Identification of and correction for publication bias*

Isaiah Andrews† Maximilian Kasy‡

March 22, 2017

Abstract

Some empirical results are more likely to be published than others. Such selective publication leads to biased estimators and distorted inference. This paper proposes 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 based on meta-studies. For known conditional publication probabilities, we propose median-unbiased estimators and associated confidence sets that correct for selective publication. We apply our methods to recent large-scale replication studies in experimental economics and psychology, and to meta-studies of the effects of minimum wages and de-worming programs.

Keywords: Publication bias, replication, meta-studies, identification

JEL Codes: C18, C12, C13

1 Introduction

Despite following the same protocols, replications of published experiments frequently find effects substantially smaller than those in the initial studies (cf. Open Science Collaboration, 2015; Camerer et al., 2016). One leading explanation for replication

*We thank Josh Angrist, Ellora Derenoncourt, Gary Chamberlain, Xavier D'Haultfoeuille, Gary King, Jesse Shapiro, Jann Spiess, and seminar participants at Brown, CREST, Harvard/MIT, Microsoft Research, and the Harvard development retreat for many helpful comments and suggestions. We also thank Paul Wolfson and Dale Belman as well as Michael Kremer for sharing their data.

†Department of Economics, MIT, iandrews@mit.edu

‡Department of Economics, Harvard University, maximiliankasy@fas.harvard.edu

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, results that confirm some prior belief or, conversely, results 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. Together, these forms of selectivity lead to severe bias in published estimates and confidence sets.

This paper provides, to the best of our knowledge, the first nonparametric identification results for the conditional publication probability as a function of the empirical results of a study. Once the conditional publication probability is known, we derive bias-corrected estimators and confidence sets. Finally, we apply the proposed methods to several empirical literatures.

Identification of publication bias Section 3 considers two approaches to identification. The first uses data from systematic replications of a collection of original studies, each of which applies the same experimental protocol to a new sample from the same population as the corresponding original study. Absent selectivity, the joint distribution of initial and replication estimates is symmetric. Asymmetries in this joint distribution nonparametrically identify conditional publication probabilities, assuming the latter only depend on the initial estimate. The second approach uses data from meta-studies. Absent selectivity, the distribution of estimates for high variance studies is a noisier version of the distribution for low variance studies. Deviations from this prediction identify conditional publication probabilities if we assume independence between the precision and true effect size across studies.

Correcting for publication bias Section 4 discusses the consequences of selective publication for statistical inference. For known selectivity, we propose median unbiased estimators and valid confidence sets for scalar parameters. These results allow valid inference on the parameters of each study, rather than merely on average effects across a given literature. The supplement extends these results and derives optimal quantile-unbiased estimators for scalar parameters of interest in the presence of nuisance parameters, as well as results on Bayesian inference.

Applications Section 5 applies the theory developed in this paper to four empirical literatures. We first 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 approach suggest that results significant at the 5% level are 10 to 50 times more likely to be published than are insignificant results, providing strong evidence of selectivity. Estimation using our meta-study approach, which uses only the originally published results, yields similar conclusions.

We then consider two settings where no replication estimates are available. The first is the literature on the impact of minimum wages on employment. Estimates based on data from the meta-study by Wolfson and Belman (2015) suggest that results finding a negative and significant effect of minimum wages on employment are four times more likely to be included in this meta-study than results finding a positive and significant effect. Second, we consider the literature on the impact of mass deworming on child body weight. Estimates based on data from the meta-study by Croke et al. (2016) find that results appear more likely to be included in this meta-study when they do not find a significant impact of deworming, though we cannot reject the null hypothesis of no selectivity.

Literature There is a large literature on publication bias; good reviews are provided by Rothstein et al. (2006) and Christensen and Miguel (2016). We will discuss some of the approaches from this literature in the context of our framework below. One popular method, used in e.g. Card and Krueger (1995) and Egger et al. (1997), regresses z-statistics on the inverse of the standard error and takes a non-zero intercept as evidence of publication bias. Our approach using meta-studies builds on related intuitions. Another approach in the literature considers the distribution of p-values or z-statistics across studies, and takes bunching, discontinuities, or non-monotonicity in this distribution as indication of selectivity or estimate inflation (cf. De Long and Lang, 1992; Brodeur et al., 2016). Other approaches include the "trim and fill" method (Duval and Tweedie, 2000) and parametric selection models (Iyengar and Greenhouse, 1988; Hedges, 1992). Some precedent for our proposed corrections to inference can be found in McCrary et al. (2016), while the parametric models in our applications are related to those of Hedges (1992). Other recent work on publication bias includes Chen and Zimmermann (2017) and Furukawa (2017).

Road map Section 2 introduces the setting we consider, as well as a running example. Section 3 presents our main identification results, and discusses approaches from the literature. Section 4 discusses bias-corrected estimators and confidence sets, assuming conditional publication probabilities are known. Section 5 presents results for our empirical applications. All proofs are given in the supplement, which also contains details of our applications, additional empirical and theoretical results, and a stylized model of optimal publication decisions.

Notation Throughout the paper, upper case letters denote random variables and lower case letters denote realizations. The latent parameter governing the distribution of observables for a given study is Θ. We condition on Θ whenever frequentist objects are considered, while unconditional expectations, probabilities, and densities integrate over the population distribution of Θ across studies. Estimates are denoted by X, while estimates normalized by their standard deviation are denoted by Z. Latent studies (published or unpublished) are indexed by i and marked by a superscript $*$, while published studies are indexed by j. Subscripts i and j will sometimes be omitted when clear from context.

2 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. Thus, different latent studies may estimate different true parameters. The case where all latent studies estimate the same parameter is nested by taking the distribution μ to be degenerate.

Conditional on the true effect, the result X_i^* in latent study i is drawn from a known continuous distribution with density $f_{X^*|\Theta^*}$. We take both X_i^* and Θ_i^* to be scalar unless otherwise noted. Studies are published if $D_i = 1$, which occurs with probability $p(X_i^*)$, and we observe the truncated sample of published studies (that is, we observe X_i^* if and only if $D_i = 1$). Publication decisions reflect both researcher and journal decisions; we do not attempt to disentangle the two. Let I_j denote the index i corresponding to the jth published study. We obtain the following model:

Definition 1 (Truncated sampling process)
Consider the following data generating process for latent (unobserved) variables.

(Θ_i^*, X_i^*, D_i) *are jointly i.i.d. across* i*, with*

$$\begin{aligned}\Theta_i^* &\sim \mu \\ X_i^*|\Theta_i^* &\sim f_{X^*|\Theta^*}(x|\Theta_i^*) \\ D_i|X_i^*, \Theta_i^* &\sim Ber(p(X_i^*))\end{aligned}$$

Let $I_0 = 0$*,* $I_j = \min\{i : \ D_i = 1, \ i > I_{j-1}\}$ *and* $\Theta_j = \Theta_{I_j}^*$*. We observe i.i.d. draws*

$$X_j = X_{I_j}^*.$$

Section 3 considers extensions of this model that allow us to identify and estimate $p(\cdot)$. Section 4 assumes $p(\cdot)$ is known, which allows us to perform inference on Θ_j when X_j is observed. Of central importance throughout is the likelihood of observing X_j given Θ_j:

Lemma 1 (Truncated likelihood)
The truncated sampling process of Definition 1 implies the following likelihood:

$$f_{X|\Theta}(x|\theta) = f_{X^*|\Theta^*,D}(x|\theta, 1) = \frac{p(x)}{E\left[p\left(X_i^*\right)|\Theta_i^* = \theta\right]} f_{X^*|\Theta^*}(x|\theta). \tag{1}$$

For fixed θ, selective publication reweights the distribution of published results by $p(\cdot)$. As we consider different values of θ for fixed x, by contrast, the likelihood is scaled by the publication probability for a latent study with true effect θ, $E\left[p\left(X_i^*\right)|\Theta_i^* = \theta\right]$.

Study-level covariates The model of Definition 1, and in particular independence between publication decisions and Θ^* given X^*, may only hold conditional on some set of observable study characteristics. For example, journals may treat studies on particular topics, or using particular research designs, differently. Likewise, the distribution of true effects may differ across these categories. In such cases, we can condition our analysis on these variables and apply our approach separately to papers with different topics, research designs, and so on. For simplicity of notation, however, we suppress such additional conditioning.

2.1 An illustrative example

To illustrate our setting we consider a simple example to which we will return throughout the paper. A journal receives a stream of studies $i = 1, 2, \ldots$ reporting experimental estimates $Z_i^* \sim N(\Theta_i^*, 1)$ of treatment effects Θ_i^*, where each experiment examines a different treatment. We denote the estimates by Z^* rather than X^* here to emphasize that they can be interpreted as z-statistics. Denote the distribution of treatment effects across latent studies by μ. Normality is in many cases a plausible asymptotic approximation; $\text{Var}(Z^*|\Theta^*) = 1$ is a scale normalization. The journal publishes studies with Z_i^* in the interval $[-1.96, 1.96]$ with probability $p(Z_i^*) = .1$, while results outside this interval are published with probability $p(Z_i^*) = 1$. These values correspond to our estimates based on the economics lab experiments data of Camerer et al. (2016) discussed in Section 5.2 below. This publication policy reflects a preference for "significant results," where a two-sided z-test rejects the null hypothesis $\Theta^* = 0$ at the 5% level. This journal is ten times more likely to publish significant results than insignificant ones. This selectivity results in publication bias: published results, whose distribution is given by Lemma 1 above, tend to over-estimate the magnitude of the treatment effect. Published confidence intervals under-cover the true parameter value for small values of Θ and over-cover for somewhat larger values. This is demonstrated by Figure 1, which plots the median bias, $med(\hat{\Theta}_j|\Theta_j = \theta) - \theta$, of the usual estimator $\hat{\Theta}_j = Z_j$, as well as the coverage of the conventional 95% confidence interval $[Z_j - 1.96, Z_j + 1.96]$.

2.2 Alternative data generating processes

To clarify the implications of our model, we contrast it with two alternative data generating processes.

Observability The setup of Definition 1 assumes that we only observe the draws X^* for which $D = 1$. Alternative assumptions about observability might be appropriate, however, if additional information is 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.

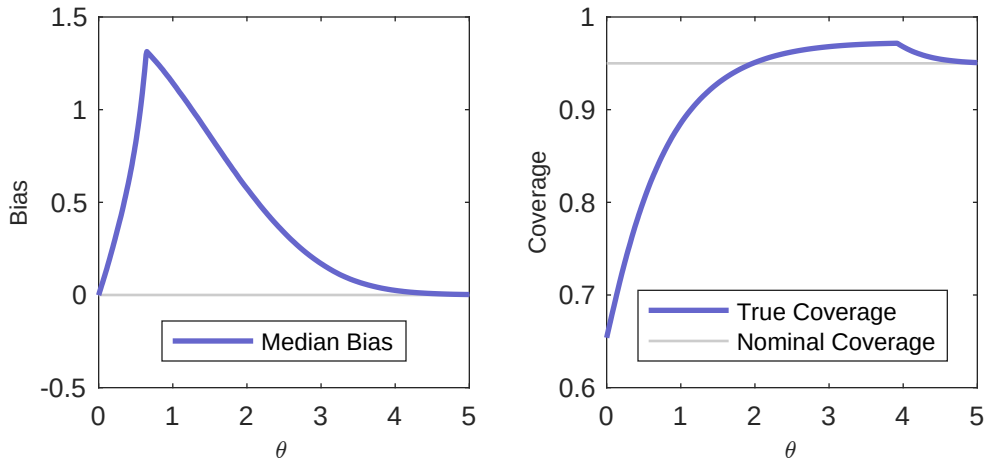

Figure 1: The left panel plots the median bias of the conventional estimator $\hat{\Theta}_j = Z_j$, while the right panel plots the true coverage of the conventional 95% confidence interval, both for $p(z) = .1 + .9 \cdot \mathbf{1}(|Z| > 1.96)$.

draws of (Y, D), where $Y = D \cdot X^*$. The corresponding censored likelihood is

$$f_{Y,D|\Theta^*}(x, d|\theta^*) = d \cdot p(x) \cdot f_{X^*|\Theta^*}(x|\theta) + (1 - d) \cdot (1 - E[D_i|\Theta_i^* = \theta^*]).$$

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

$$f_{X^*,D|\Theta^*}(x, d|\theta) = p(x)^d (1 - p(x))^{1-d} \cdot f_{X^*|\Theta^*}(x|\theta).$$

Even under these alternative observability assumptions, the truncated likelihood (1) arises as a limited information (conditional) likelihood, so identification and inference results based on this likelihood remain valid. Specifically, this likelihood conditions on publication decisions in the model with censoring, and on both publication decisions and unpublished results in the model with X^* observed. Thus, while additional information about the existence or content of unpublished studies might be used to gain additional insight, the results developed below continue to apply.

Manipulation of results Our analysis assumes that the distribution of the results X^* in latent studies given the true effects Θ^*, $f_{X^*|\Theta^*}$, is known. This implicitly restricts the scope for researchers to inflate the results of latent studies, cf. Brodeur et al. (2016). There are, however, many forms of manipulation or "p-hacking" (Simon-

sohn et al., 2014) which are accommodated by our model. In particular, if researchers conduct many independent analyses (where the results of each analysis follow known $f_{X^*|\Theta^*}$) but write up and submit only significant analyses, this is a special case of our model. More broadly, essentially any form of manipulation can be represented in a more general model where p depends on both X^* and Θ^*. This extension is discussed in Section 3.1.3 below.

3 Identifying selection

This section proposes two approaches for identifying $p(\cdot)$. The first uses systematic replication studies. By a "replication" we mean what Clemens (2015) terms a "reproduction," obtained by applying the same experimental protocol or analysis to a new sample from the same population as the original study. For each published X in a given set of studies, such replications provide an independent estimate X^r governed by the same parameter Θ as the original study. Under the assumption that selectivity operates only on X and not on X^r, we prove nonparameteric identification of $p(\cdot)$ up to scale. Under the additional assumption of normally distributed estimates we also establish identification of the latent distribution μ of true effects Θ^*.

The second approach considers meta-studies where there is variation across published studies in the standard deviation σ of normally distributed estimates X of Θ, where normality can again be understood as arising from the usual asymptotic approximations. Under the assumption that the standard deviation σ^* is independent of Θ^* in the population of latent studies, and that publication probabilities are a function of the z-statistic $Z^* = X^*/\sigma$ alone, we again show nonparametric identification of $p(\cdot)$ up to scale, as well as of μ.

Identification based on systematic replication studies is considered in Section 3.1. Identification based on meta-studies is considered in Section 3.2. In both sections, we return to our treatment effect example to illustrate results and develop intuition. Approaches in the literature, including meta-regressions and bunching of p-values, are discussed in the context of our assumptions in Section 3.3.

3.1 Systematic replication studies

We first consider the case of systematic replication studies, where both X^* and X^{*r} are drawn independently from the same distribution $f_{X^*|\Theta^*}$, conditional on Θ^*. In this setting the joint density $f_{X^*,X^{*r}}$, integrating out Θ^*, is symmetric in its arguments. Deviations from symmetry of f_{X,X^r} identify $p(\cdot)$ up to scale. We then extend this result in several ways, allowing different sample sizes for the original and replication studies as well as selection on Θ.

3.1.1 The symmetric baseline case

We extend the model in Definition 1 above to incorporate a conditionally independent replication draw X^{*r} which is observed whenever X^* is. The key implications of our model are symmetry of the joint distribution of (X^*, X^{*r}), and that selectivity of publication operates only on X^* and not on X^{*r}. The latter 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.

Definition 2 (Replication data generating process)
Consider the following data generating process for latent (unobserved) variables. $(\Theta_i^*, X_i^*, D_i, X_i^{*r},)$ *are jointly i.i.d. across* i*, with*

$$\begin{aligned}
\Theta_i^* &\sim \mu \\
X_i^*|\Theta_i^* &\sim f_{X^*|\Theta^*}(x|\Theta_i^*) \\
D_i|X_i^*, \Theta_i^* &\sim Ber(p(X_i^*)) \\
X_i^{*r}|D_i, X_i^*, \Theta_i^* &\sim f_{X^*|\Theta^*}(x|\Theta_i^*).
\end{aligned}$$

Let $I_0 = 0$*,* $I_j = \min\{i : \ D_i = 1, \ i > I_{j-1}\}$ *and* $\Theta_j = \Theta_{I_j}$*. We observe i.i.d. draws of*

$$(X_j, X_j^r) = (X_{I_j}^*, X_{I_j}^{*r}).$$

The next result extends Lemma 1 to derive the joint density of (X, X^r).

Lemma 2 (Replication Density)
Consider the setup of Definition 2. In this setup, the conditional density of (X, X^r)

given Θ *is*

$$\begin{aligned} f_{X,X^r|\Theta}(x, x^r|\theta) &= f_{X^*,X^{*r}|\Theta^*,D}(x, x^r|\theta, 1) \\ &= \frac{p(x)}{E[p(X_i^*)|\Theta_i^* = \theta]} f_{X^*|\Theta^*}(x|\theta)\, f_{X^*|\Theta^*}(x^r|\theta)\,. \end{aligned}$$

The marginal density of (X, X^r) *is*

$$f_{X,X^r}(x, x^r) = \frac{p(x)}{E[p(X_i^*)]} \int f_{X^*|\Theta^*}(x|\theta_i^*)\, f_{X^*|\Theta^*}(x^r|\theta_i^*)\, d\mu(\theta_i^*).$$

This lemma immediately implies that any asymmetries in the joint distribution of X, X^r must arise from the publication probability $p(\cdot)$. In particular,

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

whenever the denominators on either side are non-zero. Using this fact, we prove that $p(\cdot)$ is nonparametrically identified up to scale.

