\hat{\Theta}_j = Z_j$, while the right panel plots the true coverage of the conventional 95% confidence interval, both for $p(z) = .1 + .9 \cdot \mathbf{1}(|Z| > 1.96)$.

function as a function of ω to construct a quantile-unbiased estimator. Let us define $\hat{\omega}_\alpha(z)$ as the solution to

$$F_{Z|\Omega}(z|\hat{\omega}_\alpha(z)) = \alpha \in (0,1), \tag{2}$$

so z lies at the α-quantile of the distribution implied by $\hat{\omega}_\alpha(z)$. Using the monotonicity properties of $F_{Z|\Omega}$, we prove that $\hat{\omega}_\alpha(Z)$ is an α-quantile unbiased estimator for Ω.

PROPOSITION 1: *Suppose that $p(z) > 0$ for all z, and $p(\cdot)$ is almost everywhere continuous. Then $\hat{\omega}_\alpha(z)$ as defined in (2) exists, is unique, and is continuous and strictly increasing for all z. Furthermore, $\hat{\omega}_\alpha(Z)$ is α-quantile unbiased for Ω under the truncated sampling setup of Definition 1,*

$$P(\hat{\omega}_\alpha(Z) \leq \omega | \Omega = \omega, \Sigma = \sigma) = \alpha \textit{ for all } \omega.$$

These results allow straightforward frequentist inference that corrects for selective publication. In particular, using Proposition 1 we can consider the median-unbiased estimator $\hat{\omega}_{\frac{1}{2}}(z)$ for ω, as well as the equal-tailed level $1-\alpha$ confidence interval $\left[\hat{\omega}_{\frac{\alpha}{2}}(Z), \hat{\omega}_{1-\frac{\alpha}{2}}(Z)\right]$. This estimator and confidence set fully correct the bias and coverage distortions induced by selective publication. In the special case where insignificant results are published with probability zero while significant results are published with probability one, our corrected confidence sets exclude zero if and only if the test of McCrary et al. (2016) rejects.

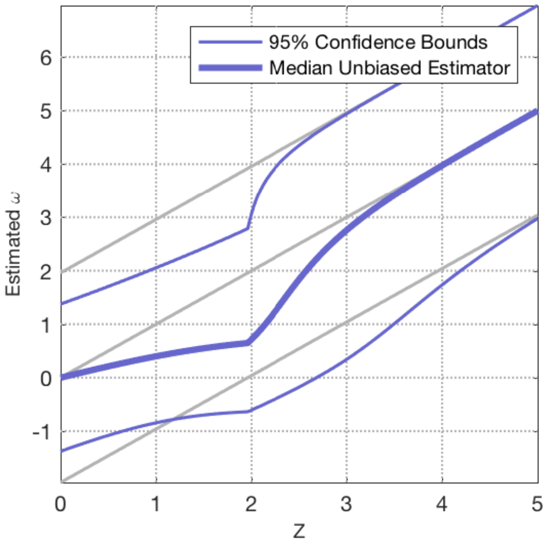

Figure 2. Bias correction

Note: This figure plots 95% confidence bounds and the median unbiased estimator for the normal model where results that are significant at the 5% level are ten times more likely to be published than are insignificant results. The usual (uncorrected) estimator and confidence bounds are plotted in grey for comparison.

Illustrative example (continued) To illustrate these results, we return to the treatment effect example discussed above. Figure 2 plots the median unbiased estimator, as well as upper and lower 95% confidence bounds, as a function of Z, again for the case with $p(Z^*) = 1$ when $|Z^*| > 1.96$ and $p(Z^*) = .1$ otherwise. We see that the median unbiased estimator lies below the usual estimator $\hat{\omega} = Z$ for small positive Z but that the difference is eventually decreasing in Z. The truncation-corrected confidence interval shown in Figure 2 has exactly correct coverage, is smaller than the usual interval for small Z, wider for moderate values Z, and essentially the same for $Z \geq 5$.

II. Identifying selection

This section proposes two approaches for identifying $p(\cdot)$. The first uses systematic replication studies, while the second uses meta-studies.

A. *Systematic replication studies*

The following proposition extends the model in Definition 1 above to incorporate a conditionally independent replication draw X^{r*} which is observed whenever

X^* is. The key assumption for this proposition is that selectivity of publication operates only on X^* and not on X^{r*}. This assumption is plausible for systematic replication studies such as Open Science Collaboration (2015) and Camerer et al. (2016), but may fail in non-systematic replication settings, for instance if replication studies are published only when they "debunk" prior published results.

PROPOSITION 2 (Nonparametric identification using replication experiments): *Consider the data generating process of Definition 1. Assume that for each latent study there exist a replication estimate and standard error* (X^{r*}, Σ^{r*}) *with*

$$X^{r*}|\Theta^*, \Sigma^{r*}, \Sigma^*, D, X^* \sim N(\Theta^*, \Sigma^{*r2}),$$

where we again observe the replication estimate and standard error only for published studies. Then $p(\cdot)$ *is identified up to scale, and* μ_Θ *is identified as well.*

INTUITION Consider the setup of Proposition 2, and define $Z^r = X^r/\Sigma$, that is as the *replication* estimate normalized by the *original* standard error. Assume for the moment that $\Sigma^{r*} = \Sigma^*$, so that the replication estimate X^{r*} has the same variance as X^*. Under these assumptions, the marginal density of (Z, Z^r) is

$$f_{Z,Z^r}(z, z^r) = \frac{p(z)}{E[p(Z^*)]} \int \varphi(z-\omega)\varphi(z^r-\omega)d\mu_\Omega(\omega). \tag{3}$$

This expression immediately implies that any asymmetries in the joint distribution of (Z, Z^r) must be due to the publication probability $p(\cdot)$. In particular,

$$\frac{f_{Z,Z^r}(b,a)}{f_{Z,Z^r}(a,b)} = \frac{p(b)}{p(a)},$$

whenever the denominators on either side are non-zero. Proposition 2 uses this identity to show that $p(\cdot)$ is nonparametrically identified up to scale.[7] That $p(\cdot)$ is only identified up to scale is intuitive: Equation (1) above shows that the scale of $p(\cdot)$ does not affect the distribution of published results, and Equation (3) shows that the same remains true once we add replication results. Hence, the scale of $p(\cdot)$ is both unnecessary for bias corrections and unidentified without data on unpublished results.

In general the replication standard error Σ^{r*} will differ from the original variance Σ^*, which takes us out of the symmetric framework. Additionally, the distribution of Σ^{r*} might depend on Z^*. Such dependence is present if power calculations are used to determine replication sample sizes, as in both Open Science Collaboration (2015) and Camerer et al. (2016). In that case, Σ^{r*} is positively related to the magnitude of Z^*, but conditionally unrelated to Θ^*. The proof of Proposition

[7] Note that this argument does not use normality of Z and Z^r, and thus generalizes to other estimator distributions.

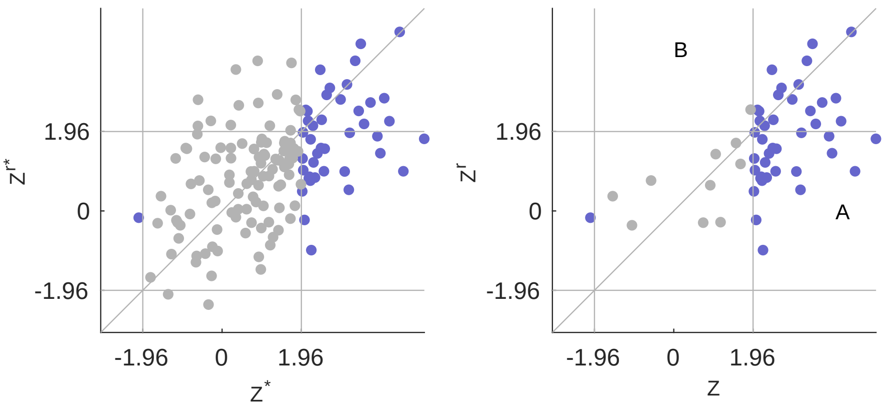

Figure 3. Identification using systematic replication studies

Note: This figure illustrates the effect of selective publication in the replication experiments setting using simulated data, where selection is on statistical significance, as described in the text. The left panel shows the joint distribution of a random sample of latent estimates and replications; the right panel shows the subset which are published. Results where the original estimates are significantly different from zero at the 5% level are plotted in blue, while insignificant results are plotted in grey.

2 shows that identification carries over to this setting, since we can recover the symmetric setting by (de)convolution of Z^r with normal noise.