Theorem 1 (Nonparametric identification using replication experiments)
Consider the setup for replication experiments of Definition 2, and assume that the support of $f_{X^*,X^{*r}}$ *is of the form* $A \times A$ *for some measurable set* A*. In this setup* $p(\cdot)$ *is nonparametrically identified on* A *up to scale.*

Testable restrictions The density derived in Lemma 2 shows that the model of Definition 2 implies testable restrictions. Specifically, define $h(a, b) = \log(f_{X,X^r}(b, a)) - \log(f_{X,X^r}(a, b))$. By Lemma 2, $h(a, b) = \log(p(b)) - \log(p(a))$, and therefore

$$h(a, b) + h(b, c) + h(c, a) = 0$$

for any three values a, b, c. One could construct a nonparametric test of the model based on these restrictions and an estimate of f_{X,X^r}. In the applications below we opt for an alternative approach. We test restrictions on an identified model which nests the setup of Definition 2, detailed in Section 3.1.3 below.

Illustrative example (continued) To illustrate our identification approach using replication studies, we return to the illustrative numerical example introduced in

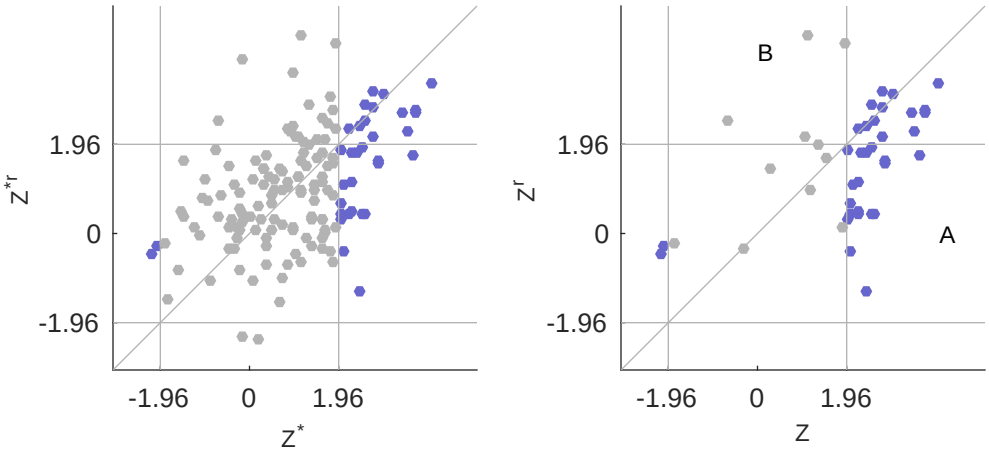

Figure 2: This figure illustrates the effect of selective publication in the replication experiments setting using simulated data, where selection is on statistical significance, as described in the text. The left panel shows the joint distribution of a random sample of latent estimates and replications; 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.

Section 2.1. In this setting, suppose that the true effect Θ^* is distributed $N(1,1)$ across latent studies. As before, assume that Z^* is $N(\Theta^*,1)$ distributed conditional on Θ^*, that $p(Z^*) = 1$ when $|Z^*| > 1.96$, and that $p(Z^*) = .1$ otherwise. Hence, results that are significantly different from zero at the 5% level based on a two-sided z-test are ten times as likely to be published as insignificant results.

This setting is illustrated in Figure 2. 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) which 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.

3.1.2 Generalizations and practical complications

In practice we need to modify the assumptions above to fit our applications, where the sample size for the replication often differs from that in the initial study, and the sign of the initial estimate X is normalized to be positive. We thus extend our identification results to accommodate these issues.

Differing variances To account for the impact of differing sample sizes on the distribution of X^{*r} relative to X^*, we need to be more specific about the form of these distributions. We assume that both X^* and X^{*r} are normally distributed unbiased estimates of the same latent parameter Θ^*, and that their variances are known. The assumption of approximate normality with known variance is already implicit in the inference procedures used in most applications. Since we require normality of only the final estimate from each study, rather than the underlying data, this assumption can be justified using standard asymptotic results even in settings with non-normal data, heteroskedasticity, clustering, or other features commonly encountered in practice. Normalizing the variance of the initial estimate to one yields the following setup, where we again denote the estimate by Z rather than X to emphasize the normalization of the variance.

$$\begin{aligned}
\Theta_i^* &\sim \mu \\
Z_i^* | \Theta_i^* &\sim N(\Theta_i^*, 1) \\
D_i | Z_i^*, \Theta_i^* &\sim Ber(p(Z_i^*)) \\
\sigma_i^* | Z_i^*, D_i, \Theta_i^* &\sim f_{\sigma|Z^*} \\
Z_i^{*r} | \sigma_i^*, Z_i^*, D_i, \Theta_i^* &\sim N(\Theta_i^*, \sigma_i^{*2})
\end{aligned} \tag{2}$$

We use σ to denote both the standard deviation as a random variable and the realized standard deviation. We again assume that results are published whenever $D_i = 1$, so that

$$f_{Z,Z^r,\sigma}(z, z^r, \sigma) = f_{Z^*,Z^{*r},\sigma^*|D}(z, z^r, \sigma|1).$$

Allowing the replication variance σ_i^* to differ from one takes us out of the symmetric framework of Definition 2. Display 2 also allows the possibility that the distribution of σ_i^* might depend on Z_i^*. Dependence of σ_i^* on Z_i^* is present, for example, 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_i^* is positively related to Z_i^*, but conditionally unrelated to Θ_i^*.

The following corollary states that identification carries over to this setting. The proof relies on the fact that we can recover the symmetric setting by (de)convolution of Z^r with normal noise, given Z and σ, which then allows us to apply Theorem 1. The assumption of normality further allows recovery of μ, the distribution of Θ^*.

Corollary 1
Consider the setup for replication experiments in display (2)*. Suppose we observe i.i.d. draws of* (Z, Z^r)*. In this setup* $p(\cdot)$ *is non-parametrically identified on* $\mathbb{R}$ *up to scale, and* μ *is identified as well.*

Normalized sign A further complication is that the sign of the estimates Z in our replication datasets is normalized to be positive, with the sign of Z^r adjusted accordingly. The following corollary shows that under this sign normalization identification of $p(\cdot)$ still holds, so long as $p(\cdot)$ is symmetric.

Corollary 2
Consider the setup for replication experiments of display (2)*. Assume additionally that* $p(\cdot)$ *is symmetric,* $p(z) = p(-z)$*, and that* $f_{\sigma|Z^*}(\sigma|z) = f_{\sigma|Z^*}(\sigma| - z)$ *for all* z*. Suppose that we observe i.i.d. draws of*

$$(W, W^r) = \text{sign}(Z) \cdot (Z, Z^r).$$

In this setup $p(\cdot)$ *is non-parametrically identified on* $\mathbb{R}$ *up to scale, and the distribution of* $|\Theta^*|$ *is identified as well.*

3.1.3 Selection depending on Θ^* given X^*

Selection of an empirical result X for publication might depend not only on X but also on other empirical findings reported in the same manuscript, or on unreported results obtained by the researcher. If that is the case, our assumption that publication decisions are independent of true effects conditional on reported results, $D \perp \Theta^*|X^*$, may fail. Allowing for a more general selection probability $p(X^*, \Theta^*)$, we can still identify $f_{X|\Theta}$, which is the key object for bias-corrected inference as discussed in

Section 4. Consider the following setup.

$$\begin{aligned}
\Theta_i^* &\sim f_{\Theta^*} \\
Z_i^*|\Theta_i^* &\sim N(\Theta_i^*, 1) \\
D_i|Z_i^*, \Theta_i^* &\sim Ber(p(Z_i^*, \Theta_i^*)) \\
\sigma_i^*|D_i, Z_i^*, \Theta_i^* &\sim f_{\sigma|Z^*} \\
Z_i^{*r}|\sigma_i^*, D_i, Z_i^*, \Theta_i^* &\sim N(\Theta_i^*, \sigma_i^2)
\end{aligned} \tag{3}$$

Assume again that results are published whenever $D_i = 1$. The assumption $D_i|Z_i^*, \Theta_i^* \sim Ber(p(Z_i^*, \Theta_i^*))$ is the key generalization relative to the setup considered before. This allows publication decisions to depend on both the reported estimate and the true effect, and allows a wide range of models for the publication process. In particular, this accommodates models where publication decisions depend on a variety of additional variables, including alternative specifications and robustness checks not reported in the replication dataset. Publication probabilities conditional on Z^* and Θ^* then implicitly average over these variables, resulting in additional dependence on Θ^*. For a simple example of this form, see Section C of the supplement.

Theorem 2
Consider the setup for replication experiments of display (3)*. In this setup $f_{Z|\Theta}$ is non-parametrically identified.*

The proof of Theorem 2 implies that the joint density $f_{Z,Z^r,\sigma,\Theta}$ is identified. Under the assumptions of display (3) the joint density of (Z, Z^r, σ, Θ) is

$$f_{Z,Z^r,\sigma,\Theta}(z, z^r, \sigma, \theta) = \frac{p(z,\theta)}{E[p(Z^*, \Theta^*)]} \varphi(z-\theta) \tfrac{1}{\sigma} \varphi\left(\tfrac{z^r-\theta}{\sigma}\right) f_{\sigma|Z^*}(\sigma|z) \tfrac{d\mu}{d\nu}(\theta),$$

where we use ν to denote a dominating measure on the support of Θ. Without further restrictions $p(z, \theta)$ is not identified; we can always divide $p(z, \theta)$ by some function $g(\theta)$ and multiply $\frac{d\mu}{d\nu}(\theta)$ by the same function to get an observationally equivalent model. Theorem 2 implies, however, that $p(z, \theta)$ is identified up to a normalization given θ, since

$$\frac{f_{Z|\Theta}(z,\theta)}{f_{Z^*|\Theta^*}(z,\theta)} = \frac{p(z,\theta)}{E[p(Z^*, \Theta^*)|\Theta^* = \theta]}.$$

We can for instance impose $\sup_z p(z, \theta) = 1$ for all θ to get an identified model. In our

applications we consider a parametric version of this model and test $p(z, \theta) = p(z)$ as a specification check on our baseline model.

3.2 Meta-studies

We next consider identification using meta-studies. Suppose that studies report normally distributed estimates X^* with mean Θ^* and standard deviation σ^*, and that selectivity of publication is based on the z-statistic $Z^* = X^*/\sigma^*$. The key identifying assumption is that Θ^* is statistically independent of σ^* across studies, so studies with larger sample sizes do not have systematically different estimands. Under this assumption, the distribution of X^* conditional on a larger value $\sigma^* = \sigma_1$ is equal to the convolution of normal noise of variance $\sigma_1^2 - \sigma_2^2$ with the distribution of X^* conditional on a smaller value $\sigma^* = \sigma_2$. Deviations from this equality for the observed distribution $f_{X|\sigma}$ identify $p(\cdot)$ up to scale.

Definition 3 (Meta-study data generating process)
Consider the following data generating process for latent (unobserved) variables. $(\sigma_i^*, \Theta_i^*, X_i^*, D_i)$ *are jointly i.i.d. across* i*, such that*

$$\begin{aligned}
\sigma_i^* &\sim \mu_\sigma \\
\Theta_i^* | \sigma_i^* &\sim \mu_\Theta \\
X_i^* | \Theta_i^*, \sigma_i^* &\sim N(\Theta_i^*, \sigma_i^{*2}) \\
D_i | X_i^*, \Theta_i^*, \sigma_i^* &\sim Ber(p(X_i^*/\sigma_i^*))
\end{aligned}$$

Let $I_0 = 0$*,* $I_j = \min\{i : \ D_i = 1, \ i > I_{j-1}\}$ *and* $\Theta_j = \Theta_{I_j}$*. We observe i.i.d. draws of*

$$(X_j, \sigma_j) = (X_{I_j}^*, \sigma_{I_j}^*).$$

Define $Z_i^* = \frac{X_i^*}{\sigma_i^*}$ *and* $Z_j = \frac{X_j}{\sigma_j}$*.*

A key object for identification of $p(\cdot)$ in this setting is the conditional density $f_{Z|\sigma}$.

Lemma 3 (Meta-study density)
Consider the setup of definition 3. The conditional density of Z *given* σ *is*

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

We build on Lemma 3 to prove our main identification result for the meta-studies setting. Lemma 3 implies that, for $\sigma_2 > \sigma_1$,

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

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, which would then immediately allow us to identify $p(\cdot)$. The proof of Theorem 3 builds on this idea, considering $\partial_\sigma \log(f_{Z|\sigma}(z|\sigma))$.

Theorem 3 (Nonparametric identification using meta-studies)
Consider the setup for experiments with independent variation in σ, described by Definition 3. Suppose that the support of σ contains an open interval. Then $p(\cdot)$ is identified up to scale, and μ is identified as well.

Illustrative example (continued) As before, assume that Θ^* is $N(1,1)$ distributed. Suppose further that σ^* is independent of Θ^* across latent studies, and that X^* is $N(\Theta^*, \sigma^*)$ distributed conditional on Θ^*, σ^*. Let $p(X^*/\sigma^*) = 1$ when $|X^*/\sigma^*| > 1.96$, $p(X^*/\sigma^*) = .1$ otherwise. Thus, results which differ significantly from zero at the 5% level are again ten times as likely to be published as insignificant results. This setting is illustrated in Figure 3. 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) which are published, where 90% of statistically insignificant draws are deleted.

Compare what happens for two different values of the standard deviation σ, marked by A and B in Figure 3. 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.

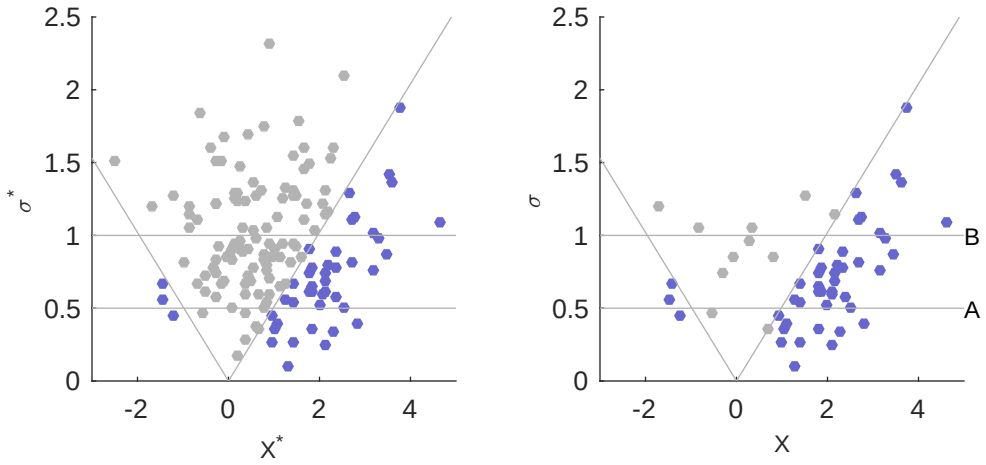

Figure 3: This figure illustrates the effect of selective publication in the meta-studies setting using simulated data, where selection is on statistical significance, as described in the text. The left panel shows a random sample of latent estimates; the right panel shows the subset of estimates which are published. Results which are significantly different from zero at the 5% level are plotted in blue, while insignificant results are plotted in grey.

Normalized sign In some of our applications the sign of the reported estimates X is again normalized to be positive. The following corollary shows that $p(\cdot)$ remains identified under this sign normalization provided it is symmetric in its argument.

Corollary 3

Consider the setup of Definition 3. Assume additionally that $p(\cdot)$ is symmetric, i.e., $p(x/\sigma) = p(-x/\sigma)$. Suppose that we observe i.i.d. draws of $(|X|, \sigma)$. In this setup $p(\cdot)$ is non-parametrically identified on $\mathbb{R}$ up to scale, and the distribution of $|\Theta^*|$ is identified as well.

3.3 Relation to approaches in the literature

Various approaches to detect selectivity and publication bias have been proposed in the literature. We briefly analyze some of the these approaches in our framework. First, we discuss to what extent we should expect the results of significance tests to "replicate" in a sense considered in the literature, and show that the probability of such replication may be low even in the absence of publication bias. Second, we discuss meta-regressions, and show that while they provide a valid test of the null of no selectivity under our meta-study assumptions, they are difficult to interpret under the alternative. Third, we consider approaches based on the distribution of p-values

or z-statistics, and analyze the extent to which bunching or discontinuities of this distribution provide evidence for selectivity or inflation of estimates.

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 suggests, however, 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 in display (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 $\sigma^* \equiv 1$. The probability that a result replicates in the sense described above is

$$\begin{aligned} &P(Z^{*r} \cdot sign\{Z^*\} > 1.96 || Z^*| > 1.96) \\ &\qquad = \frac{P(Z^{*r} < -1.96,\ Z^* < -1.96) + P(Z^{*r} > 1.96,\ Z^* > 1.96)}{P(Z^* < -1.96) + P(Z^* > 1.96)} \\ &\qquad\qquad = \frac{\int \left[\Phi(-1.96-\theta)^2 + \Phi(-1.96+\theta)^2\right] d\mu(\theta)}{\int \left[\Phi(-1.96-\theta) + \Phi(-1.96+\theta)\right] d\mu(\theta)}. \end{aligned}$$

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 $|\Theta^*| > M$ with probability one for some large M, then the probability of replication is approximately one. Thus, the probability that results replicate in this sense gives little indication of whether selective publication or some other source of bias for published research is present unless we either restrict the distribution of true effects or observe replication frequencies less than 0.025. Strengths and weaknesses of alternative measures of replication are discussed in Simonsohn (2015), Gilbert et al. (2016), and Patil and Peng (2016).

Meta-regressions A popular test for publication bias in meta-studies (cf. Card and Krueger, 1995; Egger et al., 1997) 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. The following lemma is immediate.

Lemma 4

Under the assumptions of Definition 3, if $p(\cdot)$ is constant then

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

As this lemma confirms, meta-regressions can be used to construct tests for the null of no publication bias. In particular, absent publication bias $\beta_0 = 0$ and $\gamma_1 = 0$, so tests for these null hypotheses allow us to test the hypothesis of no publication bias, though there are some forms of selectivity against which such tests have no power. As also noted in the previous literature, absent publication bias the coefficients β_1 and γ_0 recover the average of Θ^* in the population of latent studies. While these coefficients are sometimes interpreted as selection-corrected estimates of the mean effect across studies (cf. Doucouliagos and Stanley, 2009; Christensen and Miguel, 2016), 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 $E[\Theta^*]$.[1] To illustrate the resulting complications, we discuss a simple example with one-sided significance testing in Section B of the supplement.

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. For example, Simonsohn et al. (2014) analyze whether the distribution of p-values in a given literature is right- or left-skewed. Brodeur et al. (2016) compiled 50,000 test results from all papers published in the American Economic Review, the Quarterly Journal of Economics, and the Journal of Political Economy between 2005 and 2011, and analyze their distribution to draw conclusions about distortions in the research process.

Under our model, absent selectivity of the publication process the distribution f_Z is equal to f_{Z^*}. If we additionally assume that $Z^*|\Theta^* \sim N(\Theta^*, 1)$ and $\Theta^* \sim \mu$, this

[1] 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.

implies that

$$f_Z(z) = f_{Z^*}(z) = (\pi * \varphi)(z) = \int \varphi(z-\theta) d\mu(\theta).$$

This model has testable implications, and requires that the deconvolution of f_Z with a standard normal density φ yield a probability measure μ. This implies that the density f_{Z^*} is infinitely differentiable. If selectivity is present, by contrast, then

$$f_Z(z) = \frac{p(z)}{E[p(Z^*)]} \cdot f_{Z^*}(z),$$

and any discontinuity of $f_Z(z)$ (for instance at critical values such as $z = 1.96$) identifies a corresponding discontinuity of $p(z)$ and indicates the presence of selectivity:

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

If we impose that $p(\cdot)$ is a step function, for example, then this argument allows us to identify $p(\cdot)$ up to scale.

The density f_{Z^*} also 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 that in particular $f_{Z^*}(z) \leq \varphi(0)$ and $f''_{Z^*}(z) \geq \varphi''(0) = -\varphi(0)$ for all z. Spikes in the distribution of Z thus likewise indicate the presence of selectivity or inflation.

Unlike our model, which focuses on selection, Brodeur et al. (2016) are interested in potential inflation of test results by researchers, and in particular in non-monotonicities of f_Z which can not be explained by monotone publication probabilities $p(z)$ alone. They construct tests for such non-monotonicities based on parametrically estimated distributions f_{Z^*}.

4 Corrected inference

This section derives median unbiased estimators and valid confidence sets for scalar parameters θ assuming $p(\cdot)$ is known. The supplement extends these results to derive optimal estimators for scalar components of vector-valued θ, and analyzes Bayesian inference under selective publication. While our identification results in the last section relied on an empirical Bayes perspective, which assumed that Θ_i^* was drawn from some distribution μ, this section considers standard frequentist results which

hold conditional on Θ.

Selective publication reweights the distribution of X by $p(\cdot)$. To obtain valid estimators and confidence sets, we need to correct for this reweighting. To define these corrections denote the cdf for published results X given true effect Θ by $F_{X|\Theta}$. For $f_{X|\Theta}$, the density of published results derived in Lemma 1,

$$F_{X|\Theta}(x|\theta) = \int_{-\infty}^{x} f_{X|\Theta}(\tilde{x}|\theta)d\tilde{x} = \frac{1}{E[p(X^*)|\Theta^* = \theta]} \int_{-\infty}^{x} p(\tilde{x}) f_{X^*|\Theta^*}(\tilde{x}|\theta)d\tilde{x}.$$

For many distributions $f_{X^*|\Theta^*}$, and in particular in the leading normal case (see Lemma 5 below) this cdf is strictly decreasing in θ. Using this fact, we can adapt an approach previously applied by, among others, D. Andrews (1993) and Stock and Watson (1998) and invert the cdf as a function of θ to construct a quantile-unbiased estimator. In particular, if we define $\hat{\theta}_\alpha(x)$ as the solution to

$$F_{X|\Theta}\left(x|\hat{\theta}_\alpha(x)\right) = \alpha \in (0,1), \tag{4}$$

then $\hat{\theta}_\alpha(X)$ is an α-quantile unbiased estimator for θ.

Theorem 4
If for all x, $F_{X|\Theta}(x|\theta)$ is continuous and strictly decreasing in θ, tends to one as $\theta \to -\infty$, and tends to zero as $\theta \to \infty$, then $\hat{\theta}_\alpha(x)$ as defined in (4) exists, is unique, and is continuous and strictly increasing for all x. If, further, $F_{X|\Theta}(x|\theta)$ is continuous in x for all θ then $\hat{\theta}_\alpha(X)$ is α-quantile unbiased for θ under the truncated sampling setup of Definition 1,

$$P\left(\hat{\theta}_\alpha(X) \leq \theta | \Theta = \theta\right) = \alpha \text{ for all } \theta.$$

If $f_{X^*|\Theta^*}(x|\theta)$ is normal, as in our applications, then the assumptions of Theorem 4 hold whenever $p(x)$ is strictly positive for all x and almost everywhere continuous.

Lemma 5
If the distribution of latent draws X^* conditional on (Θ^*, σ^*) is $N(\Theta^*, \sigma^{*2})$,

$$f_{X^*|\Theta^*,\sigma^*}(x|\theta,\sigma) = \frac{1}{\sigma}\varphi\left(\frac{x-\theta}{\sigma}\right),$$

$p(x) > 0$ for all x, and $p(\cdot)$ is almost everywhere continuous, then the assumptions of

Theorem 4 are satisfied.

These results allow straightforward frequentist inference that corrects for selective publication. In particular, using Theorem 4 we can consider the median-unbiased estimator $\hat{\theta}_{\frac{1}{2}}(X)$ for θ, as well as the equal-tailed level $1-\alpha$ confidence interval

$$\left[\hat{\theta}_{\frac{\alpha}{2}}(X), \hat{\theta}_{1-\frac{\alpha}{2}}(X)\right].$$

This estimator and confidence set fully correct the bias and coverage distortions induced by selective publication. Other selection-corrected confidence intervals are also possible in this setting. For example, provided the density $f_{X^*|\Theta^*}(x|\theta)$ belongs to an exponential family one can form confidence intervals by inverting uniformly most powerful unbiased tests as in Fithian et al. (2014). Likewise, one can consider alternative estimators, such as the weighted average risk-minimizing unbiased estimators considered in Mueller and Wang (2015), or the MLE based on the truncated likelihood $f_{X|\Theta}$.

Illustrative example (continued) To illustrate these results, we return to the treatment effect example discussed above. Figure 4 plots the median unbiased estimator, as well as upper and lower 95% confidence bounds as a function of X for the same publication probability $p(\cdot)$ considered above. We see that the median unbiased estimator lies below the usual estimator $\hat{\theta} = X$ for small positive X but that the difference is eventually decreasing in X. The truncation-corrected confidence interval shown in Figure 4 has exactly correct coverage, is smaller than the usual interval for small X, wider for moderate values X, and essentially the same for $X \geq 5$.

Figure 4 provides useful guidance for readers of published papers interested in the magnitude of true effects. Suppose that the illustrative example is a reasonable approximation of how selection works in practice, as our empirical findings below suggest is the case for experimental economics. Then the following "rule of thumb" adjustments correspond roughly to median-unbiased estimates. (i) If reported effects are close to zero, or very far from zero (z-statistics bigger than 4), then these estimates can be taken at face value. (ii) In intermediate ranges, magnitudes should be adjusted downwards. A reported z-statistic of 1 should be taken to indicate an effect (relative to the standard error) of about 0.4. A reported z-statistic of 2 should be taken to indicate an effect of about 0.7, and a reported z-statistic of 3 should be taken to

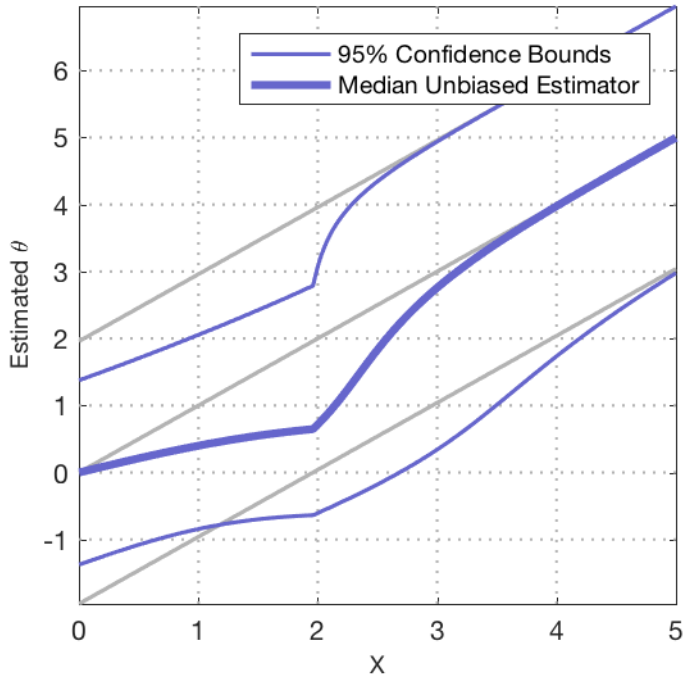

Figure 4: This figure plots frequentist 95% confidence bounds and the median unbiased estimator for the normal model where results that are significant at the 5% level are published with probability one, while insignificant results are published with probability 10%. The usual (uncorrected) estimator and confidence bounds are plotted in grey for comparison.

indicate an effect of about 2.75. Likewise, two-sided tests reject zero when z-statistics are larger than about 2.7 in absolute value.

We do not recommend adjusting publication standards to reflect these corrections. If publication probabilities in this example were based on more stringent critical values, for instance, then the corrections discussed above would become invalid. Instead, the purpose of these corrections is to allow readers of published research to draw valid inferences, taking the publication rule as given. The publication rule itself can then be chosen on other grounds, for example to maximize social welfare or provide incentives to researchers. We briefly discuss the question of optimal publication rules in the conclusion, as well as in Section I supplement.

In this example, our approach is closely related to the correction for selective publication proposed by McCrary et al. (2016). There, the authors propose conservative

tests derived under an extreme form of publication bias in which insignificant results are never published. If we consider testing the null hypothesis that θ is equal to zero, and calculate our equal-tailed confidence interval under the publication probability $p(\cdot)$ implied by the model of McCrary et al. (2016), then our confidence interval contains zero if and only if the test of McCrary et al. (2016) fails to reject.

5 Applications

This section uses the results developed above to estimate the degree of selectivity in several empirical literatures. Our nonparametric identification results imply identification of both $p(\cdot)$ and μ. The sample sizes in our applications are limited, however, so for estimation we specify parsimonious parametric models for both the conditional publication probability $p(\cdot)$ and the distribution μ of true effects across latent studies, which we then fit by maximum likelihood.

We begin by introducing the parametric specifications we consider. We then discuss our results based on the experimental economics replications of Camerer et al. (2016), the experimental psychology replications of Open Science Collaboration (2015), the minimum-wage meta-study of Wolfson and Belman (2015), and the deworming meta-study of Croke et al. (2016).

5.1 Likelihood and parametric specifications

5.1.1 Systematic replications

Under the replication setup of display (2), the marginal density of Z, Z^r, σ is

$$f_{Z,Z^r,\sigma}(z, z^r, \sigma) = \frac{p(z) \int \varphi(z-\theta) \cdot \frac{1}{\sigma}\varphi\left(\frac{z^r-\theta}{\sigma}\right) d\mu(\theta)}{\iint p(z') \cdot \varphi(z'-\theta) dz' d\mu(\theta)} f_{\sigma^*|Z^*}(\sigma|z). \tag{5}$$

Denoting the total number of observations by J, the joint likelihood of the observed sample $((z_1, z_1^r, \sigma_1), ..., (z_J, z_J^r, \sigma_J))$ is $\mathcal{L}(p, \mu) = \prod_{j=1}^{J} f_{Z,Z^r,\sigma}(z_j, z_j^r, \sigma_j)$. To fit a given model, we maximize this likelihood with respect to $p(\cdot)$ and μ. Since $f_{\sigma^*|Z^*}$ enters multiplicatively, it plays no role in maximum likelihood estimation of $p(\cdot)$ and μ. Hence, we drop this term from the likelihood used in estimation.

To model $p(\cdot)$, similar to Hedges (1992) we consider step functions

$$p(z) \propto \sum_{k=1}^{K} \beta_{p,k} \cdot \mathbf{1}\left(\zeta_{k-1} \leq z < \zeta_k\right),$$

where $-\infty = \zeta_0 < \zeta_1 < \ldots < \zeta_K = \infty$ are fixed cutoffs. Since $p(\cdot)$ is only identified up to scale, we normalize $\beta_{p,K} = 1$ and estimate $\beta_{p,1}, ..., \beta_{p,K-1}$. Thus $\beta_{p,k}$ can be interpreted as the publication probability for a latent study with Z^* between ζ_{k-1} and ζ_k, relative to a latent study with $Z^* \geq \zeta_{K-1}$. Finally, to model μ we assume that Θ^* is normally distributed with mean $\bar{\theta}$ and variance τ^2, and hence that

$$(Z^*, Z^{*r}, \sigma^*) \sim N\left(\begin{pmatrix}\bar{\theta}\\ \bar{\theta}\end{pmatrix}, \begin{pmatrix}\tau^2+1 & \tau^2\\ \tau^2 & \tau^2+\sigma^{*2}\end{pmatrix}\right) \cdot f_{\sigma^*|Z^*}$$

Under these assumptions the likelihood is available in closed form, simplifying estimation.

Sign normalization As noted in the discussion preceding Corollary 2, the sign of the initial estimate is normalized to be positive in both of our replication datasets.[2] In these applications, we thus follow the approach of Corollary 2 and assume that $p(\cdot)$ is symmetric around zero. We conduct estimation based on the normalized z-statistics $(W, W^r) = \text{sign}(Z) \cdot (Z, Z^r)$ using the marginal likelihood

$$f_{W,W^r,\sigma}(w, w^r, \sigma) = f_{Z,Z^r,\sigma}(w, w^r, \sigma) + f_{Z,Z^r,\sigma}(-w, -w^r, \sigma).$$

In this setting, Corollary 2 implies that $\beta_1, ...\beta_{k-1}$ and τ are identified, while $\bar{\theta}$ is identified up to sign. Identification of $\bar{\theta}$ is irregular at $\bar{\theta} = 0$, however, in the sense that the Fisher information for this parameter is zero, yielding nonstandard asymptotic behavior for the maximum likelihood estimator. Since a modal true effect of zero seems reasonable for the experimental economics and psychology settings we consider, we fix $\bar{\theta} = 0$ in these specifications. If we instead estimate this parameter, the MLE is exactly zero in all our sign-normalized applications.

[2]Each study in these datasets considers a different treatment, so the relative signs of effects across studies are arbitrary. Hence, setting the sign of the initial estimate in each study to be positive has the desirable effect of ensuring invariance to the sign normalization chosen by the authors of each study.

Specification test As noted in Section 3.1.3, replication data allows us to identify models where conditional publication probabilities may depend on both Z^* and Θ^*. We use these models to check our baseline specifications. Note that in principle any model that nests the null of no dependence of $p(\cdot)$ on Θ^* can be used to construct a valid test of this null. The specific model we consider determines where power is directed. In the supplement we introduce a model where publication decisions depend both on Z^* and on whether a 5% z-test based on an unobserved independent normal estimate rejects $\Theta^* = 0$. This yields a conditional publication probability of the form

$$p\left(z,\theta\right) = \sum_{k=1}^{K} \left(\beta_{p,k} + \gamma_{p,k} \cdot \Psi(\theta)\right) \cdot 1\left\{\zeta_{k-1} \leq z < \zeta_k\right\}, \tag{6}$$

for

$$\Psi(\theta) = \frac{\Phi(1.96-\theta) - \Phi(-1.96-\theta) - \Phi(1.96) + \Phi(-1.96)}{\Phi(1.96) + \Phi(-1.96)},$$

where Φ is the standard normal distribution function. This model implies that the publication probability is $\beta_{p,k}$ when Z^* is in bracket k and Θ^* is zero, while the publication probability is approximately $\beta_{p,k} + \gamma_{p,k}$ when Z^* is in bracket k and $|\Theta^*|$ is large. Setting $\gamma_p = 0$ recovers our baseline model, so testing $H_0 : \gamma_p = 0$ allows us to test our baseline specifications.

5.1.2 Meta-studies

In the meta-study context, the marginal likelihood of (X, σ) is

$$f_{X,\sigma}(x,\sigma) = \frac{p(\frac{x}{\sigma}) \cdot \int \varphi\left(\frac{x-\theta}{\sigma}\right) d\mu(\theta)}{\int p(\frac{x'}{\sigma}) \cdot \varphi\left(\frac{x'-\theta}{\sigma}\right) dx' d\mu(\theta)} f_\sigma^*(\sigma). \tag{7}$$

Again denoting the total number of observations by J, this yields joint likelihood $\mathcal{L}(p,\mu) = \prod_{j=1}^{J} f_{X,\sigma}(x_j,\sigma_j)$, which we again use to estimate $p(\cdot)$ and μ. As before, f_σ enters multiplicatively and need not be specified. Also as before, we consider step function specifications for $p(\cdot)$ and assume that Θ^* is $N(\bar{\theta},\tau^2)$ distributed, so

$$(X^*,\sigma^*) \sim N(\bar{\theta},\tau^2+\sigma^2) \cdot f_\sigma(\sigma^*).$$

Under these assumptions, the marginal likelihood (7) is again available in closed form.

Sign normalization In contexts where the sign of the initial estimate has been normalized to be positive, we follow the analog of the approach described above, restricting $p(\cdot)$ to be symmetric and conducting estimation based on $|X| = W \cdot \sigma$ and σ. Identification of $\bar{\theta}$ is again irregular at zero, and we fix it at $\bar{\theta} = 0$.

Note that meta-regressions, as discussed in section 3.3, do not yield a valid test of the null of no selectivity when using sign-normalized data. Regressions of $|X|$ on σ can have a non-zero slope even when $p(\cdot)$ is constant, and regressions of $|Z|$ on $1/\sigma$ can have a non-zero intercept. For this reason, we do not discuss meta-regression results in sign-normalized applications.