Illustrative example (continued) To illustrate our identification approach using replication studies, we return to the illustrative example introduced in Section I. In this setting, suppose that the normalized true effect Ω^* is distributed $N(1, 1)$ across latent studies. As before, assume that $p(Z^*) = 1$ when $|Z^*| > 1.96$, and that $p(Z^*) = .1$ otherwise. Assume finally that $\Sigma^{r*} = \Sigma^* = 1$, so original and replication estimates both have variance one.

This setting is illustrated in Figure 3. The left panel of this figure shows 100 random draws (Z^*, Z^{r*}); draws where $|Z^*| \leq 1.96$ are marked in grey, while draws where $|Z^*| > 1.96$ are marked in blue. The right panel shows the subset of draws (Z, Z^r) that are published. These are the same draws as (Z^*, Z^{r*}), except that 90% of the draws for which Z^* is statistically insignificant are deleted.

Our identification argument in this case proceeds by considering deviations from symmetry around the diagonal $Z = Z^r$. Let us compare what happens in the regions marked A and B. In A, Z is statistically significant but Z^r is not; in B it is the other way around. By symmetry of the data generating process, the latent (Z^*, Z^{r*}) fall in either area with equal probability. The fact that the observed (Z, Z^r) lie in region A substantially more often than in region B thus provides evidence of selective publication, and the exact deviation of the distribution of (Z, Z^r) from symmetry identifies $p(\cdot)$ up to scale.

B. Meta-studies

Our approach using meta-studies restricts the model in Definition 1 by assuming that Θ^* is statistically independent of Σ^* across latent studies, so studies with smaller standard errors do not have systematically different estimands. This is a strong assumption, but is imposed by many popular meta-analysis techniques including in meta-regression (see Section IV.B) and the "trim and fill" method (Duval and Tweedie, 2000). This assumption holds trivially if Θ^* is constant across latent studies. In our applications with replication data, estimates for $p(\cdot)$ based on this assumption are similar to those based on our replication approach, lending further support to this method.

PROPOSITION 3 (Nonparametric identification using meta-studies): *Consider the data generating process of Definition 1. Assume additionally that Σ^* and Θ^* are independent, and that the support of Σ contains an open interval. Then $p(\cdot)$ is identified up to scale, and μ_Θ is identified as well.*

INTUITION Consider the setup of Proposition 3. The conditional density of Z given Σ is

$$f_{Z|\Sigma}(z|\sigma) = \frac{p(z)}{E[p(Z^*)|\Sigma^* = \sigma]} \int \varphi(z - \theta/\sigma) d\mu_\Theta(\theta).$$

This implies that, for $\sigma_2 > \sigma_1$,

$$\frac{f_{Z|\Sigma}(z|\sigma_2)}{f_{Z|\Sigma}(z|\sigma_1)} = \frac{E[p(Z^*)|\Sigma^* = \sigma_1]}{E[p(Z^*)|\Sigma^* = \sigma_2]} \cdot \frac{\int \varphi(z - \theta/\sigma_2) d\mu_\Theta(\theta)}{\int \varphi(z - \theta/\sigma_1) d\mu_\Theta(\theta)}, \tag{4}$$

where the first term on the right hand side does not depend on z. Since $f_{Z|\Sigma}(z|\sigma_2)/f_{Z|\Sigma}(z|\sigma_1)$ is identified, this suggests we might be able to invert this equality to recover μ_Θ, which would then allow us to identify $p(\cdot)$. The proof of Proposition 3 builds on this idea.

ILLUSTRATIVE EXAMPLE (CONTINUED) As before, assume that Θ^* is $N(1, 1)$ distributed, that $p(Z^*) = 1$ when $|Z^*| > 1.96$, and that $p(Z^*) = .1$ otherwise. Suppose further that Σ^* is independent of Θ^* across latent studies. This setting is illustrated in Figure 4. The left panel of this figure shows 100 random draws (X^*, Σ^*); draws where $|X^*/\Sigma^*| \leq 1.96$ are marked in grey, while draws where $|X^*/\Sigma^*| > 1.96$ are marked in blue. The right panel shows the subset of draws (X, Σ) that are published, where 90% of statistically insignificant draws are deleted.

Compare what happens for two different values of the standard error Σ, marked by A and B in Figure 4. By the independence of Σ^* and Θ^*, the distribution of X^* for larger values of Σ^* is a noised up version of the distribution for smaller values of Σ^*. To the extent that the same does not hold for the distribution of published X given Σ, this must be due to selectivity in the publication process.

In this example, statistically insignificant observations are "missing" for larger values Σ. Since publication is more likely when $|X^*/\Sigma^*| > 1.96$, the estimated values X tend to be larger on average for larger values of Σ, and the details of how the conditional distribution of X given Σ varies with Σ will again allow us to identify $p(\cdot)$ up to scale.

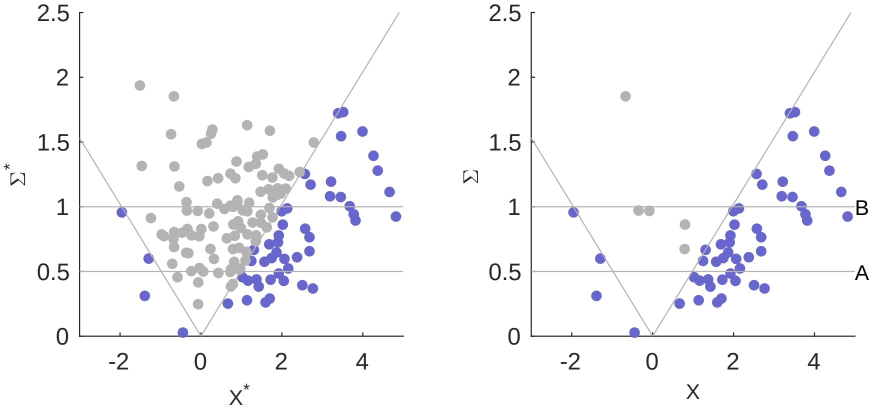

Figure 4. Identifcation using meta-studies

Note: This figure illustrates the effect of selective publication in the meta-studies setting using simulated data, where selection is on statistical significance, as described in the text. The left panel shows a random sample of latent estimates; the right panel shows the subset of estimates which are published. Results which are significantly different from zero at the 5% level are plotted in blue, while insignificant results are plotted in grey.

C. *Estimation*

The sample sizes in our applications are limited, which makes fully nonparameteric estimation impractical. In the supplement we build on our identification arguments to derive GMM estimators that assume a functional form for the conditional publication probability $p(\cdot)$ but are nonparametric in the distribution μ of true effects. For simplicity and ease of exposition, however, in the main text we specify parsimonious parametric models for both $p(\cdot)$ and μ which we fit by maximum likelihood, similar to Hedges (1992). Our nonparametric identification results suggest that there is hope for estimation robust to functional form assumptions, and this is borne out by the similarity of the maximum likelihood estimates reported here to the GMM results reported in the supplement.

We consider step function models for $p(\cdot)$, with jumps at conventional critical values, and possibly at zero. Since $p(\cdot)$ is only identified up to scale, we impose the normalization $p(z) = 1$ for $z > 1.96$ throughout. This is without loss of generality, since $p(\cdot)$ is allowed to be larger than 1 for other cells. We assume different parametric models for the distribution of latent effects Θ^*, discussed case-by-case below. In our first two applications the sign of the original estimates is normalized to be positive.[8] We denote these normalized estimates by $W = |Z|$, and in these settings we impose that $p(\cdot)$ is symmetric.

III. Applications

This section applies the results developed above to estimate the degree of selectivity in three empirical literatures. We first consider data from the large scale replication studies Camerer et al. (2016) and Open Science Collaboration (2015), which examine experimental studies in economics and psychology, respectively. We then turn to the meta-study Wolfson and Belman (2015) on the effect of the minimum wage on employment. We consider two additional applications in the supplement, using replication data from Camerer et al. (2018) on social-science experiments and meta-study data from Croke et al. (2016) on the effect of deworming.

Plausibility of identifying assumptions The results of Section II imply nonparametric identification of both $p(\cdot)$ and μ_Θ. Our approach using replication data is based on the assumption that selection for publication depends only on the original estimates and not on the replication estimates. This assumption is highly plausible by design in the two replication settings we consider, which use data from systematic replication studies. These studies pre-specify and replicate a large number of results published in a given time period and set of journals, and report all replication results together.

Our approach using meta-studies is based on the assumption that studies on a given topic with different standard errors do not have systematically different estimands. While we cannot guarantee validity of this assumption by design, its plausibility is enhanced by our finding that it yields estimates very similar to the approach based on replication studies in all our applications where both apply (Camerer et al. (2016), Open Science Collaboration (2015), and Camerer et al. (2018)). Variants of this assumption (or the strictly stronger assumption that Θ is constant) are common in existing meta-studies.

Finally, for both approaches we assume that conditional on (Θ^*, Σ^*) estimates are approximately normal, consistent with the inference methods used in the underlying studies.

A. *Economics laboratory experiments*

Our first application uses data from a recent large-scale replication of experimental economics papers by Camerer et al. (2016). The authors replicated all

[8] The studies in these datasets consider different outcomes, so the relative signs of effects across studies are arbitrary. Setting the sign of the initial estimate in each study to be positive ensures invariance to the sign normalization chosen by the authors of each study.

18 between-subject laboratory experiment papers published in the American Economic Review and Quarterly Journal of Economics between 2011 and 2014.[9] Further details on the selection and replication of results can be found in Camerer et al. (2016), while details on our handling of the data are discussed in the supplement.

A strength of this dataset for our purposes, beyond the availability of replication estimates, is the fact that it replicates results from all papers in a particular subfield published in two leading economics journals over a fixed period of time. This mitigates concerns about the selection of which studies to replicate. Moreover, since the authors replicate 18 such studies, it seems likely that they would have published their results regardless of what they found, consistent with our assumption that selection operates only on the initial studies and not on the replications.

A caveat to the interpretation of our results is that Camerer et al. (2016) select the most important statistically significant finding from each paper, as emphasized by the original authors, for replication. This selection changes the interpretation of $p(\cdot)$, which has to be interpreted as the probability that a result was published *and* selected for replication. In this setting, our corrected estimates and confidence intervals provide guidance for interpreting the headline results of published studies. For consistency with the rest of the paper, however, we continue to shorthand $p(\cdot)$ as the publication probability.

Histogram Before we discuss our formal estimation results, consider the distribution of originally published estimates $W = |Z|$, shown by the histogram in the left panel of Figure 5. This histogram suggests a large jump in the density $f_W(\cdot)$ at the cutoff 1.96, and thus a corresponding jump in the publication probability $p(\cdot)$ at the same cutoff; see Section IV.C below. Such a jump is confirmed by both our replication and meta-study approaches.

Results from replication specifications The middle panel of Figure 5 plots the joint distribution of $(W, W^r) = \text{sign}(Z) \cdot (Z, Z^r)$ in the replication data of Camerer et al. (2016). To estimate the degree of selection in these data we consider the model

$$|\Omega^*| \sim \Gamma(\kappa, \lambda), \quad p(Z) \propto \begin{cases} \beta_p & |Z| < 1.96 \\ 1 & |Z| \geq 1.96. \end{cases}$$

This assumes that the absolute value of the normalized true effect Ω^* follows a gamma distribution with shape parameter κ and scale parameter λ. This nests

[9] In their supplementary materials, Camerer et al. (2016) state that "To be part of the study a published paper needed to report at least one significant between subject treatment effect that was referred to as statistically significant in the paper." However, we have reviewed the issues of the American Economic Review and Quarterly Journal of Economics from the relevant period, and confirmed that no studies were excluded due to this restriction.

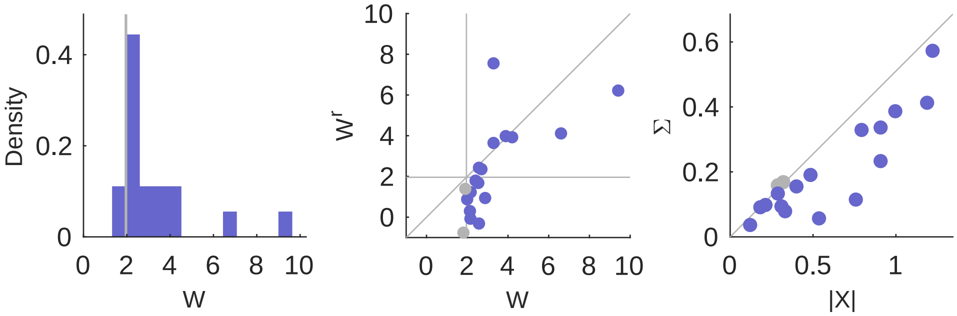

FIGURE 5. CAMERER ET AL. (2016) DATA

Note: The left panel shows a binned density plot for the normalized z-statistics $W = |X|/\Sigma$ using data from Camerer et al. (2016). The grey line marks $W = 1.96$. The middle panel plots the z-statistics W from the initial study against the estimate W^r from the replication study. The grey lines mark W and $W^r = 1.96$, as well as $W = W^r$. The right panel plots the initial estimate $|X| = W \cdot \Sigma$ against its standard error Σ. The grey line marks $|X|/\Sigma = 1.96$.

TABLE 1—SELECTION ESTIMATES FOR OPEN SCIENCE COLLABORATION (2015)

Replication			Meta-study		
κ	λ	β_p	$\tilde{\kappa}$	$\tilde{\lambda}$	β_p
0.337	2.411	0.029	1.343	0.157	0.038
(0.236)	(1.074)	(0.028)	(1.299)	(0.076)	(0.051)

Note: Selection estimates from lab experiments in economics, with robust standard errors in parentheses. The left panel reports estimates from replication specifications, while the right panel reports results from meta-study specifications. Publication probability β_p is measured relative to the omitted category of studies significant at 5% level, so an estimate of 0.029 implies that results which are insignificant at the 5% level are 2.9% as likely to be published as significant results. The parameters (κ, λ) and $(\tilde{\kappa}, \tilde{\lambda})$ are not comparable.

a wide range of cases, including χ^2 and exponential distributions, while keeping the number of parameters low. Our model for $p(\cdot)$ allows a discontinuity in the publication probability at $|Z| = 1.96$, the critical value for a 5% two-sided z-test. Fitting this model by maximum likelihood yields the estimates reported in the left panel of Table 1. Recall that β_p in this model can be interpreted as the publication probability for a result that is insignificant at the 5% level based on a two-sided z-test, relative to a result that is significant at the 5% level. These estimates therefore imply that significant results are more than thirty times more likely to be published than insignificant results. Moreover, we strongly reject the hypothesis of no selectivity, $H_0 : \beta_p = 1$.

RESULTS FROM META-STUDY SPECIFICATIONS While the Camerer et al. (2016) data include replication estimates, we can also apply our meta-study approach using just the initial estimates and standard errors. Since this approach relies on additional independence assumptions, comparing these results to those based on replication studies provides a useful check of the reliability of our meta-analysis estimates.

We begin by plotting the data used by our meta-analysis estimates in the right panel of Figure 5. We consider the model

$$|\Theta^*| \sim \Gamma(\tilde{\kappa}, \tilde{\lambda}), \quad p(Z) \propto \begin{cases} \beta_p & |Z| < 1.96 \\ 1 & |Z| \geq 1.96. \end{cases}$$

noting that Θ^* is the mean of X^*, rather than Z^*, and thus that the interpretation of $(\tilde{\kappa}, \tilde{\lambda})$ differs from that of (κ, λ) in our replication specifications. Fitting this model by maximum likelihood yields the estimates reported in the right panel of Table 1. Comparing these estimates to those in the left panel, we see that the estimates from the two approaches are similar, though the metastudy estimates suggest a somewhat smaller degree of selection. Hence, we find that in the Camerer et al. (2016) data we obtain similar results from our replication and meta-study specifications.

Bias correction To interpret our estimates, we calculate our median-unbiased estimator and confidence sets based on our replication estimate $\beta_p = .029$. Figure 6 plots the median unbiased estimator, as well as the original and adjusted confidence sets, for the 18 studies included in Camerer et al. (2016). Considering the first panel, which plots the median unbiased estimator along with the original and replication estimates, we see that the adjusted estimates track the replication estimates fairly well but are smaller than the original estimates in many cases.[10] The second panel plots the original estimate and conventional 95% confidence set in blue, and the adjusted estimate and 95% confidence set in black. As we see from this figure, twelve of the adjusted confidence sets include zero, compared to just two of the original confidence sets. Hence, adjusting for the estimated degree of selection substantially changes the number of significant results in this setting.