5.2 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 18 between-subject laboratory experiment papers published in the American Economic Review and Quarterly Journal of Economics between 2011 and 2014.[3] From each paper the most important statistically significant finding, as emphasized by the original authors, was selected for replication. Further details on the selection and replication of results can be found in Camerer et al. (2016), while details of 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 reasonable to think 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. At the same time, however, the selection of which result to replicate within each paper changes the interpretation of $p(\cdot)$, which has to be interpreted as the probability that a result was published *and* selected for replication.

[3]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 American Economic Review and Quarterly Journal of Economics from the relevant period, and confirmed that no studies were excluded due to this restriction.

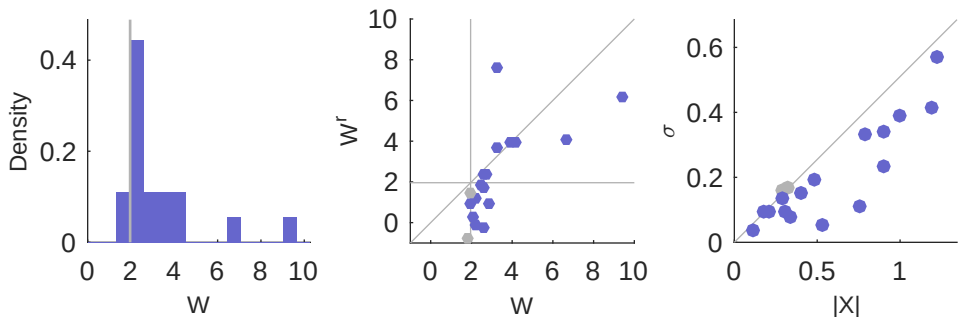

Figure 5: The left panel shows a binned density plot for the normalized z-statistics $W = |X|/\sigma$ using data from Camerer et al. (2016). The grey line marks $W = 1.96$. The middle panel plots the z-statistics W from the initial study against the estimate W^r from the replication study. The grey lines 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 $|X|/\sigma = 1.96$.

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 of a large jump in the density $f_W(\cdot)$ at the cutoff 1.96, and thus of a corresponding jump of the publication probability $p(\cdot)$ at the same cutoff; cf. the discussion in Section 3.3. Such a jump will be confirmed by the estimates from 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 in the replication data of Camerer et al. (2016), using the same conventions as in Figure 2. To estimate the degree of selection in these data we consider the model

$$\Theta^* \sim N(0, \tau^2), \quad p(Z) \propto \begin{cases} \beta_p & |Z| < 1.96 \\ 1 & |Z| \geq 1.96, \end{cases}$$

as described above. This assumes that the true effect Θ^* is mean-zero normal across latent studies, while allowing 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

Replication	
τ	β_p
2.354	0.100
(0.750)	(0.091)

Meta-study	
$\tilde{\tau}$	β_p
0.299	0.045
(0.073)	(0.045)

Table 1: Selection estimates from lab experiments in economics, with robust standard errors in parentheses. The left panel reports estimates from 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 0.1 implies that results which are insignificant at the 5% level are 10% as likely to be published as significant results. The parameters τ and $\tilde{\tau}$ are not comparable.

significant at the 5% level. These estimates therefore imply that significant results are ten times more likely to be published than insignificant results. This is the ratio we have assumed for our running example throughout this paper. Moreover, we strongly reject the hypothesis of no selectivity, $H_0 : \beta_p = 1$.

A score test of the null hypothesis $\gamma_p = 0$ in the model (6) where $p(\cdot)$ is allowed to depend on $|\Theta^*|$ yields a p-value of 0.71. We thus find no evidence that the assumption $D|Z^*, \Theta^* = p(Z^*)$ imposed in our baseline model is violated.

Results from meta-study specifications While the 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

$$\Theta^* \sim N(0, \tilde{\tau}^2), \quad p(X/\sigma) \propto \begin{cases} \beta_p & |X/\sigma| < 1.96 \\ 1 & |X/\sigma| \geq 1.96, \end{cases}$$

noting that Θ^* is now the mean of X^*, rather than Z^*, and thus that the interpretation of $\tilde{\tau}$ differs from that of τ 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, note that we estimate a similar degree of selectivity in the two specifications. Indeed, we cannot reject the hypothesis that β_p is the same in the two specifications at standard significance levels. 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 $\beta_p = .1$. 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. 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, ten 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.[4]

5.3 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, 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.

[4]Note that these adjusted confidence sets are based on the point estimate $\hat{\beta}_p$ and do not account for uncertainty in this estimate. To obtain valid confidence sets accounting for this uncertainty, one could consider Bonferroni-corrected versions of these adjusted confidence sets. However, such corrections would only widen the adjusted confidence sets, and so increase the discrepancy in significance between the adjusted and unadjusted results.

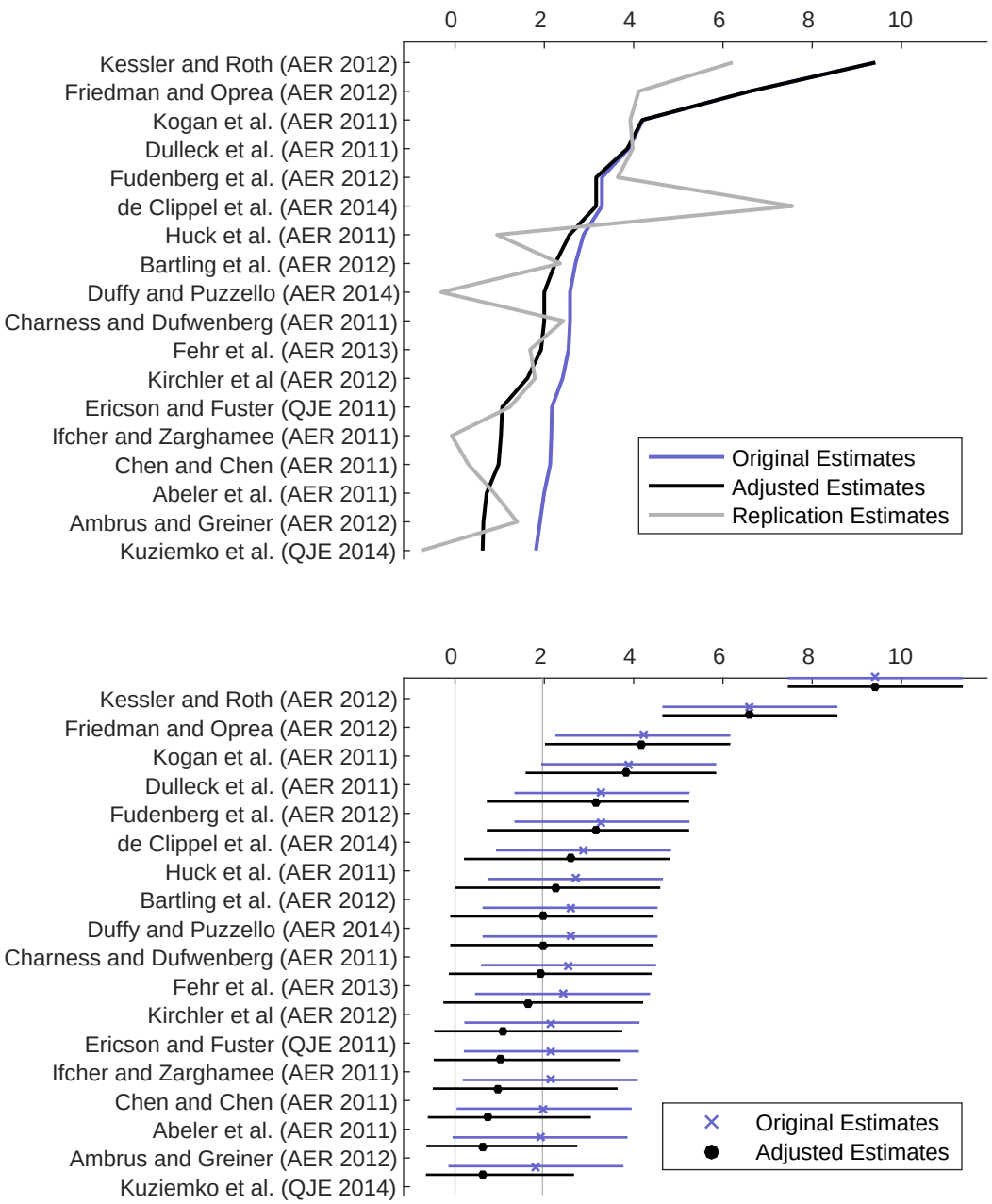

Figure 6: The top panel plots the estimates W and W^r from the original and replication studies in Camerer et al. (2016), along with the median unbiased estimate $\hat{\theta}_{\frac{1}{2}}$ based on the estimated selection model and the original estimate. The bottom panel plots the original estimate and 95% confidence interval, as well as the median unbiased estimate and adjusted 95% confidence interval $\left[\hat{\theta}_{0.025}(W), \hat{\theta}_{0.975}(W)\right]$ based on the estimated selection model.

As with the economics replications above, the systematic selection of results for replication in Open Science Collaboration (2015) is an advantage from our perspective. The fact that not all available studies were selected for replication raises the possibility of selection along this margin, but the fact that 100 of the 158 available studies were replicated limits the potential severity of selection here. Likewise, the widely followed default of replicating the final result within each study helps address concerns about the selection of which result to replicate within each paper.

A complication in this setting 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, 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.

A further complication arises from the critique of Gilbert et al. (2016), who argue that the protocols in some of the Open Science Collaboration (2015) replications differed substantially from the initial studies. To explore robustness with respect to this critique, in the supplement we report results from further restricting the sample to the subset of replications which used protocols approved by the original authors, and find roughly similar estimates, though the estimated degree of selection is smaller. For further discussion, as well other details of our analysis, please see the supplement.

Histogram Consider now the distribution of originally published estimates W, shown by the histogram in the left panel of Figure 7. This histogram is suggestive of 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 will again be confirmed by the estimates from both our replication and meta-study approaches.

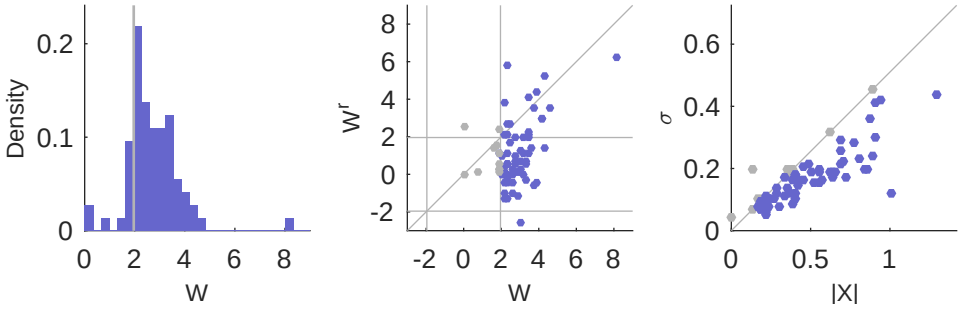

Figure 7: The left panel shows a binned density plot for the normalized z-statistics $W = |X|/\sigma$ using data from Open Science Collaboration (2015). The grey line marks $W = 1.96$. The middle panel plots the z-statistics W from the initial study against the estimate W^r from the replication study. 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 $|X|/\sigma = 1.96$.

Replication		
τ	$\beta_{p,1}$	$\beta_{p,2}$
1.252	0.021	0.294
(0.195)	(0.012)	(0.128)

Meta-study		
$\tilde{\tau}$	$\beta_{p,1}$	$\beta_{p,2}$
0.252	0.025	0.375
(0.041)	(0.015)	(0.166)

Table 2: Selection estimates from lab experiments in psychology, with robust standard errors in parentheses. The left panel reports estimates from 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 τ and $\tilde{\tau}$ are not comparable.

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). We fit the model

$$\Theta^* \sim N(0, \tau^2), \quad p(Z) \propto \begin{cases} \beta_{p,1} & |Z| < 1.64 \\ \beta_{p,2} & 1.64 \leq |Z| < 1.96 \\ 1 & |Z| \geq 1.96. \end{cases}$$

This model again assumes that the true effect Θ^* is mean-zero normal 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 almost fifty times more likely to be published than results that are insignificant at the 10% level, and over three 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.

A score test of the null hypothesis $\gamma_p = 0$ in the model where $p(\cdot)$ is allowed to depend on $|\Theta^*|$ yields a p-value of 0.3. Thus, we again find no evidence that the assumption $D|Z^*, \Theta^* = p(Z^*)$ imposed in our baseline model is violated.

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 quite similar.

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 the fit is worse for studies with larger original z-statistics. Our adjustments again dramatically change the number of significant results, with 62 of the 73 original 95% confidence sets excluding zero, and only 21 of the adjusted confidence sets (not displayed) doing the same.

5.4 Effect of minimum wage on employment

Our third 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) consider analyses of the effect of minimum wages on employment which use US data and were published or circulated as working papers after the year 2000. They collect estimates from all studies fitting their criteria which 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).

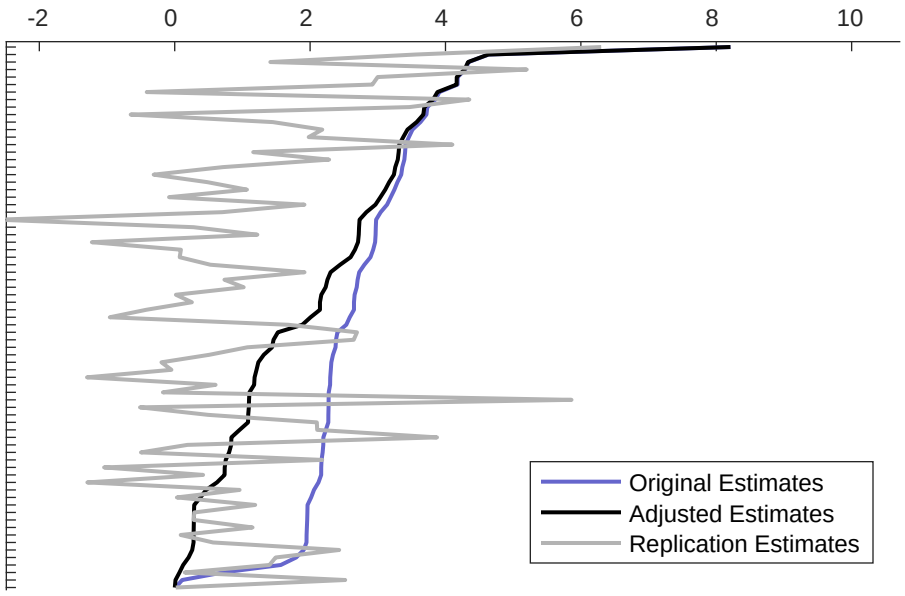

Figure 8: This figure plots the estimates W and W^r from the original and replication studies in Open Science Collaboration (2015), along with the median unbiased estimate $\hat{\theta}_{\frac{1}{2}}$ based on the estimated selection model and the original estimate.

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

Multiple estimates per study A complication arises in this application, relative to those considered so far, due to the presence of multiple estimates per study. Moreover, it is difficult to argue that a given estimate in each of these studies constitutes the "main" estimate, so restricting attention to a single estimate per study seems arbitrary. This raises issues for both inference and identification.

For inference, it is implausible that estimate standard-error pairs X_j, σ_j 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

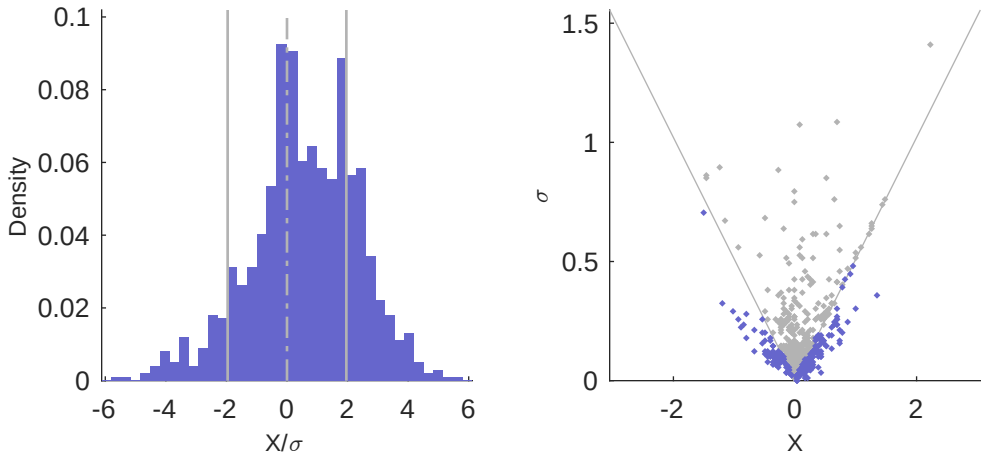

Figure 9: The left panel shows a binned density plot for the z-statistics X/σ in the Wolfson and Belman (2015) data. The solid grey lines mark $|X|/\sigma = 1.96$, while the dash-dotted grey line marks $X/\sigma = 0$. The right panel plots the estimate X against its standard error σ. The grey lines mark $|X|/\sigma = 1.96$.

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 X_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 X_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_i^* within each study are statistically independent (relative to the population of estimands in the literature under consideration). As discussed in Section 3.1.3, however, if these assumptions fail our model is misspecified.

Histogram Consider first the distribution of 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 of 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.

As a first check, we run meta-regressions as discussed in section 3.3, clustering standard errors at the study-level. A regression of X on σ yields a slope of 0.406 with a standard error of 0.369. A regression of Z on $1/\sigma$ yields an intercept of 0.343 with a standard error of 0.281. Both of these estimates are indicative of selection favoring results findings a negative effect of minimum wages on employment, but neither allows us to reject the null of no selection at conventional significance levels.