B. *Psychology laboratory experiments*

Our second application is to data from Open Science Collaboration (2015), who conducted a large-scale replication of experiments in psychology. The authors considered studies published in three leading psychology journals, Psychological Science, Journal of Personality and Social Psychology, and Journal of Experimental Psychology: Learning, Memory, and Cognition, in 2008. They assigned papers to replication teams on a rolling basis, with the set of available papers determined by publication date. Ultimately, 158 articles were made available for replication,

[10] Note, however, that even for $p(\cdot)$ known it is not the case that the conditional median of Z^r given Z is equal to the adjusted estimate. Indeed, the conditional distribution of Z^r given Z does not depend on $p(\cdot)$.

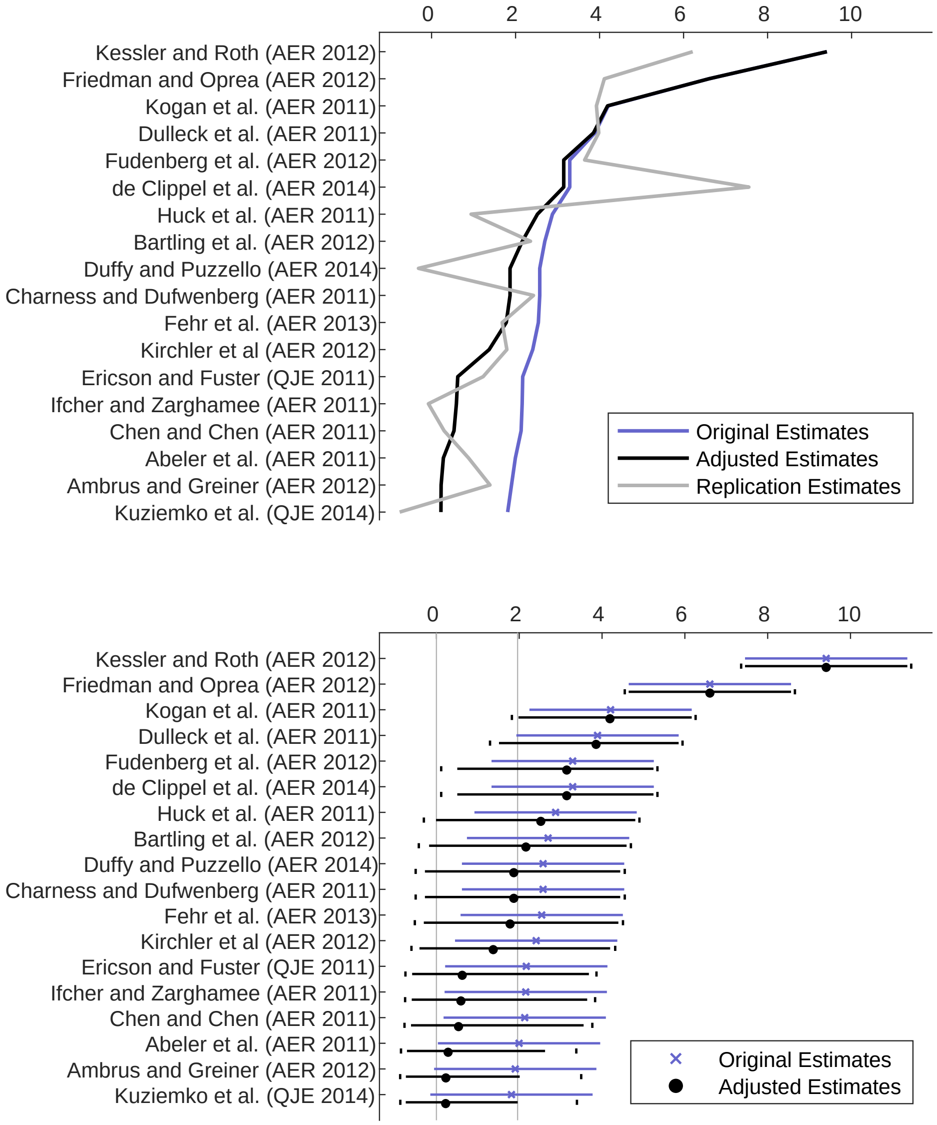

Figure 6. Adjusted estimates for Camerer et al. (2016)

Note: The top panel plots the estimates W and W^r from the original and replication studies in Camerer et al. (2016), along with the median unbiased estimate $\hat{\theta}_{\frac{1}{2}}$ based on the estimated selection model and the original estimate. The bottom panel plots the original estimate and 95% confidence interval, as well as the median unbiased estimate and adjusted 95% confidence interval $\left[\hat{\theta}_{0.025}\left(W\right),\hat{\theta}_{0.975}\left(W\right)\right]$ based on the estimated selection model. Adjusted intervals not accounting for estimation error in the selection model are plotted with solid lines, while endpoints for intervals accounting for estimation error are marked with "|" – see Section B.1 of the supplement.

111 were assigned, and 100 of those replications were completed in time for inclusion in Open Science Collaboration (2015). Replication teams were instructed to replicate the final result in each article as a default, though deviations from this default were made based on feasibility and the recommendation of the authors of the original study. Ultimately, 84 of the 100 completed replications consider the final result of the original paper.

As with the economics replications above, the systematic selection of results for replication in Open Science Collaboration (2015) is an advantage from our perspective. A complication in this setting, however, is that not all of the test statistics used in the original and replication studies are well-approximated by z-statistics (for example, some of the studies use χ^2 test statistics with two or more degrees of freedom). To address this, we limit attention to the subset of studies which use z-statistics or close analogs thereof, leaving us with a sample of 73 studies. Specifically, we limit attention to studies using z- and t-statistics, or χ^2 and F-statistics with one degree of freedom (for the numerator, in the case of F-statistics), which can be viewed as the squares of z- and t-statistics, respectively. To explore sensitivity of our results to denominator degrees of freedom for t- and F-statistics, in the supplement we limit attention to the 52 observations with denominator degrees of freedom of at least 30 in the original study and find quite similar results.

Histogram The distribution of originally published estimates W is shown by the histogram in the left panel of Figure 7. This histogram suggests a large jump in the density $f_W(\cdot)$ at the cutoff 1.96, as well as possibly a jump at the cutoff 1.64, and thus of corresponding jumps of the publication probability $p(\cdot)$ at the same cutoffs. Such jumps are again confirmed by the estimates from both our replication and meta-study approaches.

Results from replication specifications The middle panel of Figure 7 plots the joint distribution of W, W^r in the replication data of Open Science Collaboration (2015). Relative to the plot for Camerer et al. (2016), we see a larger fraction of studies where $W > 1.96$ for the original study while $W^r < 1.96$ in the replication study (8 of the 18 of studies in Open Science Collaboration (2015), compared to 43 of the 73 studies in Camerer et al. (2016)).[11] This could be due to differences in selection or to differences in the distribution of effects. To disentangle these issues, we fit the model

$$|\Omega^*| \sim \Gamma(\kappa, \lambda), \quad p(Z) \propto \begin{cases} \beta_{p,1} & |Z| < 1.64 \\ \beta_{p,2} & 1.64 \leq |Z| < 1.96 \\ 1 & |Z| \geq 1.96. \end{cases}$$

[11] Indeed 12 of the 73 studies in Open Science Collaboration (2015) have $W > 3$ and $W^r < 1.96$, while none of those in Camerer et al. (2016) do.

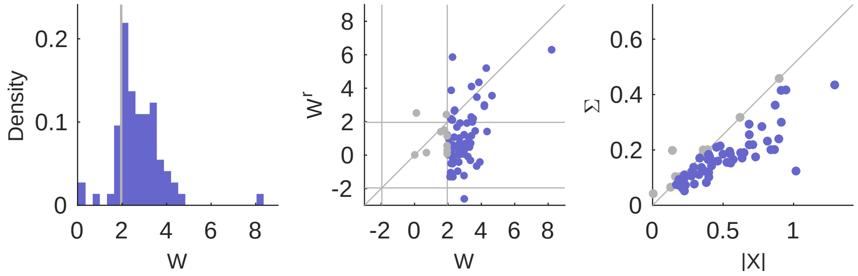

Figure 7. Open Science Collaboration (2015) data

Note: The left panel shows a binned density plot for the normalized z-statistics $W = |X|/\Sigma$ using data from Open Science Collaboration (2015). The grey line marks $W = 1.96$. The middle panel plots the z-statistics W from the initial study against the estimate W^r from the replication study. The grey lines mark $|W|$ and $|W^r| = 1.96$, as well as $W = W^r$. The right panel plots the initial estimate $|X| = W \cdot \Sigma$ against its standard error Σ. The grey line marks $|X|/\Sigma = 1.96$.