We next consider the model

$$\Theta^* \sim N(\bar{\theta}, \tilde{\tau}^2), \quad p(X/\sigma) \propto \begin{cases} \beta_{p,1} & X/\sigma < -1.96 \\ \beta_{p,2} & -1.96 \leq X/\sigma < 0 \\ \beta_{p,3} & 0 \leq X/\sigma < 1.96 \\ 1 & X/\sigma \geq 1.96. \end{cases}$$

Unlike in our previous applications, we allow the probability of publication to depend on the sign of the z-statistic X/σ 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 or negative effect of the minimum wage on employment. Our estimates based on these data are reported in Table 3, where we find that publication probabilities are monotonically increasing in Z. In particular, recalling that positive estimates X indicate a negative effect of the minimum wage on employment, our estimates suggest that studies that find a negative and significant effect of the minimum wage on employment at the 5% level are over four times more likely to published than studies that find a positive and significant effect, over twice as likely to be published as studies that find a positive but insignificant effect, and over 35% more likely to be published than estimates that which find a negative but insignificant effect.

These results are consistent with the meta-analysis results 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.

Since the studies in this application estimate related parameters, it is also interest-

$\bar{\theta}$	$\tilde{\tau}$	$\beta_{p,1}$	$\beta_{p,2}$	$\beta_{p,3}$
-0.024	0.122	0.225	0.424	0.738
(0.053)	(0.038)	(0.118)	(0.207)	(0.291)

Table 3: Meta-study estimates from minimum wage data, with clustered standard errors in parentheses. Publication probabilities β_p measured relative to omitted category of estimates positive and significant at 5% level.

ing to consider the estimate $\bar{\theta}$ for the mean effect in the population of latent estimates. The point estimate suggests that the average latent study finds a small positive effect of the minimum wage on employment, though the estimated $\bar{\theta}$ is quite small relative to both its standard error and the estimated standard deviation τ across specifications. This contrasts with the "naive" average effect we would get by ignoring selectivity and estimating our model subject to the constraint $\beta_p = (1, \ldots, 1)$. This yields a "naive" estimate for $\bar{\theta}$ of .038, with a standard error of .025, suggesting a negative average estimate of the effect of minimum wages on employment. To link either of these estimates to the "true" effect of minimum wages on employment, however, we would then need to take a stand on the credibility of the underlying studies.

5.5 Deworming meta-study

Our final application is to data from the recent meta-study Croke et al. (2016) on the effect of mass drug administration for deworming on child body weight. They collect results from randomized controlled trials which report child body weight as an outcome, and focus on intent-to-treat estimates from the longest follow-up reported in each study. They include all studies identified by the previous review of Taylor-Robinson et al. (2015), as well as additional trials identified by Welch et al. (2017). They then extract estimates as described in Croke et al. (2016) and obtain a final sample of 22 estimates drawn from 20 studies, which we take as the basis for our analysis. For further discussion of sample construction, see Taylor-Robinson et al. (2015), Croke et al. (2016), and Welch et al. (2017). To account for the presence of multiple estimates in some studies, we again cluster by study.

Histogram Consider first the distribution of the normalized estimates Z, shown by the histogram in the left panel of Figure 10. Given the small sample size of 22 esti-

mates, this histogram should not be interpreted too strongly. That said, the density of Z appears to jump up at 0, which suggests selection toward positive estimates.

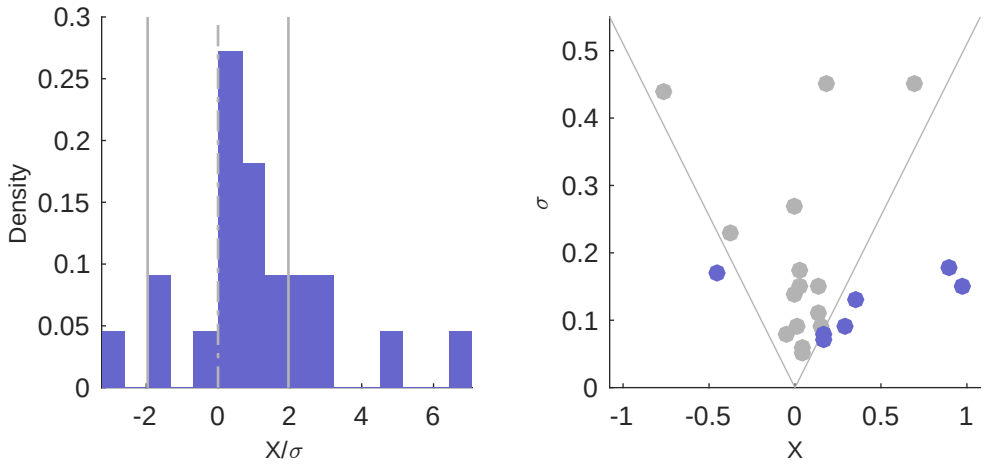

Figure 10: The left panel shows a binned density plot for the z-statistics X/σ in the Croke et al. (2016) data. The solid grey lines mark $|X|/\sigma = 1.96$, while the dash-dotted grey line marks $X/\sigma = 0$. The right panel plots the estimate X against its standard error σ. The grey lines mark $|X|/\sigma = 1.96$.

Results from meta-study specifications The right panel of Figure 10 plots the joint distribution of X, the estimated intent to treat effect of mass deworming on child weight, along with the standard error σ in the Croke et al. (2016) data.

As a first check, we again run meta-regressions as discussed in Section 3.3, clustering standard errors by study. A regression of X on σ yields a slope of -0.296 with a standard error of 0.917. A regression of Z on $1/\sigma$ yields an intercept of 0.481 with a standard error of 0.889. Neither of these estimates allows rejection of the null of no selection at conventional significance levels.

We next consider the model

$$\Theta^* \sim N(\bar{\theta}, \tilde{\tau}^2), \quad p(X/\sigma) \propto \begin{cases} \beta_p & |X/\sigma| < -1.96 \\ 1 & |X/\sigma| \geq 1.96, \end{cases}$$

where we constrain the function $p(\cdot)$ to be symmetric to limit the number of free parameters, which is important since we have only 22 observations. Fitting this model yields the estimates reported in Table 4. The point estimates here suggest

that statistically significant results are less likely to be included in the meta-study of Croke et al. (2016) than are insignificant results.

However, the standard errors are quite large, and the difference in publication (inclusion) probabilities between significant and insignificant results is itself not significant at conventional levels, so there is no basis for drawing a firm conclusion here. Likewise, the estimated $\bar{\theta}$ suggests a positive average effect in the population, but is not significantly different from zero at conventional levels.

In the supplement we report results based on alternative specifications which allow the function $p(\cdot)$ to be asymmetric. These specifications generate significant estimates of selection, and in particular suggest selection against negative estimates, but we hesitate to interpret these results due to the small sample size and multiple specifications considered.

Our findings here are potentially relevant in the context of the controversial debate surrounding mass deworming; see for instance Clemens and Sandefur (2015). The point estimates for our baseline specification suggest that insignificant results have a higher likelihood of being included in Croke et al. (2016) relative to significant ones. In light of the large standard errors and limited robustness to changing the specification of $p(\cdot)$, however, these findings should not be interpreted too strongly.

$\bar{\theta}$	$\tilde{\tau}$	β_p
0.190	0.343	2.514
(0.120)	(0.128)	(1.872)

Table 4: Meta-study estimates from deworming data, with robust standard errors in parentheses. Publication probabilities β_p measured relative to omitted category of studies significant at 5% level.

6 Conclusion

This paper contributes to the literature in three ways. First, we provide nonparametric identification results for selectivity (in particular, the conditional publication probability) as a function of the empirical findings of a study. Second, we provide methods to calculate bias-corrected estimators and confidence sets when the form of selectivity is known. Third, we apply the proposed methods to several literatures, documenting the varying scale and kind of selectivity.

Implications for applied researchers What can applied 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 in this literature, and to apply appropriate corrections to individual estimates, tests, and confidence sets. We will provide code on our webpages which implements the proposed methods for a flexible family of selection models.[5]

Second, when reading empirical research, readers may wish to apply some "rule of thumb" corrections to the published point estimates and confidence sets. Based on our finding that publication probabilities increase by a factor of 10 for experimental papers when exceeding the 5% significance threshold, the following corrections would be appropriate (cf. Figure 4 in Section 4): If reported effects are close to zero, or very far from zero (z-statistic bigger than 4), then these estimates can be taken at face value. In intermediate ranges, magnitudes should be adjusted downwards, so that for instance a reported z-statistic of 2 should be taken to indicate an effect (relative to the standard error) of about 0.7.

It should be emphasized that we do not advocate using more stringent critical values in the publication process, in a possible effort to obtain correct size control. If more stringent values were to be systematically applied, 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. This might for instance be achieved by instituting some form of result-blind 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. Consider for instance a setting where policy decisions are made based on published findings, policy makers have a limited capacity to read publications, and journal editors maximize the same social welfare function as policy makers. In a stylized model of such a setting, detailed in the supplement, we show that expected social welfare is maximized by publishing the results which allow policy makers to update the most relative to their

[5]In progress.

prior beliefs. The corresponding publication rule favors the publication of surprising findings, thus violating non-selectivity. This stylized model omits many features of practical interest, however, and a more general theory of optimal publication is of considerable interest for future research.

References

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

Baricz, Á. (2008). Mills' ratio: Monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications*, 340(2):1362–1370.

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.

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

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

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. and Sandefur, J. (2015). Mapping the worm wars: What the public should take away from the scientific debate about mass deworming.

Clemens, M. A. (2015). The meaning of failed replications: a review and proposal. *Journal of Economic Surveys*, Forthcoming.

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.

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.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

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

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

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

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

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).

Iyengar, S. and Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, pages 109–117.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

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

Mueller, U. and Wang, Y. (2015). Nearly weighted risk minimal unbiased estimation. Unpublished Manuscript.

Murphy, G. M. (2011). *Ordinary differential equations and their solutions*. Courier Corporation.

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.

Pfanzagl, J. (1994). *Parametric Statistical Theory*. De Gruyter.

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

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. 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).

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.

Taylor-Robinson, D. C., Maayan, N., Soares-Weiser, K., Donegan, S., and Garner, P. (2015). Cochrane database of systematic reviews. 7.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Welch, V. A., Ghogomu, E., Hossain, A., Awasthi, S., Bhutta, Z. A., Cumberbatch, C., Fletcher, R., McGowan, J., Krishnaratne, S., Kristjansson, E., Sohani, S., Suresh, S., Tugwell, P., White, H., and Wells, G. A. (2017). Mass deworming to improve developmental health and wellbeing of children in low-income and middle-income countries: a systematic review and network meta-analysis. *The Lancet Global Health*, 5(1):e40–e50.

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

Yekutieli, D. (2012). Adjusted bayesian inference for selected parameters. *Journal of the Royal Statistical Society Series B*, 74(3):515–541.

Supplement to the paper

Identification of and correction for publication bias

Isaiah Andrews Maximilian Kasy

March 22, 2017

This appendix contains proofs and supplementary results for the paper "Identification of and correction for publication bias." Section A collects proofs for the results stated in the main text. Section B considers the behavior of meta-regression coefficients, discussed in Section 3.3 of the main text, in a simple example. Section C states a simple model for selection on both Z^* and a latent variable V^*, and derives the conditional publication probability $p(z, \theta)$ stated in Section 5.1.1 of the main text. Section D provides details on the empirical applications discussed in main text, while Section E reports additional results. Section F provides corrected inference plots, analogous to Figure 4 of the main text, based on our psychology lab experiment, minimum wage, and deworming applications. Section G generalizes the inference results discussed in the main text to multivariate normal settings, while Section H discusses the effect of selective publication on Bayesian inference. Finally, Section I discusses optimal selection in a stylized model.

A Proofs

Proof of Lemma 1: By construction,

$$\begin{aligned} f_{X|\Theta}(x|\theta) &= f_{X^*|\Theta,D}(x|\theta, 1) \\ &= \frac{P(D_i = 1|X_i^* = x, \Theta_i^* = \theta)}{P(D_i = 1|\Theta_i^* = \theta)} \cdot f_{X^*|\Theta^*}(x|\theta) \\ &= \frac{p(x)}{E\left[p\left(X_i^*\right)|\Theta_i^* = \theta\right]} \cdot f_{X^*|\Theta^*}(x|\theta). \end{aligned}$$

□

Proof of Lemma 2: The conditional density follows by the same argument used to derive the truncated likelihood in Lemma 1. As for the marginal density, by construction,

$$\begin{aligned} f_{X,X^r}(x,x^r) &= f_{X^*,X^{*r}|D_i}(x,x^r|d=1) \\ &= \frac{P(D_i=1|X_i^*=x,X_i^{*r}=x^r)}{P(D_i=1)} \cdot f_{X^*,X^{*r}}(x,x^r) \\ &= \frac{p(x)}{E[p(X_i^*)]} f_{X^*,X^{*r}}(x,x^r), \end{aligned}$$

and, since $X_i^* \perp X^{*r}|\Theta_i^*$,

$$f_{X^*,X^{*r}}(x,x^r) = \int f_{X^*|\Theta^*}(x|\theta_i^*)\, f_{X^*|\Theta^*}(x^r|\theta_i^*)\, d\mu(\Theta_i^*).$$

□

Proof of Theorem 1: The marginal likelihood f_{X,X^r} derived in Lemma 2 satisfies

$$f_{X,X^r}(a,b) \cdot p(b) = f_{X,X^r}(b,a) \cdot p(a)$$

for all a, b.

Let $(a,b) \in A \times A$ be any point such that $f_{X,X^r}(a,b) > 0$, so that in particular $p(a) > 0$. By the assumptions on the support of $f_{X^*,X^{*r}}$ and the data generating process this implies that $f_{X,X^r}(a,c) > 0$ for all $c \in A$.

This in turn implies that

$$p(c) = p(a) \cdot \frac{f_{X,X^r}(c,a)}{f_{X,X^r}(a,c)}$$

for all $c \in A$, where $p(a)$ is the only unknown on the right hand side. We thus find that $p(x)$ is identified up to scale. □

Proof of Corollary 1: In the case where $\sigma \equiv 1$, this is a special case of Theorem 1, and the claim immediately follows. (Note that (Z^*, Z^{*r}) has full support $\mathbb{R}^2$.) We will show that we can reduce the case where $\sigma \not\equiv 1$ to this special case. Let $\tilde{Z}$ be such that

$$\tilde{Z}_i|Z_i^*, D_i, \Theta_i^* \sim N(\Theta_i^*, 1).$$

If $f_{\tilde{Z}|Z}$ is identified, we are done. Note that

$$f_{\tilde{Z}|Z} = f_{\Theta|Z} * \varphi,$$

and

$$f_{Z^r|Z,\sigma} = f_{\Theta|Z,\sigma} * \varphi_\sigma.$$

Based on the last equation, $f_{\Theta|Z,\sigma}$ is identified using deconvolution (this is a standard result; see for instance Wasserman 2006, Chapter 10.1). We then get

$$f_{\Theta|Z} = \int f_{\Theta|Z,\sigma} f_{\sigma|Z} d\sigma,$$

and identification of $p(\cdot)$ follows.

To show identification of μ, note that knowledge of $p(\cdot)$ up to scale allows us to recover the density f_{Z^*} via

$$f_{Z^*}(z) = \frac{E[p(Z^*)]}{p(z)} f_Z(z).$$

Deconvolution then identifies μ, since $f_{Z^*} = \mu * \varphi$. □

Proof of Corollary 2: Let $S_i^* = \pm 1$ with probability 0.5, independently of $(Z_i^*, Z_i^{*r}, \sigma_i^*, \Theta_i^*)$, and $S_j = S_{I_j}^*$. Define

$$(V, V^r) = S \cdot (W, W^r).$$

We show that (V, V^r) satisfies the assumptions of Corollary 1, from which the claim then follows.

Define $\tilde{S}^* = S^* \cdot \text{sign}(Z^*)$, so that $(V, V^r) = \tilde{S} \cdot (Z, Z^r)$, and define $\tilde{\Theta}^* = \tilde{S}^* \cdot \Theta^*$. Since $\tilde{S}$ is independent of (Z, Z^r, σ, Θ), we get

$$\tilde{\Theta}^* \sim \tilde{\mu} = \tfrac{1}{2}(\mu_{\Theta^*} + \mu_{-\Theta^*})$$

and

$$f_{V,V^r,\sigma}(v, v^r, \sigma) = p(v) \cdot f_{\sigma|Z^*}(\sigma|v) \cdot \frac{\int \varphi(v-\theta) \cdot \frac{1}{\sigma}\varphi\left(\frac{v^r-\theta}{\sigma}\right) d\tilde{\mu}(\theta)}{\int p(v') \cdot \varphi(v'-\theta) dv' d\tilde{\mu}(\theta)}.$$

This has the exact same form as the density of (Z, Z^r, σ) under the symmetric measure

$\tilde{\mu}$. The claim follows, since identification of $\tilde{\mu}$ implies identification of the distribution of $|\Theta^*|$. $\square$

Proof of Theorem 2: Under the setup considered, using the implied conditional independence assumptions we get

$$\begin{aligned} f_{Z^r|Z,\sigma}(z^r, z, \sigma) &= \int f_{Z^{*r}|\sigma,Z^*,D,\Theta^*}(z^r|\sigma, z, 1, \theta) f_{\Theta^*|\sigma,Z^*,D}(\theta|\sigma, z, 1) d\theta \\ &= \int \varphi_\sigma(z^r - \theta) f_{\Theta^*|Z^*,D}(\theta|z, 1) d\theta \\ &= (f_{\Theta|Z} * \varphi_\sigma)(z^r|z). \end{aligned}$$

By deconvolution, this immediately implies that we can identify $f_{\Theta|Z}$. Since f_Z is directly identified, Bayes' rule yields the desired result via

$$f_{Z|\Theta}(z|\theta) = \frac{f_{\Theta|Z}(\theta|z) \cdot f_Z(z)}{\int f_{\Theta|Z}(\theta|z') \cdot f_Z(z') dz'}.$$

$\square$

Proof of Theorem 3: Assume w.l.o.g. that $\sigma = 1$ lies in the interior of the support of σ, and let

$$h(z) = f_{Z^*|\sigma^*}(z|1).$$

If $h(\cdot)$ is identified, then so are $p(\cdot)$ and μ. We will show that $h(\cdot)$ is identified. Once $h(z)$ is identified, we get $p(z)$ as before, since the truncated conditional density of Z is given by

$$f_{Z|\sigma}(z|\sigma) = \frac{p(z)}{E[p(Z^*)|\sigma]} f_{Z^*|\sigma^*}(z|\sigma), \tag{8}$$

and thus

$$p(z) = const. \cdot \frac{f_{Z|\sigma}(z|1)}{h(z)}.$$

We can further identify μ by deconvolution given h, since $h = \mu * \varphi$.