Table 2—Selection estimates for Open Science Collaboration (2015)

Replication				Meta-study			
κ	λ	$\beta_{p,1}$	$\beta_{p,2}$	$\tilde{\kappa}$	$\tilde{\lambda}$	$\beta_{p,1}$	$\beta_{p,2}$
0.311	1.314	0.009	0.205	0.974	0.153	0.017	0.306
(0.118)	(0.296)	(0.005)	(0.087)	(0.549)	(0.053)	(0.009)	(0.135)

Note: Selection estimates from lab experiments in psychology, with robust standard errors in parentheses. The left panel reports estimates from replication specifications, while the right panel reports results from meta-study specifications. Publication probabilities β_p are measured relative to the omitted category of studies significant at 5% level. The parameters (κ, λ) and $(\tilde{\kappa}, \tilde{\lambda})$ are not comparable.

This model again assumes that the absolute value of the normalized true effect $|\Omega^*|$ follows a gamma distribution across latent studies. Given the larger sample size, we consider a slightly more flexible model than before and allow discontinuities in the publication probability at the critical values for both 5% and 10% two-sided z-tests.

Fitting this model by maximum likelihood yields the estimates reported in the left panel of Table 2. These estimates imply that results that are significantly different from zero at the 5% level are over a hundred times more likely to be published than results that are insignificant at the 10% level, and nearly five times more likely to be published than results that are significant at the 10% level but insignificant at the 5% level. We strongly reject the hypothesis of no selectivity.

These results do not indicate a large difference in the degree of selection relative to the Camerer et al. (2016) data.[12] They suggest, however that the distribution

[12] If we instead estimate the model only with a discontinuity at the 5% level (as in the Camerer et al. (2016) data), we estimate $\beta_p = 0.024$ with standard error of 0.009.

of $|\Omega^*|$ may be substantially smaller, with $E[|\Omega^*|] = 0.41$ (standard error 0.1) in the Open Science Collaboration (2015) data compared to $E[|\Omega^*|] = 0.81$ (standard error 0.39) in the Camerer et al. (2016) data. These estimates for $E[|\Omega^*|]$ are noisy but suggest that the larger number of studies with $W > 1.96$ and $W^r < 1.96$ in Open Science Collaboration (2015) may be due to differences in the distribution of true effects, rather than to differences in the degree of selection.

Our results for this setting are roughly consistent with those of Johnson et al. (2017), who independently consider the Open Science Collaboration (2015) data and likewise estimate a step function model for $p(\cdot)$, but allow a discontinuity only at the 5% significance level. Johnson et al. (2017) estimate that insignificant results are only about 0.5% as likely to be published as are significant results. The Johnson et al. specifications for μ_Ω allow the possibility that $Pr\{\Omega^* = 0\} > 0$ and they estimate that $\Omega^* = 0$ about 90% of the time. Similarly, our estimated gamma distribution has mode equal to zero.

Results from meta-study specifications As before, we re-estimate our model using our meta-study specifications, and plot the joint distribution of estimates and standard errors in the right panel of Figure 7. Fitting the model yields the estimates reported in the right panel of Table 2. As in the last section, we find that the meta-study and replication estimates are broadly similar, though the meta-study estimates again suggest a somewhat more limited degree of selection

Approved replications Gilbert et al. (2016) argue that the protocols in some of the Open Science Collaboration (2015) replications differed substantially from the initial studies. These arguments were disputed by many of the Open Science Collaboration (2015) authors in Anderson et al. (2016), who note that many of the replications used protocols approved in advance by the authors of the underlying papers. In Section B.6.2 of the supplement we report results based on the subset of approved replications and find roughly similar estimates, though the estimated degree of selection is smaller.

Bias corrections To interpret our results, we plot our median-unbiased estimates based on the Open Science Collaboration (2015) data in Figure 8. We see that our adjusted estimates track the replication estimates fairly well for studies with small original z-statistics, though unlike in Figure 6 differences are larger for studies with larger original z-statistics.[13]

Our adjustments again dramatically change the number of significant results, with 62 of the 73 original 95% confidence sets excluding zero, and only 28 of the adjusted confidence sets (not displayed) doing the same.

[13] Since we have sorted on the original estimates, patterns of this sort can arise from mean reversion.

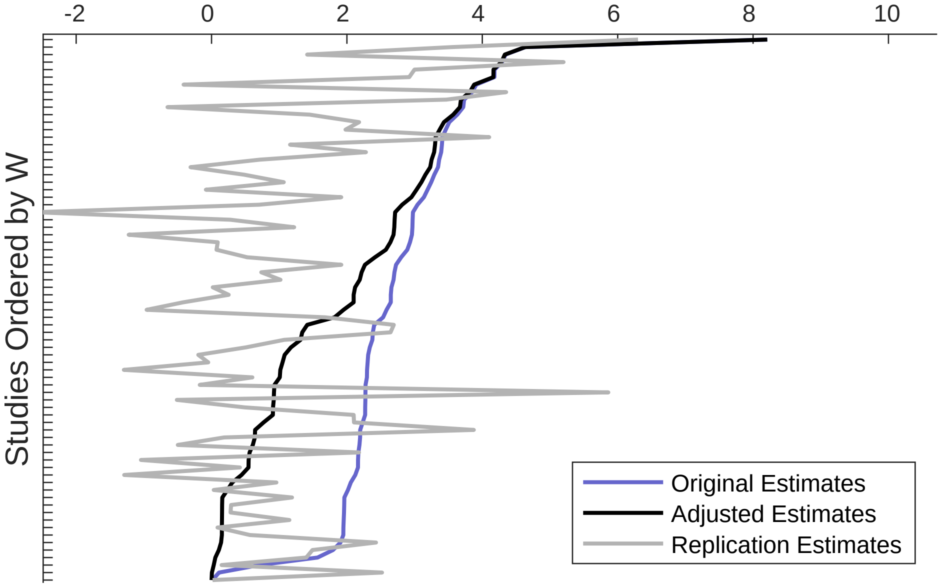

FIGURE 8. ADJUSTED ESTIMATES FOR OPEN SCIENCE COLLABORATION (2015)

Note: This figure plots the estimates W and W^r from the original and replication studies in Open Science Collaboration (2015), along with the median unbiased estimate $\hat{\theta}_{\frac{1}{2}}$ based on the estimated selection model and the original estimate.

C. *Effect of minimum wage on employment*

Our final application uses data from Wolfson and Belman (2015), who conduct a meta-analysis of studies on the elasticity of employment with respect to the minimum wage. In particular, Wolfson and Belman (2015) collect analyses of the effect of minimum wages on employment that use US data and were published or circulated as working papers after the year 2000. They collect estimates from all studies fitting their criteria that report both estimated elasticities of employment with respect to the minimum wage and standard errors, resulting in a sample of a thousand estimates drawn from 37 studies, and we use these estimates as the basis of our analysis. For further discussion of these data, see Wolfson and Belman (2015).

Since the Wolfson and Belman (2015) sample includes both published and unpublished papers, we evaluate our estimators based on both the full sample and the sub-sample of published estimates. We find qualitatively similar answers for the two samples, so we report results based on the full sample here and discuss results based on the subsample of published estimates in the supplement. We

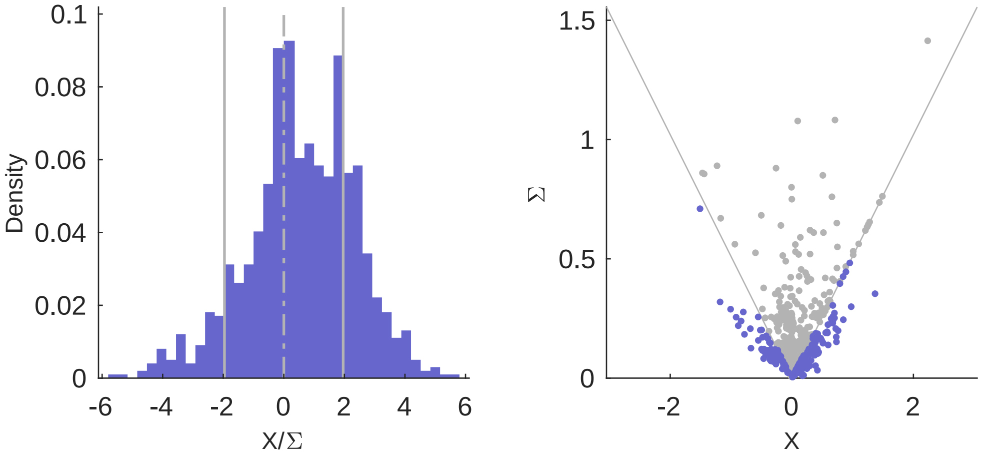

Figure 9. Wolfson and Belman (2015) data

Note: The left panel shows a binned density plot for the z-statistics X/Σ in the Wolfson and Belman (2015) data. The solid grey lines mark $|X|/\Sigma = 1.96$, while the dash-dotted grey line marks $X/\Sigma = 0$. The right panel plots the estimate X against its standard error Σ. The grey lines mark $|X|/\Sigma = 1.96$.

define X so that $X > 0$ indicates a negative effect of the minimum wage on employment.