A second order ODE for $h(\cdot)$. Let $\pi = 1/\sigma$ be the precision of an estimate. Differentiating the log of expression (8) for the truncated density at $\pi = 1$ yields

$$g(z) = \partial_\pi \log f_{Z|\sigma}(z|1) = C_1 + \partial_\pi \log f_{Z^*|\sigma^*}(z|1) \tag{9}$$

for a constant C_1. Note how, as we differentiate $\log f_{Z|\sigma}(z|1)$ with respect to π at a given value z, the term $p(z)$ drops out of the resulting equation. The function g is identified under our assumptions.

Recall now that the definition of the standard normal density implies $\varphi'(z) = -z\varphi(z)$. The density $f_{X^*|\sigma^*}$ is given by $\mu * \varphi_\sigma$, and thus $f_{Z^*|\sigma^*}(z|1/\pi) = \int \varphi(z - \theta\pi)\, d\mu(\theta)$, which implies

$$\begin{aligned}
\partial_z f_{Z^*|\sigma^*}(z|1) &= -\int (z-\theta)\,\varphi(z-\theta)\, d\mu(\theta) \\
\partial_z^2 f_{Z^*|\sigma^*}(z|1) &= -f_{Z^*|\sigma^*}(z|1) + \int (z-\theta)^2\,\varphi(z-\theta)\, d\mu(\theta) \\
\partial_\pi f_{Z^*|\sigma^*}(z|1) &= \int \theta\,(z-\theta)\,\varphi(z-\theta)\, d\mu(\theta) \\
&= -\left[f_{Z^*|\sigma^*}(z|1) + z \cdot \partial_z f_{Z^*|\sigma^*}(z|1) + \partial_z^2 f_{Z^*|\sigma^*}(z|1)\right],
\end{aligned}$$

from which we conclude

$$h''(z) = (C_1 - 1 - g(z)) \cdot h(z) - z \cdot h'(z). \tag{10}$$

Equation (10) is a second order linear homogenous ordinary differential equation.

Two free parameters Given the initial conditions $h(0) = h_0$ and $h'(0) = h_0'$, and given C_1, the solution to this equation exists and is unique, because all coefficients are continuous in z; cf. Murphy (2011). Furthermore, the general solution to this differential equation can be written in the form $h(z, C_1, h_0, h_0') = h_0 \cdot h_1(z, C_1) + h_0' \cdot h_2(z, C_1)$, where the functions $h_1(\cdot)$ and $h_2(\cdot)$ are determined by equation (10); cf. Murphy (2011), chapter B. This leaves three free parameters to be determined, C_1, h_0 and h_0'. The constraint $\int h(z)dz = 1$ pins down h_0 or h_0' given the other two parameters, so that there remain two free parameters.

We next turn to the second derivative $k(\cdot)$ defined by

$$k(z) = \partial_\pi^2 \log f_{Z|\sigma}(z|1) = C_2 + \partial_\pi^2 \log f_{Z^*|\sigma^*}(z|1),$$

which is identified under our assumptions, just like $g(\cdot)$. Calculations similar to those for the first derivative with respect to π yield the fourth order differential equation

$$h^{(4)}(z) = \left(k(z) - C_2 + (g(z) - C_1)^2 - 2\right) h(z) - 4zh'(z) - (z^2+5)h''(z) - 2zh^{(3)}(z). \tag{11}$$

To complete this proof, we now (i) derive the fourth order differential equation (11) and (ii) show that it allows us to pin down the remaining free parameters. We provide further discussion immediately following the proof.

Derivation of the fourth order ODE for $h(\cdot)$ Differentiating $\log f_{Z^*|\sigma^*}$ twice yields

$$\partial_\pi^2 \log f_{Z^*|\sigma^*}(z|1) = \frac{\partial_\pi^2 f_{Z^*|\sigma^*}(z|1)}{h(z)} - (g(z) - C_1)^2,$$

so that

$$\partial_\pi^2 f_{Z^*|\sigma^*}(z|1) = h(z) \cdot \left(k(z) - C_2 + (g(z) - C_1)^2\right).$$

From $f_{Z^*|\sigma^*}(z|1/\pi) = \int \varphi(z - \theta\pi)\, d\mu(\theta)$ we note that

$$\partial_\pi^2 f_{Z^*|\sigma^*}(z|1) = \int \left(-\theta^2 + \theta^2 (z-\theta)^2\right) \varphi(z-\theta)\, d\mu(\theta).$$

We furthermore have

$$\begin{aligned}
h^{(3)} &= -3h'(z) - \int (z-\theta)^3 \varphi(z-\theta)\, d\mu(\theta) \\
h^{(4)} &= -3h''(z) - 3\int (z-\theta)^2 \varphi(z-\theta)\, d\mu(\theta) + \int (z-\theta)^4 \varphi(z-\theta)\, d\mu(\theta) \\
&= -6h''(z) - 3h(z) + \int (z-\theta)^4 \varphi(z-\theta)\, d\mu(\theta).
\end{aligned}$$

Comparing coefficients on θ between $\partial_\pi^2 f_{Z^*|\sigma^*}$ and the derivatives of $h(\cdot)$, we get the fourth order differential equation (11).

The fourth order ODE pins down the remaining free parameters Our proof is complete once we have shown that there is at most one set of values C_1, C_2, h_0 and h_0' such that the resulting h satisfies the two differential equations (10) and (11). Differentiating equation (10) three times yields

$$\begin{aligned}
h''(z) = {} & (-1 + C_1 - g(z))h(z) & -zh'(z) & & \\
h^{(3)}(z) = {} & -g'(z)h(z) & +(-2 + C_1 - g(z))h'(z) & -zh''(z) & \\
h^{(4)}(z) = {} & -g''(z)h(z) & -2g'(z)h'(z) & +(-3 + C_1 - g(z))h''(z) & -zh^{(3)}(z) \\
h^{(5)}(z) = {} & -g^{(3)}(z)h(z) & -3g''(z)h'(z) & -3g'(z)h''(z) & \\
 & & & +(-4 + C_1 - g(z))h^{(3)}(z) & -zh^{(4)}(z),
\end{aligned}$$

and differentiating equation (11) yields

$$\begin{aligned}
h^{(4)}(z) = {} & \left(-2 - C_2 + (-C_1 + g(z))^2 + k(z)\right) h(z) & -4zh'(z) & \\
 & & -\left(5 + z^2\right) h''(z) & -2zh^{(3)}(z), \\
h^{(5)}(z) = {} & \left(2(-C_1 + g(z))g'(z) + k'(z)\right) h(z) & +\left(-6 - C_2 + (C_1 - g(z))^2 + k(z)\right) h'(z) & \\
 & -6zh''(z) & +\left(-7 - z^2\right) h^{(3)}(z) & -2zh^{(4)}(z).
\end{aligned}$$

We can iteratively eliminate the derivatives of $h(\cdot)$ from these equations by substitution. After doing so, we divide by $h(z)$, which is possible since $h(z) > 0$ for all z by construction. This yields the following equation involving the constants C_1 and C_2, but not involving the function $h(\cdot)$ or any of its derivatives:

$$\begin{aligned}
C_1^2 + C_2^2 + g(z)^2 + k(z)^2 - z^2 g'(z)^2 + 4k(z)g''(z) + 3g''(z)^2 & \\
- 2C_2 \left(g(z) + k(z) + 2g''(z)\right) + 2g(z) \left(k(z) + 2\left(g'(z)^2 + g''(z)\right)\right) & \\
+ C_1 \left(2C_2 - 2\left(g(z) + k(z) + 2\left(g'(z)^2 + g''(z)\right)\right)\right) - 2g'(z)g^{(3)}(z) = 2g'(z)k'(z) &
\end{aligned}$$

This equation again has to hold for all z. Differentiating twice with respect to z yields new equations where the constants C_1 and C_2 enter only linearly, and we can explicitly solve for them.[6]

Substituting the solutions C_1 and C_2 back into one of the first order differential equations we obtained by substitution and elimination of higher order derivatives above, we obtain a solution for h_0' given h_0. Given h_0, h_0' and the constants C_1 and C_2, equation (10) yields a unique solution $h(z)$ for all z. Rescaling any solution $h(\cdot)$ by a constant again yields a solution by linearity of the differential equations. h_0 is finally pinned down by the constraint $\int h(z)dz = 1$. $\square$

[6]The resulting expressions are unwieldy and so are omitted here, but are available on request.

Remarks:

- The proof of Theorem 3 shows that our model is overidentified. If we consider higher order derivatives of equations (10) and (11), or alternatively evaluate them at different values z, we obtain infinitely many restrictions on a finite number of free parameters.

- The proof of identification is considerably simplified if we restrict the model to a normal distribution for Θ^*, $\Theta^* \sim N(\bar{\mu}, \tau^2)$, which implies $Z^*|\sigma^* = 1 \sim N(\bar{\mu}, \tau^2 + 1)$, and thus $h(z) = const. \cdot \exp\left(-\frac{1}{2(\tau^2+1)}(z-\bar{\mu})^2\right)$. Denoting $e(z) = \partial_z \log h(z)$, we can rewrite equation (10) as

$$e'(z) = C_1 - g(z) - 1 - ze(z) - e^2(z),$$

  while the normality assumption yields $e(z) = -(z-\bar{\mu})/(\tau^2+1)$ and $e'(z) = -\frac{1}{(\tau^2+1)}$. Plugging in yields

$$-\tfrac{1}{(\tau^2+1)} = C_1 - g(z) - 1 + z\tfrac{z-\bar{\mu}}{(\tau^2+1)} - \left(\tfrac{z-\bar{\mu}}{(\tau^2+1)}\right)^2.$$

  Evaluating this equation at different values z pins down τ^2 and $\bar{\mu}$.

- The proof of Theorem 3 could be equivalently stated in terms of linear operators rather than differential equations. In particular, equations (10) and (11) are equivalent to the following two equations, indexed by z and linear in μ,

$$\int \left[\theta\left(z-\theta\right) - \left(g(z) - C_1\right)\right] \varphi\left(z-\theta\right) d\mu(\theta) = 0$$
$$\int \left[\left(-\theta^2 + \theta^2\left(z-\theta\right)^2\right) - \left(k(z) - C_2 + \left(g(z) - C_1\right)^2\right)\right] \varphi\left(z-\theta\right) d\mu(\theta) = 0$$

  Identification is then equivalent to the "completeness condition" that there is at most one probability measure μ in the orthocomplement of the span of the functions of θ

$$\left[\theta\left(z-\theta\right) - \left(g(z) - C_1\right)\right] \varphi\left(z-\theta\right) \text{ and}$$
$$\left[\left(-\theta^2 + \theta^2\left(z-\theta\right)^2\right) - \left(k(z) - C_2 + \left(g(z) - C_1\right)^2\right)\right] \varphi\left(z-\theta\right).$$

Proof of Corollary 3: The proof proceeds like the proof of Corollary 2. Let $S_i^* = \pm 1$ with probability 0.5, independently of $(X_i^*, \sigma_i^*, \Theta_i^*)$, and $S_j = S_{I_j}^*$. Define $V = S \cdot |X|$. We show that (V, σ) satisfies the assumptions of Theorem 3, from which the claim then follows.

Define $\tilde{S}^* = S^* \cdot \text{sign}(X^*)$, so that $V = \tilde{S} \cdot X$, and define $\tilde{\Theta}^* = \tilde{S}^* \cdot \Theta^*$. Since $\tilde{S}$ is independent of (Z, σ, Θ), we get $\tilde{\Theta}^* \sim \tilde{\mu} = \frac{1}{2}(\mu_{\Theta^*} + \mu_{-\Theta^*})$ and

$$f_{V/\sigma|\sigma}(z|\sigma) = \frac{p(z) \cdot \int \varphi(z - \theta/\sigma) d\tilde{\mu}(\theta)}{\int p(z')\varphi(z' - \theta/\sigma) dz' d\tilde{\mu}(\theta)}.$$

This has the exact same form as the density of Z given σ under the symmetric measure $\tilde{\mu}$. The claim follows, where we again use the fact that identification of $\tilde{\mu}$ implies identification of the distribution of $|\Theta^*|$. □

Proof of Theorem 4 For the first claim, note that since $F_{X|\Theta}(x|\theta)$ tends to zero as $\theta \to -\infty$ and tends to one as $\theta \to \infty$, for any x and any $\alpha \in (0, 1)$ there exist $\theta_l(x)$ and $\theta_u(x)$ such that

$$F_{X|\Theta}(x|\theta_u(x)) < \alpha < F_{X|\Theta}(x|\theta_l(x)),$$

where since $F_{X|\Theta}(x|\theta)$ is decreasing in θ we know that $\theta_l(x) < \theta_u(x)$. Thus, since $F_{X|\Theta}(x|\theta)$ is continuous in θ, the intermediate value theorem implies that there exists $\hat{\theta}_\alpha(x) \in (\theta_l(x), \theta_u(x))$ such that $F_{X|\Theta}\left(x|\hat{\theta}_\alpha(x)\right) = \alpha$. Since $F_{X|\Theta}(x|\theta)$ is strictly decreasing we know this $\hat{\theta}_\alpha(x)$ is unique, while its strict monotonicity and continuity likewise follow from strict monotonicity and continuity of $F_{X|\Theta}$ in both arguments.

For the second claim, note that since $F_{X|\Theta}(x|\theta)$ is strictly decreasing in θ, we have $\hat{\theta}_\alpha(x) \leq \theta$ if and only if $F_{X|\Theta}(x|\theta) \leq \alpha$. Continuity of $F_{X|\Theta}(x|\theta)$ in x, however, means that X is continuously distributed conditional on $\Theta = \theta$ for all θ, and thus that $F_{X|\Theta}(X|\theta)$ is uniformly distributed conditional on $\Theta = \theta$. Thus,

$$P\left(F_{X|\Theta}(x|\theta) \leq \alpha | \Theta = \theta\right) = \alpha,$$

so

$$P\left(\hat{\theta}_\alpha(X) \leq \theta | \Theta = \theta\right) = \alpha \text{ for all } \theta,$$

as we aimed to show. □

Proof of Lemma 5 Under the stated assumptions, Lemma 1 implies that X is continuously distributed under all $\theta \in \mathbb{R}$, with density given by (1). To prove the strict monotonicity of $F_{X|\Theta}(x|\theta)$ in θ, we adapt the proof of Lemma A.1 in Lee et al. (2016).

In particular, note that for $x_1 > x_0$ and $\theta_1 > \theta_0$,

$$\frac{f_{X|\Theta}(x_1|\theta_1)}{f_{X|\Theta}(x_0|\theta_1)} > \frac{f_{X|\Theta}(x_1|\theta_0)}{f_{X|\Theta}(x_0|\theta_0)},$$

as can be verified from multiplying out these expressions. This means, however, that

$$f_{X|\Theta}(x_1|\theta_1)f_{X|\Theta}(x_0|\theta_0) > f_{X|\Theta}(x_1|\theta_0)f_{X|\Theta}(x_0|\theta_1).$$

Integrating both sides with respect to x_0 from $-\infty$ to $x < x_1$, and with respect to x_1 from x to ∞, we obtain that

$$(1 - F_{X|\Theta}(x|\theta_1))F_{X|\Theta}(x|\theta_0) > (1 - F_{X|\Theta}(x|\theta_0))F_{X|\Theta}(x|\theta_1),$$

and thus that $F_{X|\Theta}(x|\theta_0) > F_{X|\Theta}(x|\theta_1)$. Since this argument applies for all x and all θ_0, θ_1, we have shown that $F_{X|\Theta}(x|\theta)$ is strictly decreasing in θ for all x.

To prove that $F_{X|\Theta}(x|\theta) \to 0$ as $\theta \to \infty$, note that by our assumption that $p(x)$ is almost everywhere continuous, for any x_0 there exists a point $x_1 > x_0$, and an open neighborhood $(x_1 - \varepsilon, x_1 + \varepsilon)$ of x_1 such that $p(\cdot)$ is continuous on the closure of this neighborhood, and $x_0 < x_1 - 2\varepsilon$. Note, however, that for $\theta > x_1 + \varepsilon$, $f_{X|\Theta}(x|\theta)$ for $x \leq x_0$ is bounded above by $\varphi((x - \theta)/\sigma)/(\sigma \cdot E[p(X)|\Theta^* = \theta])$. On the other hand, the infimum of $f_{X|\Theta}(x|\theta)$ over $(x_1 - \varepsilon, x_1 + \varepsilon)$ is bounded below by $p_l \cdot \varphi((x_1 - \varepsilon - \theta)/\sigma)/(\sigma \cdot E[p(X)|\Theta^* = \theta])$ for

$$p_l = \inf_{x \in [x_1 - \varepsilon, x_1 + \varepsilon]} p(x) > 0.$$

Integrating and taking the ratio, we see that

$$\frac{P(x \leq x_0|\Theta = \theta)}{P(x \in (x_1 - \varepsilon, x_1 + \varepsilon)|\Theta = \theta)} \leq \frac{\Phi((x_0 - \theta)/\sigma)}{2\varepsilon p_l \cdot \varphi((x_1 - \varepsilon - \theta)/\sigma)/\sigma}.$$

This expression can in turn be bounded above by

$$\frac{\Phi((x_0-\theta)/\sigma)}{2\varepsilon p_l \cdot \varphi((x_0-\theta)/\sigma)/\sigma},$$

which is proportional to Mill's ratio and tends to zero and $\theta \to \infty$ (see, for example, Baricz (2008)). This immediately implies that $F_{X|\Theta}(x_0|\theta) \to 0$, as we aimed to show. The claim that $F_{X|\Theta}(x|\theta) \to 1$ as $\theta \to \infty$ can be proved analogously. □

B Meta-regression coefficients

In Section 3.3 of the main text we discussed meta-regressions. We noted that under our assumptions meta-regressions deliver a valid test of the null of no selectivity. We also noted, however, that in the presence of selectivity the function $E[Z|1/\sigma = \pi]$ is in general non-linear, and the slope of the best linear predictor cannot be interpreted as a selection-corrected estimate of $E[\Theta^*]$.