Histogram Consider first the distribution of the normalized estimates Z, shown by the histogram in the left panel of Figure 9. This histogram is somewhat suggestive of jumps in the density $f_Z(\cdot)$ around the cutoffs -1.96, 0, and 1.96, and thus of corresponding jumps in the publication probability $p(\cdot)$ at the same cutoffs; these jumps seem less pronounced than in our previous applications, however.

Results from meta-study specifications For this application we do not have any replication estimates, and so move directly to our meta-study specifications. The right panel of Figure 9 plots the joint distribution of X, the estimated elasticity of employment with respect to decreases in the minimum wage, and the standard error Σ in the Wolfson and Belman (2015) data.

We consider the model

$$\Theta^* \sim \bar{\theta} + t(\nu) \cdot \tilde{\tau}, \quad p(Z) \propto \begin{cases} \beta_{p,1} & Z < -1.96 \\ \beta_{p,2} & -1.96 \leq Z < 0 \\ \beta_{p,3} & 0 \leq Z < 1.96 \\ 1 & Z \geq 1.96. \end{cases}$$

Since the data are not sign-normalized, we model Θ^* using a t distribution with degrees of freedom $\tilde{\nu}$ and location and scale parameters $\bar{\theta}$ and $\tilde{\tau}$, respectively. Unlike in our previous applications, we allow the probability of publication to depend on the sign of the z-statistic Z rather than just on its absolute value. This is important, since it seems plausible that the publication prospects for a study could differ depending on whether it found a positive ($X < 0$) or negative ($X > 0$) effect of the minimum wage on employment.

Our estimates based on these data are reported in Table 3, where we find that results which are insignificant at the 5% level are about 30% as likely to be published as are significant estimates finding a negative effect of the minimum wage on employment. Our point estimates also suggest that studies finding a positive and significant effect of the minimum wage on employment may be less likely to be published, but this estimate is quite noisy and we cannot reject the hypothesis that selection depends only on significance and not on sign. Unlike our other results, this is sensitive to the details of the specification: if we instead restrict the distribution of true effects Θ^* to be normal, our estimate for $\beta_{p,1}$ drops to 0.225 with a standard error of 0.118. On the other hand, our GMM approach discussed in Section C.1 of the supplement returns a $\beta_{p,1}$ estimate of 1.174 with a standard error of 0.417.

Table 3—Selection estimates for Wolfson and Belman (2015)

$\bar{\theta}$	$\tilde{\tau}$	$\tilde{\nu}$	$\beta_{p,1}$	$\beta_{p,2}$	$\beta_{p,3}$
0.018	0.019	1.303	0.697	0.270	0.323
(0.009)	(0.011)	(0.279)	(0.350)	(0.111)	(0.094)

Note: Meta-study estimates from minimum wage data, with standard errors clustered by study in parentheses. Publication probabilities β_p measured relative to omitted category of estimates positive and significant at 5% level.

Since the studies in this application estimate related parameters, it is interesting to consider the estimate $\bar{\theta}$ for the mean effect in the population of latent estimates. The point estimate is small but significantly different from zero at the 5% level, and suggests that the average latent study finds a small negative effect of the minimum wage on employment. This effect is about half as large as the "naive" average effect $\bar{\theta}$ we would estimate by ignoring selectivity, .041 with a standard error of 0.011.

These results are consistent with the meta-analysis estimates of Wolfson and Belman (2015), who found evidence of some publication bias towards a negative employment effect, as well as the results of Card and Krueger (1995), who focused on an earlier, non-overlapping set of studies.

Multiple estimates A complication arises in this application, relative to those considered so far, due to the presence of multiple estimates per study. Since it is difficult to argue that a given estimate in each of these studies constitutes the "main" result, restricting attention to a single estimate per study would be arbitrary. This somewhat complicates inference and identification.

For inference, it is implausible that estimate standard-error pairs (X, Σ) are independent within study. To address this, we cluster our standard errors by study.

For identification, the problem is somewhat more subtle. Our model assumes that the latent parameters Θ_i^* and Σ_i^* are statistically independent across estimates i, and that D_i is independent of (Θ_i^*, Σ_i^*) conditional on Z_i^*. It is straightforward to relax the assumption of independence across i, provided the marginal distribution of $(\Theta_i^*, \Sigma_i^*, X_i^*, D_i)$ is such that D_i remains independent of (Θ_i^*, Σ_i^*) conditional on Z_i^*. This conditional independence assumption is justified if we believe that both researchers and referees consider the merits of each estimate on a case-by-case basis, and so decide whether or not to publish each estimate separately. Alternatively, it can also be justified if the estimands Θ^* within each study are statistically independent (relative to the population of estimands in the literature under consideration).

IV. Alternative approaches

Many approaches to detecting selectivity and publication bias have been proposed in the literature. Good reviews are provided by Rothstein et al. (2006) and Christensen and Miguel (2016). In this section we analyze some of these approaches through the lens of our framework and relate them to our results.

A. *Should results "replicate?"*

The findings of recent systematic replication studies such as Open Science Collaboration (2015) and Camerer et al. (2016) are sometimes interpreted as indicating an inability to "replicate the results" of published research. In this setting, a "result" is understood to "replicate" if both the original study and its replication find a statistically significant effect in the same direction. The share of results which replicate in this sense is prominently discussed in Camerer et al. (2016). Our framework shows that the probability of replication in this sense might be low even without selective publication or other sources of bias.

Consider the setup for replication experiments of Proposition 2, with constant publication probability $p(\cdot)$, so that publication is not selective and $f_{Z,Z^r} = f_{Z^*,Z^{r*}}$. For illustration, assume further that $\Sigma^* = \Sigma^{r*}$ with probability 1. For Φ the standard normal distribution function, the probability that a result replicates in the sense described above is

$$P(Z^{r*} \cdot sign\{Z^*\} > 1.96 || Z^*| > 1.96) = \frac{\int \left[\Phi(-1.96 - \omega)^2 + \Phi(-1.96 + \omega)^2\right] d\mu_\Omega(\omega)}{\int \left[\Phi(-1.96 - \omega) + \Phi(-1.96 + \omega)\right] d\mu_\Omega(\omega)}.$$

If the true effect is zero in all studies then this probability is 0.025. If the true effect in all studies is instead large, so that $|\Omega^*| > M$ with probability one for some large M, then the probability of replication is approximately one. Thus, any replication probability between 0.025 and one is consistent with no selection, and low replication frequencies are not necessarily indicative of selective publication, but could instead be due to a large share of small true effects. Strengths and weaknesses of alternative measures of replication are discussed in Simonsohn (2015) and Patil and Peng (2016).

B. *Meta-regressions*

A popular test for publication bias in meta-studies (cf. Card and Krueger, 1995; Egger et al., 1997) is based on meta-regression, which uses regressions of either of the following forms:

$$E^*[X|1,\Sigma] = \gamma_0 + \gamma_1 \cdot \Sigma, \qquad E^*\left[Z|1,\tfrac{1}{\Sigma}\right] = \beta_0 + \beta_1 \cdot \tfrac{1}{\Sigma},$$

where we use E^* to denote best linear predictors. Under the assumptions of Proposition 3, if $p(\cdot)$ is constant then it follows immediately that

$$E^*[X|1,\Sigma] = E[\Theta^*], \quad E^*\left[Z|1,\tfrac{1}{\Sigma}\right] = E[\Theta^*] \cdot \tfrac{1}{\Sigma}.$$

Hence, testing that either $\gamma_1 = 0$ or $\beta_0 = 0$ delivers a valid test for the null hypothesis of no selectivity, though there are some forms of selectivity against which such tests have no power.

For our minimum wage application, a regression of X on Σ yields an intercept of 0.006 (standard error 0.038) and a slope of 0.408 (standard error 0.372).[14] A regression of Z on $1/\Sigma$ yields an intercept of 0.343 (standard error 0.283) and a slope of 0.018 (standard error 0.009). In particular, neither of these regressions allows to reject the null of no selectivity at a 5% level, in contrast to the estimates discussed in Section III.C.