To see this, consider the following simple example. Suppose that $\Theta^* \equiv \bar{\theta} > 0$, so there is no parameter heterogeneity across latent studies, and that $p(Z) = \mathbf{1}(Z > z^c)$, so there is strict selection on significant, positive effects. Let $\varepsilon \sim N(0,1)$, and let h be the inverse Mill's ratio, $h(x) = \frac{\varphi(x)}{1-\Phi(x)}$. Then

$$E[Z|1/\sigma = \pi] = E[\pi\bar{\theta} + \varepsilon | \pi\bar{\theta} + \varepsilon > z^c] = \pi\bar{\theta} + h\left(z^c - \pi\bar{\theta}\right).$$

Confirming the general point made above, this is a nonlinear function of π, and the slope and intercept of the best linear predictor approximating this function both depend on the distribution of π (that is, of σ). If σ takes on only small values, the Mill's ratio term is negligible, and $E^*[Z|1/\sigma = \pi] \approx \pi\bar{\theta}$. If σ takes on only large values, a first order approximation around $\pi = 0$ yields

$$E^*[Z|1/\sigma = \pi] \approx h(z^c) + \bar{\theta}(1 - h'(z^c)) \cdot \pi.$$

This shows in particular that the slope, which in this example equals $\bar{\theta}(1 - h'(z^c))$, is in general different from the average effect $\bar{\theta}$, so that meta-regressions cannot be expected to deliver bias-corrected estimates of $E[\Theta^*]$.

C Latent selection model

The baseline model we consider assumes that $E[D = 1|X^*, \Theta^*] = p(X^*)$, so that there is no dependence of publication probabilities on the latent parameter given X^*. In the context of systematic replication studies with normally distributed estimates, however, we showed that a more general class of models which allows for dependence of $p(\cdot)$ on Θ^* is identified. In Section 5.1.1 we introduced a parametric specification for such a more general model, which we then estimate to provide a specification check for our baseline model.

The parametric specification introduced in Section 5.1.1 can be derived as follows. Assume that publication decisions are based on

$$\begin{pmatrix} Z^* \\ V^* \end{pmatrix} | \Theta^* \sim N\left(\begin{pmatrix} \Theta^* \\ \Theta^* \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right),$$

where V^* is a second, independent estimate of the true effect Θ^*, with the same variance as Z^*. Assume further that

$$D_i | Z_i^*, V_i^*, \Theta_i^* \sim Ber\left(p\left(Z_i^*, V_i^*\right)\right),$$

so publication decisions are based on Z_i^* and V_i^*. Since V_i^* is unobserved, integrating over its distribution gives publication probabilities of the form $p(Z^*, \Theta^*)$.

We want our specification for $p\left(z, v\right)$ to nest our baseline specifications,

$$p\left(z\right) = \sum_{k=1}^{K} \beta_{p,k} 1\left\{\zeta_{k-1} \leq z < \zeta_k\right\}.$$

To ensure this, we consider the generalized specification

$$\begin{array}{rc} p\left(z, v\right) = & \sum_{k=1}^{K} \tilde{\beta}_{p,k}^{1} 1\left\{\zeta_{k-1} \leq z < \zeta_k, |v| \geq \zeta_V\right\} \\ & + \sum_{k=1}^{K} \tilde{\beta}_{p,k}^{0} 1\left\{\zeta_{k-1} \leq z < \zeta_k, |v| < \zeta_V\right\}, \end{array}$$

which allows publication probabilities to depend on whether two-sided z-tests based on the latent variable v reject $\Theta^* = 0$. Integrating over the distribution of V^* yields

the following specification for $p(z,\theta)$:

$$p(z,\theta) = \begin{array}{c} \sum_{k=1}^{K} \tilde{\beta}_{p,k}^{1} 1\{\zeta_{k-1} \leq z < \zeta_k\} \left(1 - \tilde{\Psi}(\zeta_V, \theta)\right) \\ + \sum_{k=1}^{K} \tilde{\beta}_{p,k}^{0} 1\{\zeta_{k-1} \leq z < \zeta_k\} \tilde{\Psi}(\zeta_V, \theta), \end{array}$$

where

$$\tilde{\Psi}(\zeta_V, \theta) = Pr\{|V| < \zeta_V | \Theta^* = \theta\} = \Phi(\zeta_V - \theta) - \Phi(-\zeta_V - \theta).$$

As noted in the main text, $p(z,\theta)$ is only nonparametrically identified up to a normalization for each value θ. Analogous to our baseline specifications, here we impose the normalization $\tilde{\beta}_{p,K}^{1} = \tilde{\beta}_{p,K}^{0} = 1$. To obtain the specification discussed in Section 5.1.1, we then define

$$\beta_{p,k} = \tilde{\beta}_{p,k}^{1} + \tilde{\Psi}(\zeta_V, 0) \cdot (\tilde{\beta}_{p,k}^{0} - \tilde{\beta}_{p,k}^{1}),$$

$$\gamma_{p,k} = \left(\tilde{\beta}_{p,k}^{1} - \tilde{\beta}_{p,k}^{0}\right) \cdot \tilde{\Psi}(\zeta_V, 0),$$

and

$$\Psi(\zeta_V, \theta) = \frac{\tilde{\Psi}(\zeta_V, \theta) - \tilde{\Psi}(\zeta_V, 0)}{-\tilde{\Psi}(\zeta_V, 0)},$$

which yields the specification

$$p(z,\theta) = \sum_{k=1}^{K} (\beta_{p,k} + \gamma_{p,k} \cdot \Psi(\zeta_V, \theta)) \cdot 1\{\zeta_{k-1} \leq z < \zeta_k\}.$$

Note that our normalization now implies that $\beta_{p,K} = 1$ and $\gamma_{p,K} = 0$. For our specification tests we set $\zeta_V = 1.96$, corresponding to a 5% test based on V^*.

D Application details

In this section, we give additional details on our applications in Section 5 of the main text and discuss how we cast the data of Camerer et al. (2016) and Open Science Collaboration (2015) into our framework.

D.1 Details for economics laboratory experiments

We first discuss our results based on data from Camerer et al. (2016). To apply our approach, we need z-statistics and standard errors for both the original and replication studies. We first collect p-values and standardized effect sizes from table S1 in the supplement to Camerer et al. (2016). Some of the p-values are censored below at .001, so for these studies we also collect the original estimates and standard errors from the replication reports posted online by Camerer et al.[7] and recompute the censored p-values. We then construct z-statistics by inverting the p-value transformation, where $z = \Phi^{-1}(1-p/2)$. To obtain effect size estimates, we apply the Fisher transformation to standardized effect sizes reported by Camerer et al. Dividing these estimates by the z-statistics finally recovers the standard error.

We can infer the sign of the z-statistics from the sign of the standardized effect. Since signs are arbitrary and not comparable across studies, however, we normalize all signs to be positive.

D.2 Details for psychology laboratory experiments

To apply our approach to the data from Open Science Collaboration (2015), we again need z-statistics and standard errors for both the original and replication studies. We draw the inputs for all of these calculations from the RPPdataConverted spreadsheet posted online by the Open Science Collaboration.[8] Since Open Science Collaboration (2015) report p-values for both the original and replication studies, we invert the p-value transform to obtain z statistics. We use the p-values reported in their columns T.pval.USE.O and T.pval.USE.R for the original and replication studies, respectively. Since some of the p-values in this application are based on one-sided tests, we account for this in the inversion step. To compute effect size estimates, we again apply the Fisher transformation to the standardized effect sizes (columns T.r.O and T.r.R of RPPdataConverted for the original and replication studies, respectively), and then divide these estimates by the z-statistics to construct standard errors.

[7] Available at `https://experimentaleconreplications.com/replicationreports.html`, accessed September 3, 2016.

[8] Available at `https://osf.io/ytpuq/files/`, accessed January 19, 2017.

E Additional empirical results

E.1 Additional results for psychology laboratory experiments

Here we report results based on two alternative specifications for the psychology replication data from Open Science Collaboration (2015). First, we limit attention to studies with a large number of denominator degrees of freedom. Second, we limit attention to studies where the replication protocols were approved by the original authors.

Denominator degrees of freedom As noted in the main text, our baseline analysis of the Open Science Collaboration (2015) data focuses on studies that use z- or t-statistics (or the squares of these statistics). Our analysis then treats these statistics as approximately normal. A potential problem here is that t-distributions with a small number of degrees of freedom behave differently from normal distributions, and in particular have heavier tails. While the smallest degrees of freedom in the Open Science Collaboration (2015) data is seven, this concern may still lead us to worry about the validity of our approach in this setting. To address this concern, in Table 5 we report parameter estimates using the replication and meta-study specifications discussed in Section 5.3, where

$$p(Z) \propto \begin{cases} \beta_{p,1} & |Z| < 1.64 \\ \beta_{p,2} & 1.64 \leq |Z| < 1.96 \\ 1 & |Z| \geq 1.96, \end{cases}$$

except that we now limit attention to the 52 observations with denominator degrees of freedom at least 30 in the original and study.[9] As these results make clear, our results are broadly similar for this restricted sample and for the full data.

[9]We screen only on the degrees of freedom in the original study since sample sizes, and thus degrees of freedom, in the replication studies depend on the results in the initial study. Hence, screening on replication degrees of freedom has the potential to introduce additional selection on the results of the original study. That said, screening on degrees of freedom for both the original and replication studies yields a sample of 49 studies and extremely similar results to those reported here.

Replication			Meta-study		
τ	$\beta_{p,1}$	$\beta_{p,2}$	$\tilde{\tau}$	$\beta_{p,1}$	$\beta_{p,2}$
1.181	0.019	0.217	-0.221	0.030	0.327
(0.257)	(0.014)	(0.125)	(0.046)	(0.021)	(0.190)

Table 5: Selection estimates from lab experiments in psychology, restricted to observations with denominator degrees of freedom at least 30, with 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 omitted category of studies significant at 5% level.

Approved replications As discussed in the main text, Gilbert et al. (2016) argue that some of the replications in Open Science Collaboration (2015) deviated substantially from the protocol of the original studies, which might lead to a violation of our assumption that the replication and original results are generated by the same underlying parameter Θ. Before conducting their replications, however, Open Science Collaboration (2015) asked the authors of each original study to review the proposed replication protocol, and recorded whether the original authors endorsed the replication protocol. We can thus partly address this critique by limiting attention to the subset of studies where the replication was endorsed by the authors of the original study. Re-estimating the specifications of Section 5.3 on the 51 endorsed replications, we obtain the estimates reported in Table 6. These estimates suggest a somewhat smaller degree of selection than our baseline estimates, consistent with a higher rate of replication for approved replications, but are broadly similar to our other estimates.

Replication			Meta-study		
τ	$\beta_{p,1}$	$\beta_{p,2}$	$\tilde{\tau}$	$\beta_{p,1}$	$\beta_{p,2}$
1.385	0.038	0.512	0.272	0.042	0.621
(0.272)	(0.024)	(0.239)	(0.055)	(0.027)	(0.300)

Table 6: Selection estimates from lab experiments in psychology, approved replications, with 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 omitted category of studies significant at the 5% level.

E.2 Additional results for minimum wage meta-study

As noted in the main text, the data from Wolfson and Belman (2015) include estimates from both published and working papers. While our analysis in the main text uses

the full data, Table 7 reports estimates of the model

$$\Theta^* \sim N(\bar{\theta}, \tilde{\tau}^2), \quad p(X/\sigma) \propto \begin{cases} \beta_{p,1} & X/\sigma < -1.96 \\ \beta_{p,2} & -1.96 \leq X/\sigma < 0 \\ \beta_{p,3} & 0 \leq X/\sigma < 1.96 \\ 1 & X/\sigma \geq 1.96 \end{cases}$$

based on the subset of published papers, consisting of 705 estimates drawn from 31 studies. As in the main text we cluster our standard errors at the study level. The resulting estimates are broadly similar to those obtained on the full sample.

$\bar{\theta}$	$\tilde{\tau}$	$\beta_{p,1}$	$\beta_{p,2}$	$\beta_{p,3}$
-0.019	0.145	0.345	0.453	0.651
(0.045)	(0.034)	(0.171)	(0.230)	(0.250)

Table 7: Meta-study selection estimates from minimum wage data, published studies, with standard errors in parentheses. Publication probability β_p is measured relative to omitted category of studies estimating a positive effect significant at the 5% level.

E.3 Additional results for deworming meta-study

In the main text, we report estimates for the deworming data of Croke et al. (2016) based on a specification that restricts $p(\cdot)$ to be symmetric around zero. To complement those results, here we consider the more flexible specification

$$\Theta^* \sim N(\bar{\theta}, \tau^2), \quad p(X/\sigma) \propto \begin{cases} \beta_{p,1} & X/\sigma < -1.96 \\ \beta_{p,2} & -1.96 \leq X/\sigma < 0 \\ \beta_{p,3} & 0 \leq X/\sigma < 1.96 \\ 1 & X/\sigma \geq 1.96. \end{cases}$$

Results based on this specification are reported in Table 8. These estimates differ substantially from those reported in the main text, and suggest strong selectivity against negative estimates, particularly negative and significant estimates. However, as can be seen from Figure 10 in the main text there is only a single negative and statistically significant estimate in the sample, so the reliability of conventional asymptotic

approximations here is highly suspect.

$\bar{\theta}$	$\tilde{\tau}$	$\beta_{p,1}$	$\beta_{p,2}$	$\beta_{p,3}$
-0.714	0.521	0.008	0.151	1.299
(0.626)	(0.206)	(0.025)	(0.207)	(1.113)

Table 8: Meta-study selection estimates from deworming wage data, flexible specification, with standard errors in parentheses. Publication probability β_p is measured relative to omitted category of studies estimating a positive effect significant at the 5% level.

To reduce the number of free parameters, we estimate a version of the model which does not allow discontinuities in $p(\cdot)$ based on statistical significance, but only based on the sign of the estimate,

$$\Theta^* \sim N(\bar{\theta}, \tau^2), \quad p(X/\sigma) \propto \begin{cases} \beta_p & X/\sigma < 0 \\ 1 & X/\sigma \geq 0. \end{cases}$$

Fitting this model to the data yields the estimates reported in Table 9. These estimates suggest strong selectivity on the sign of the estimated effect, where positive effects are estimated to be ten times more likely to be published than negative effects. While this is consistent with the distribution of observations in Figure 10, our choice of this specification was driven by our results in Table 8. Given that this is a form of specification search, it suggests that conventional asymptotic approximations may be unreliable here, and thus that these results should be treated with caution.

$\bar{\theta}$	$\tilde{\tau}$	β_p
-0.217	0.365	0.094
(0.156)	(0.103)	(0.099)

Table 9: Meta-study selection estimates from deworming wage data, restricted asymmetric specification, with standard errors in parentheses. Publication probability β_p is measured relative to omitted category of studies estimating a positive effect significant at the 5% level.

F Inference corrections based on estimates

In this section, we plot our median unbiased estimators and corrected confidence sets, analogous to Figure 4 of the paper, based on the selection estimates from our

applications. The estimates based on the Camerer et al. (2016) data match those used to generate Figure 4, so we do not plot this again. Corrections based on replication estimates from the Open Science Collaboration (2015) data are plotted in Figure 11. Corrections based on estimates from the Croke et al. (2016) data are plotted in figure 12. Finally, corrections based on estimates using data from Wolfson and Belman (2015) are reported in Figure 13.

G Inference in multivariate normal models

In this section, we extend the frequentist inference results developed in the main text to cases where publication decisions are based not just on a scalar, but instead on a normally distributed vector of estimates. Let X_i^* represent the estimates from study i, and assume that

$$X_i^* | \Theta_i^* \sim N\left(\Theta_i^*, \Sigma\right)$$

for Σ known. Assume that Σ is constant across latent studies i; the generalization to the case where latent study i has variance Σ_i^* is immediate. Since X_i^* is a vector, Σ is a matrix. We thus get the following density for X^* given Θ^*:

Assumption 1
The distribution $f_{X^*|\Theta^*}(x|\theta)$ is multivariate normal with mean θ and variance Σ:

$$f_{X^*|\Theta^*}(x|\theta) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\theta)' \Sigma^{-1} (x-\theta)\right).$$

We consider inference on $\Gamma = v'\Theta$ for a known non-zero vector v, treating the other elements of Θ, denoted Ω, as nuisance parameters. To conduct inference on the ith element of Θ we can simply take v to be the ith standard basis vector. To illustrate our results, we consider the example of difference in differences estimation, with selection on both statistical significance and a test for parallel trends.

G.1 Illustrative example: difference in differences

Suppose we observe data from two states, $s \in \{1, 2\}$ over three time periods $t \in \{1, 2, 3\}$. Denote the average outcome for residents of state s at time t by Y_{st}, and

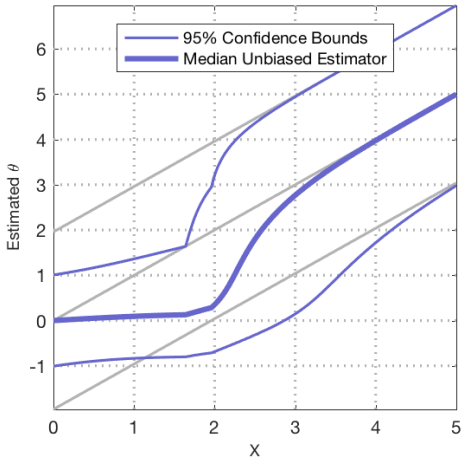

Figure 11: This figure plots frequentist 95% confidence bounds and the median unbiased estimator for the selection estimates based on replication data from Open Science Collaboration (2015). The usual (uncorrected) estimator and confidence bounds are plotted in grey for comparison.