Absent publication bias, γ_0 and β_1 recover the average of Θ^* in the population of latent studies. Our estimates of γ_0 and β_1 are 0.006 and 0.018. These coefficients are sometimes interpreted as selection-corrected estimates of the mean effect across studies (cf. Doucouliagos and Stanley, 2009; Christensen and Miguel, 2016), but this interpretation is potentially misleading in the presence of publication bias. In particular, the conditional expectation $E[X|1,\Sigma]$ is nonlinear in both Σ and $1/\Sigma$, which implies that β_0, γ_1 are generally biased as estimates of $E[\Theta^*]$.[15] We discuss a simple example with one-sided significance testing in Section D.1 of the supplement.

A variety of generalizations to meta-regression have been proposed in the literature, including by Stanley and Doucouliagos (2014), who propose to use power-

[14] The sign normalization in our economics and psychology lab experiment applications means that meta-regression does not apply in these settings.

[15] Stanley (2008) and Doucouliagos and Stanley (2009) note this bias but suggest that one can still use $H_0 : \gamma_1 = 0$ to test the hypothesis of zero true effect if there is no heterogeneity in the true effect Θ^* across latent studies.

weighted meta-regressions to increase robustness to selective publication, and Stanley et al. (2017) who consider non-linear meta-regressions. Meta-regressions have also been widely used in applications, including by Carter et al. (2017), Havránek (2015), and Ioannidis et al. (2017).[16]

C. *The distribution of p-values and z-statistics*

Another approach in the literature considers the distribution of p-values, or the corresponding z-statistics, across published studies (cf. De Long and Lang, 1992; Schuemie et al., 2014; Simonsohn et al., 2014; Brodeur et al., 2016, 2018). Assuming normality, there is a one-to-one mapping between the distribution of p-values P and the distribution of z-statistics Z, since $P = 1 - \Phi(Z)$ for 1-sided tests of the null hypothesis $\theta = 0$ or, equivalently, $\omega = 0$.[17] Under our model, absent selectivity in the publication process the distribution f_Z is equal to f_{Z^*}. For $Z^*|\Omega^* \sim N(\Omega^*, 1)$ and $\Omega^* \sim \mu_\Omega$, this implies that

$$f_Z(z) = f_{Z^*}(z) = (\mu_\Omega * \varphi)(z) = \int \varphi(z - \omega) d\mu_\Omega(\omega).$$

This model implies that the density f_{Z^*} is infinitely differentiable. If selectivity is present, by contrast, then $f_Z(z) = \frac{p(z)}{E[p(Z^*)]} \cdot f_{Z^*}(z)$. Any discontinuity of $f_Z(z)$ (for instance at critical values such as $z = 1.96$) thus identifies a corresponding discontinuity of the conditional publication probability $p(z)$:

$$\frac{\lim_{z \downarrow z_0} f_Z(z)}{\lim_{z \uparrow z_0} f_Z(z)} = \frac{\lim_{z \downarrow z_0} p(z)}{\lim_{z \uparrow z_0} p(z)}. \tag{5}$$

If we impose that $p(\cdot)$ is a step function, this identifies $p(\cdot)$ up to scale.

In the context of our applications, we estimate the size of this discontinuity by considering the ratio of histogram bars above and below the threshold, where we use the same bins as in Figures 5 and 7. For the application to economics laboratory experiments, we find a jump in the publication probability at 1.96 equal to 4. A two-sided test of the null of equal mass above and below the threshold gives a p-value of 0.0215. For the application to psychology laboratory experiments, we find a jump in the publication probability at 1.64 equal to 7, with a corresponding p-value of 0.0078, and a jump in the the publication probability at 1.96 equal to 2.3, with a corresponding p-value of 0.0347.

The model without selectivity, $f_Z(z) = f_{Z^*}(z) = (\mu_\Omega * \varphi)(z)$, has testable implications beyond smoothness. In particular, the density f_{Z^*} precludes excessive bunching, since for all $k \geq 0$ and all z, $\partial_z^k f_{Z^*}(z) \leq \sup_z \partial_z^k \varphi(z)$ and $\partial_z^k f_{Z^*}(z) \geq \inf_z \partial_z^k \varphi(z)$ so for example $f_{Z^*}(z) \leq \varphi(0)$ and $f''_{Z^*}(z) \geq \varphi''(0) = -\varphi(0)$ for all z.

[16] Other recent work examining selective publication in economics and finance using non meta-regression approaches includes Chen and Zimmermann (2017) and Hou et al. (2017).

[17] For two-sided tests, the mapping is between p-values and absolute z-statistics $|Z|$.

Spikes in the distribution of Z thus likewise indicate the presence of selectivity or inflation.

D. Observability

The setup of Definition 1 assumes that we only observe the draws (X^*, Σ^*) for which $D = 1$. In some cases, however, additional information may be available. First, we might know of the existence of unpublished studies, for example from experimental preregistrations, without observing their results X^*. In this case, called censoring, we observe i.i.d. draws of (Y, D), where $Y = D \cdot Z^*$.[18] The corresponding censored likelihood is

$$f_{Y,D|\Omega^*}(y, d|\omega^*) = d \cdot p(y) \cdot \varphi(y - \omega) + (1 - d) \cdot (1 - E[p(Z^*)|\Omega^* = \omega^*]).$$

Second, we might additionally observe the results Z^* from unpublished working papers as in Franco et al. (2014). The likelihood in this case is

$$f_{Z^*,D|\Omega^*}(z, d|\omega) = p(z)^d (1 - p(z))^{1-d} \cdot \varphi(z - \omega).$$

Even under these alternative observability assumptions, the truncated likelihood (1) arises as a limited information likelihood that conditions on publication decisions and/or unpublished results. Our identification and inference results therefore continue to apply.

That said, additional information allows identification of $p(\cdot)$ under weaker assumptions. With full observability of unpublished results Z^*, for example, $p(\cdot)$ is identified by simply regressing D on Z^*, cf. Franco et al. (2014).

E. Bias and Pseudo-True Values

Bruns and Ioannidis (2016) and Bruns (2017) discuss an additional way in which selectivity may increase bias in observational studies. To cast their concern into our framework, recall that we assume throughout that the distribution of X^* in latent studies is normal and centered on Θ^*. There are different ways this model can be interpreted.

A first interpretation is that Θ_i^* is the "true" parameter of interest in study i. This would for example be the case for randomized experiments where we have no reason to doubt the internal validity of each study. In this case any variation of Θ_i^* across studies i considering the same question is due to issues of external validity, for instance to different populations of experimental subjects, or to effects changing over time. In this setting our corrections yield valid estimates and confidence sets for the parameters of interest.

A second interpretation of our model is that researchers consider different estimates X^* of the same parameter. These estimates might for instance be based

[18] We could also observe the standard error Σ for published studies, but suppress this for simplicity.

on different controls, different outcome variables, different estimation methods, and so on. These estimates have expectations Θ^* that vary across specifications, so not all Θ^* correspond to the "true" effect of interest. Put differently, variation of Θ^* across studies might be due to violations of internal validity, in addition to issues of external validity.[19] Under this second interpretation, we have additional sources of bias. First, $E[\Theta] \neq E[\Theta^*]$ in general, so selection can lead to different average biases among published and latent studies. This effect can persist even as sampling noise goes to zero.[20] Second, even if we avoid this bias by using our approach to identify μ_Θ and therefore $E[\Theta^*]$, there is no guarantee that $E[\Theta^*]$ corresponds to the parameter of interest. Hence, while our corrections can undo selection bias and allow inference on either the parameter Θ in a given study or the distribution μ_Θ of Θ^* in the population of latent studies, we cannot correct deficiencies in the underlying studies.

F. Manipulation and P-hacking

Some authors consider the possibility that researchers manipulate their results (Brodeur et al., 2016; Furukawa, 2017), while others consider the selection of results within papers, which Simonsohn et al. (2014) term "p-hacking." Our primary focus in this paper is on researchers decisions whether or not to submit findings, and journal decisions whether or not to publish submissions, rather than on manipulation or p-hacking. Nonetheless, depending on the form manipulation or p-hacking takes, it may still be consistent with our baseline model.

To illustrate, consider an experimental setting where researchers run two independent versions of an experiment, or estimate two regression specifications for the same estimand. Suppose first that they decide whether to report an estimate for each experiment or specification separately. In this case our baseline model applies, save that Θ_i^* is no longer i.i.d. Suppose now alternatively that the researcher decides to always report only the more significant of the two estimates. In this case, the probability of publication of the first estimate depends on the underlying parameter via the second estimate, so publication probabilities are of the form $p(Z_i^*, \Omega_i^*)$.

To accommodate such violations of our baseline model, we discuss the extension of our approach to settings where the selection probability may depend on both Z and Ω in Section D.3 of the supplement. Given normal replication estimates X^r, we show that in this setting we can still identify enough features of the model to apply selection-corrections. We also develop specification tests for our baseline model against this more general alternative, however, and in no case do we reject our baseline model where $p(\cdot)$ does not depend on Ω given Z.

[19] If some studies are viewed as more credible than others, this highlights the value of conducting inference on Θ for individual studies, rather than merely on the distribution μ_Θ.

[20] Consider for instance the case where $E[\Theta^*] = 0$ and positive results are more likely to be published.

V. Conclusion

This paper makes three contributions relative to the existing literature on selective publication. First, we provide methods to calculate bias-corrected estimators and confidence sets when the form of selectivity is known. Second, we provide nonparametric identification results for selectivity based on replications and meta-studies. Third, we apply the proposed methods to several literatures, documenting the varying scale and kind of selectivity. In cases where both our replication and meta-study approaches apply, they yield similar conclusions.

Implications for empirical research What can researchers and readers of empirical research take away from this paper? First, when conducting a meta-analysis of the findings of some literature, researchers may wish to apply our methods to assess the degree of selectivity, and to apply appropriate corrections to individual estimates, tests, and confidence sets. We provide code on our webpages which implements the proposed methods for a flexible family of selection models.[21]

Second, our results provide guidance for how to interpret published empirical findings. In particular, if a reader has a view about how the selection process operates in a given literature, they can adjust published estimates and confidence sets as discussed in Section 4. Even if one is concerned that the selection model does not capture all sources of bias, these corrections aid interpretation by showing how much selection, considered in isolation, changes the interpretation of published results. A positive message from our results is that published estimates remain informative even when publication is quite selective.

It should be emphasized that we do not advocate adjusting publication standards to reflect our corrected critical values. If these cutoffs were to be systematically used in the publication process, this would simply entail an "arms race" of selectivity, rendering the more stringent critical values invalid again.

Optimal publication rules One might take the findings in this paper, and the debate surrounding publication bias more generally, to indicate that the publication process should be non-selective with respect to findings. Selective reporting by researchers might be eliminated by pre-analysis plans, cf. Olken (2015). Going one step further, selective publication by journals might be eliminated by result-blind review, cf. American Society of Health Economists (2015). The Journal of Development Economics now offers authors the option of pre-results review. The hope would be that non-selectivity of the publication process might restore the validity (unbiasedness, size control) of standard inferential methods.

Note, however, that optimal publication rules may depend on results. This can for instance be the case in models where policy decisions are made based on published findings. Section D.6 in the supplement provides a stylized example of

[21] We have also implemented out meta-study approach in a web app: https://maxkasy.github.io/home/metastudy/

such a setting. Alternatively, given evidence that experts can forecast experimental results quite well (cf. DellaVigna and Pope, 2018), excessively surprising findings might be interpreted as evidence of implementation problems and so weigh against publication. A broader study of the question of optimal publication from a journal's perspective can be found in Frankel and Kasy (2018).

Supplement The supplement contains a wide variety of results to complement those discussed in the main text. Section A provides proofs, while Section B gives additional details for our empirical applications and considers a range of robustness checks, including allowing publication probabilities to depend on covariates such as the journal or the year in which a paper was initially circulated. Section C derives novel GMM estimation approaches that leave the distribution of true effects unrestricted, and reports results for our applications. Section C also reports ML estimates for the Croke et al. (2016) and Camerer et al. (2018) applications. Finally, Section D reports additional theoretical results, including extensions of our identification results to allow publication probabilities to depend on Σ (to reflect a preference for precise estimates) and on Ω (to nest violations of our baseline model). This section also extends our inference results to cases where selection is driven by multiple variables, and discusses the effect of selection on Bayesian inference.

REFERENCES

American Society of Health Economists (2015). Editorial statement on negative findings. https://www.ashecon.org/american-journal-of-health-economics/.

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., ..., and Zuni, K. (2016). Response to comment on estimating the reproducibility of psychological science. *Science*, 351(6277):1037–1037.

Andrews, D. W. (1993). Exactly median-unbiased estimation of first order autoregressive/unit root models. *Econometrica*, 61(1):139–165.

Brodeur, A., Cook, N., and Heyes, A. (2018). Methods matter: P-hacking and causal inference in economics. Working Paper.

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. B. (2017). Meta-regression models and observational research. *Oxford Bulletin of Economics and Statistics*, 79(5):637–653.

Bruns, S. B. and Ioannidis, J. P. (2016). P-curve and p-hacking in observational research. *PLoS One*, 11(2):e0149144.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (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., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., and Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644.

Card, D. and Krueger, A. B. (1995). Time-series minimum-wage studies: A meta-analysis. *American Economic Review*, 85(2):238–243.

Carter, E., Schönbrodt, F., Gervais, W. M., and Hilgard, J. (2017). Correcting for bias in psychology: A comparison of meta-analytic methods. Unpublished Manuscript.

Chang, A. C. and Li, P. (2018). Is economics research replicable? sixty published papers from thirteen journals say" usually not". *Critical Finance Review*, 7.

Chen, A. Y. and Zimmermann, T. (2017). Selection bias and the cross-section of expected returns. Unpublished Manuscript.

Christensen, G. S. and Miguel, E. (2016). Transparency, reproducibility, and the credibility of economics research. NBER Working Paper No. 22989.

Clemens, M. A. (2017). The meaning of failed replications: a review and proposal. *Journal of Economic Surveys*, 31(1):326–342.

Croke, K., Hicks, J. H., Hsu, E., Kremer, M., and Miguel, E. (2016). Does mass deworming affect child nutrition? Meta-analysis, cost-effectiveness, and statistical power. Technical Report 22382, National Bureau of Economic Research.

De Long, J. B. and Lang, K. (1992). Are all economic hypotheses false? *Journal of Political Economy*, 100(6):1257–1272.

DellaVigna, S. and Pope, D. (2018). Predicting experimental results: Who knows what? *Journal of Political Economy.* Forthcoming.

Doucouliagos, H. and Stanley, T. (2009). Publication selection bias in minimum-wage research? A meta-regression analysis. *British Journal of Industrial Relations*, 47(2):406–428.

Duval, S. and Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95(449):89–98.

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, 315(7109):629–634.

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

Frankel, A. and Kasy, M. (2018). Working paper. Which Findings Should Be Published?

Furukawa, C. (2017). Unbiased publication bias: Theory and evidence. Unpublished Manuscript.

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin*, 44(1):16–23. Unpublished Manuscript.

Gertler, P., Galiani, S., and Romero, M. (2018). How to make replication the norm. *Nature*, 554:417–419.

Gilbert, D. T., King, G., Pettigrew, S., and Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277):1037.

Havránek, T. (2015). Measuring intertemporal substitution: The importance of method choices and selective reporting. *Journal of the European Economic Association*, 13(6):1180–1204.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, pages 246–255.

Hou, K., Xue, C., and Zhang, L. (2017). Replicating anomalies.

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

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

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

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112(517):1–10.

McCrary, J., Christensen, G., and Fanelli, D. (2016). Conservative tests under satisficing models of publication bias. *PloS one*, 11(2):e0149590.

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

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Patil, P. and Peng, R. D. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11(4):539–44.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641.

Rothstein, H. R., Sutton, A. J., and Borenstein, M. (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments.* John Wiley & Sons.

Schuemie, M. J., Ryan, P. B., DuMouchel, W., Suchard, M. A., and Madigan, D. (2014). Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in medicine*, 33(2):209–218.

Simonsohn, U. (2015). Small telescopes detectability and the evaluation of replication results. *Psychological Science*, 26(5):559–569.

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534–547.

Stanley, T., Doucouliagos, H., and Ioannidis, J. (2017). Finding the power to reduce publication bias. *Statistics in medicine*, 36(10):1580–1598.

Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and Statistics*, 70(103-127).

Stanley, T. D. and Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1):60–78.

Stock, J. and Watson, M. (1998). Median unbiased estimation of coefficient variance in a time-varying parameter model. *Journal of the American Statistical Association*, 93(441):349–358.

Wolfson, P. J. and Belman, D. (2015). 15 years of research on us employment and the minimum wage. *Available at SSRN 2705499.*