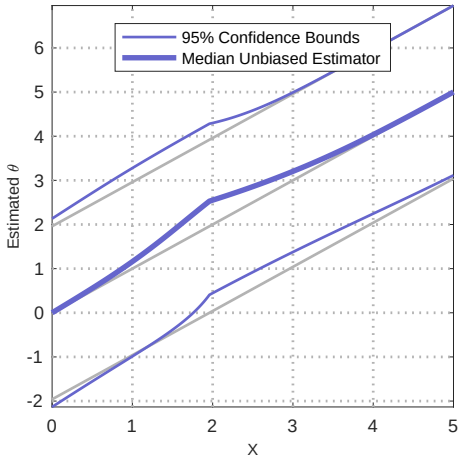

Figure 12: This figure plots frequentist 95% confidence bounds and the median unbiased estimator for the selection estimates based on replication data from Croke et al. (2016). The usual (uncorrected) estimator and confidence bounds are plotted in grey for comparison.

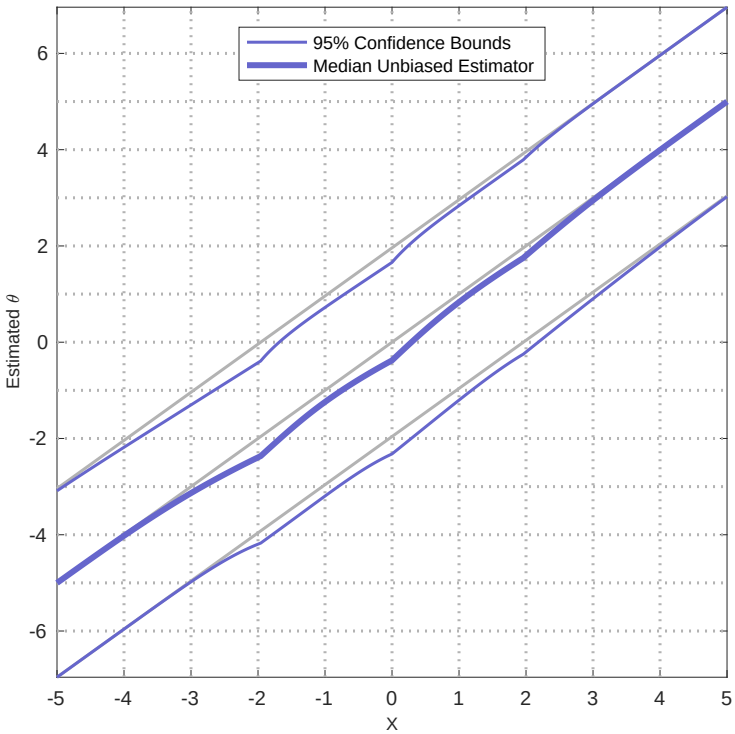

Figure 13: The figure to the left plots frequentist 95% confidence bounds and the median unbiased estimator for the selection estimates based on replication data from Wolfson and Belman (2015). The usual (uncorrected) estimator and confidence bounds are plotted in grey for comparison.

note that under regularity conditions, Y_{st} will be approximately normally distributed

$$Y_{st} \sim N\left(\mu_{st}, \sigma_{st}^2\right).$$

For simplicity we assume that Y_{st} is independent of $Y_{s't'}$ if $s \neq s'$ or $t \neq t'$.

Suppose we are interested in estimating the effect of a particular state-level policy, and let D_{st} be a dummy for the presence of the policy in state s at time t. The difference in differences model (with no control variables) assumes that

$$\mu_{st} = \alpha_s + \beta_t + D_{st}\gamma.$$

If we are interested in the effect of a policy enacted in state 1 in period 3 and nowhere else in the sample, for example, we would take

$$D_{st} = \{s = 1, t = 3\}.$$

A key identifying assumption in the difference-in-differences model is that the only source of variation in μ_{st} at the state-by-time level is the policy change of interest. In particular, while we allow state fixed effects α_s and time fixed effects β_t, we rule out state-time-specific effects other than those acting through D_{st}. This is known as the parallel trends assumption.

With only two periods of data this assumption is untestable, since we have four free parameters $(\alpha_1, \alpha_2, \beta_2, \gamma)$ and only four means $(\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})$. With data from an additional time period, however, we have five free parameters and six means and so can instead consider the model

$$\mu_{st} = \alpha_s + \beta_t + \tilde{D}_{st}\lambda + D_{st}\gamma$$

where

$$\tilde{D}_{st} = \{s = 1, t = 2\}$$

and the parallel trends assumption implies that $\lambda = 0$. Thus, given data from two states in three time periods the parallel trends assumption is testable.

Formal and informal tests of parallel trends are common in applications of difference in differences strategies. To describe a formal test in our setting, note that the

natural estimator (G, L) for (γ, λ) has a simple form,

$$(G, L) = ((X_{13} - X_{12}) - (X_{23} - X_{22}), (X_{12} - X_{11}) - (X_{22} - X_{21})).$$

To test the parallel trends assumption in this setting we again want to test that λ, the mean of L, is equal to zero.

Consider a population of latent studies with the structure just described, and let us further simplify the model by setting $\sigma_{st} = 1$ for all t. For latent estimates $X^* = (G^*, L^*)$ and latent true effects $\Theta^* = (\Gamma^*, \Lambda^*)$,

$$\left(\begin{array}{c} G^* \\ L^* \end{array} \right) \Bigg| \left(\begin{array}{c} \Gamma^* \\ \Lambda^* \end{array} \right) \sim N \left(\left(\begin{array}{c} \Gamma^* \\ \Lambda^* \end{array} \right), \left(\begin{array}{cc} 4 & 2 \\ 2 & 4 \end{array} \right) \right)$$

where the covariance matrix is known.

As in our illustrative example in the main text, assume studies that reject $\gamma = 0$ at the 5% level are ten times more likely to be published than studies that do not. In addition, assume studies that reject $\lambda = 0$ at the 5% level are ten times *less* likely to be published than studies that do not. This leads to publication probability

$$p(X) \propto 1\left\{ \frac{|G^*|}{\sigma_G} > 1.96, \frac{|L^*|}{\sigma_L} \leq 1.96 \right\} 1 + 1\left\{ \frac{|G^*|}{\sigma_G} > 1.96, \frac{|L^*|}{\sigma_L} \geq 1.96 \right\} 0.1$$

$$+1\left\{ \frac{|G^*|}{\sigma_G} \leq 1.96, \frac{|L^*|}{\sigma_L} \leq 1.96 \right\} 0.1 + 1\left\{ \frac{|G^*|}{\sigma_G} \leq 1.96, \frac{|L^*|}{\sigma_L} > 1.96 \right\} 0.01.$$

This publication rule favors studies that find significant difference in difference estimates, and disfavors studies that reject the parallel trends assumption.

To illustrate the effect of selective publication in this setting, Figure 14 plots the median bias of G as an estimator for γ (scaled by the standard deviation σ_G of G^*). Selective publication results in large bias for the conventional estimator G, which depends on both the parameter of interest γ and the nuisance parameter λ. Analogously, Figure 15 plots the coverage of the usual two-sided confidence set $G^* \pm 1.96\sigma_G$, and shows that selective publication results in substantial coverage distortions.

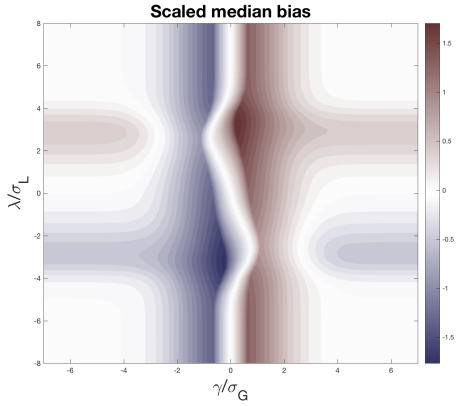

Figure 14: This figure plots the median bias of $(G)/\sigma_G$ in the difference in differences example.

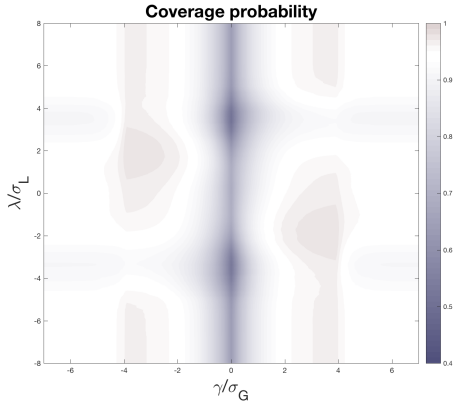

Figure 15: This figure plots the coverage of conventional 95% confidence sets in the difference in differences example.

G.2 Sufficient statistic for nuisance parameter

To conduct inference on γ, treating ω as a nuisance parameter, it will be helpful to derive a sufficient statistic for ω. Note that for $M(v)$ a $(\dim(X)-1)\times\dim(X)$ matrix such that $M(v)\left(I-\frac{\Sigma vv'}{v'\Sigma v}\right)$ has full row-rank,

$$(G(x),W(x))=\left(v'x,M(v)\left(I-\frac{\Sigma vv'}{v'\Sigma v}\right)x\right)$$

is a one-to one transformation of x. Thus $(G,W)=(G(X),W(X))$ are jointly sufficient for θ, and rather than basing inference on X we can equally well base inference on (G,W). Note moreover that for $G^*=G(X^*)$ and $W^*=W(X^*)$, $X^*\sim N(\theta,\Sigma)$ implies that

$$\begin{pmatrix} G^* \\ W^* \end{pmatrix} \sim N\left(\begin{pmatrix} \gamma \\ \omega \end{pmatrix}, \begin{pmatrix} \sigma_G^2 & 0 \\ 0 & \Sigma_W \end{pmatrix}\right) \tag{12}$$

for $\omega=M(v)\left(I-\frac{\Sigma vv'}{v'\Sigma v}\right)\theta$, $\sigma_G^2=v'\Sigma v$, and $\Sigma_W=M(v)\left(I-\frac{\Sigma vv'}{v'\Sigma v}\right)\Sigma\left(I-\frac{vv'\Sigma}{v'\Sigma v}\right)M(v)'$. Thus the conditional distribution of G^* given W^* depends only on γ,

$$G^*|W^*\sim N(\gamma,\sigma_G^*),$$

and by conditioning on W^* we can eliminate dependence on the nuisance parameter ω. This property continues to hold for the conditional distribution of published G given W, as the following lemma shows.

Lemma 6

Under Assumption 1, the conditional density $G|W,\Gamma$ is given by

$$f_{G|W,\Gamma}(g|w,\gamma) = \frac{p(g,w)}{E\left[p\left(G^*,W^*\right)|W^*=w,\Gamma^*=\gamma\right]}\frac{1}{\sigma_G}\phi\left(\frac{g-\gamma}{\sigma_G}\right) \tag{13}$$

for ϕ the standard normal density, where we use the fact that (g,w) is a one-to-one transformation of x to write $p(g,w) = p\left(x(g,w)\right)$.

Proof of Lemma 6 Note that we can draw from the conditional distribution $G|W=w,\Gamma=\gamma$ by drawing from the conditional distribution $G^*|W^*=w,\Gamma^*=\gamma$ and discarding the draw G^* with probability $1-p(G^*,w)$. The result then follows by the same argument as Lemma 1. □

Thus, we see that the conditional density of G given W depends only on the parameter of interest γ and not on the nuisance parameter ω. Hence, by conditioning on W we can eliminate the nuisance parameter and conduct inference on γ alone.

G.3 Optimal quantile-unbiased estimates

To conduct frequentist inference, we generalize the median-unbiased estimator and equal-tailed confidence set proposed in Section 4 to the present setting. Using a result from Pfanzagl (1994) we show that the resulting quantile-unbiased estimators are optimal in a strong sense.

Formally, define $\hat{\gamma}_\alpha(X)$ by

$$F_{G(X)|W(X),\Gamma}\left(G|W,\hat{\gamma}_\alpha(X)\right) = \alpha.$$

This estimator is simply the value γ such that the observed G lies at the α quantile of the corresponding conditional distribution given W. The following theorem, based on the results of Pfanzagl (1994), shows that this estimator is both quantile-unbiased and, in a strong sense, optimal in the class of quantile-unbiased estimators.

Theorem 5

Let Assumption 1 hold, and assume further that the conditional distribution of G given W is absolutely continuous for all γ and almost every W, and that the parameter space for ω given γ contains an open set for all γ. Then

1. *The estimator $\hat{\gamma}_\alpha(X)$ is level-α quantile unbiased:*

$$Pr\{\hat{\gamma}_\alpha(X) \leq \gamma | \Theta = (\gamma, \omega)\} = \alpha \text{ for all } \gamma, \omega,$$

2. *This estimator is uniformly most concentrated in the class of level-α quantile-unbiased estimators, in the sense that for any other level-α quantile unbiased estimator $\tilde{\gamma}(X)$ and any loss function $L(d, \gamma)$ that attains its minimum at $d = \gamma$ and is increasing as d moves away from γ,*

$$E[L(\hat{\gamma}_\alpha(X), \gamma) | \Theta = (\gamma, \omega)] \leq E[L(\tilde{\gamma}(X), \gamma) | \Theta = (\gamma, \omega)] \text{ for all } \gamma, \omega.$$

Proof of Theorem 5 Since the multivariate normal distribution belongs to the exponential family, we can write

$$f_{G^*, W^* | \Theta^*}(g, w | \theta) = \tilde{h}(g, w) \tilde{r}(\gamma(\theta), \omega(\theta)) \exp\left(\gamma(\theta) g + \omega(\theta)' w\right).$$

By the same argument as in the proof of Lemma 1, this implies that

$$f_{G,W|\Theta}(g, w | \theta) = h(g, w) r(\gamma(\theta), \omega(\theta)) \exp(\gamma(\theta) g) \exp\left(\omega(\theta)' w\right) \tag{14}$$

for $h(g, w) = p(g, w) \tilde{h}(g, w)$ and

$$r(\gamma, \omega) = \frac{\tilde{r}(\gamma, \omega)}{E[p(X_i^*) | \Theta_i^* = \theta(\gamma, \omega)]}.$$

The density (14) has the same structure as (5.5.14) of Pfanzagl (1994), and satisfies properties (5.5.1)-(5.5.3) of Pfanzagl (1994) as well. Part 1 of the theorem then follows immediately Theorem 5.5.9 of Pfanzagl (1994).

Part 2 of the theorem follows by using Theorem 5.5.9 of Pfanzagl (1994) along with (14) to verify the conditions of Theorem 5.5.15 of Pfanzagl (1994). □

Using this result we see that $\hat{\gamma}_{\frac{1}{2}}(X)$ is the optimal median-unbiased estimator for the parameter of interest γ. A natural level-α confidence interval to accompany this

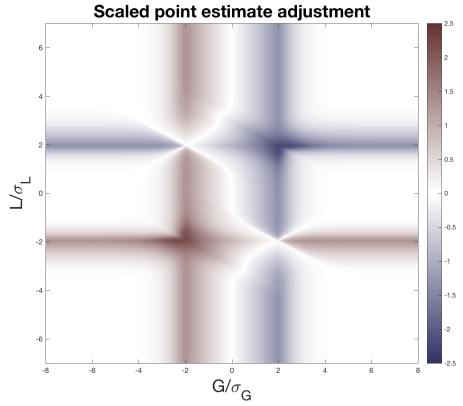

Figure 16: This figure plots the difference between the median-unbiased estiamtor $\hat{\gamma}_{\frac{1}{2}}(X)$ and the conventional estimator $\hat{G}$ in the difference-in-differences example.

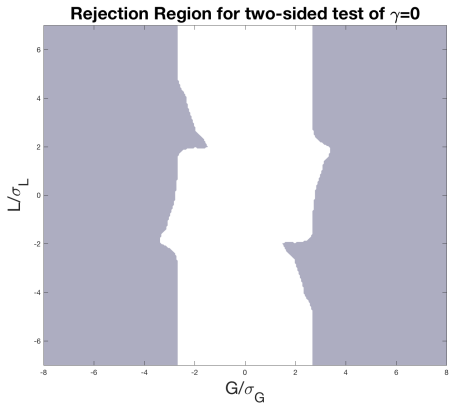

Figure 17: This figure plots the (shaded) rejection region for a 5% test of $H_0 : \gamma = 0$ based on equal-tailed confidence sets for γ in the differences in differences example.

estimator is then the equal-tailed confidence interval

$$CS = \left[\hat{\gamma}_{1-\frac{\alpha}{2}}(X), \hat{\gamma}_{\frac{\alpha}{2}}(X)\right].$$

Difference in differences example (continued) To illustrate our corrections in a multivariate setting, Figure 16 plots the difference between our median-unbiased estimator $\hat{\gamma}_{\frac{1}{2}}(X)$ and the conventional estimator $\hat{\gamma} = G$ in the difference-in-differences example. As this plot makes clear, $\hat{\gamma}_{\frac{1}{2}}(X)$ depends on both G and L. Thus, while we are interested only in the difference-in-differences parameter γ, the result for the pretest of parallel trends also plays a role in our estimate. Figure 17 plots the rejection region for a 5% test of $H_0 : \gamma = 0$ based on our equal-tailed confidence interval for γ. As this plot shows, the results of this test likewise depend on both G and L.

H Bayesian inference

In the main text we discuss the effect of selective publication on frequentist inference on θ under known $p(\cdot)$. The effect of selective publication on Bayesian inference is more subtle, and depends on the prior. Here we briefly discuss Bayesian inference on θ under known $p(\cdot)$ for two natural classes of priors. These priors can be thought of

as two extreme points of the set of relevant priors.

Definition 4 (Two classes of priors)
Consider the following two classes of prior distributions π_μ for μ:

1. *Unrelated Parameters: π_μ is a point mass at some μ, so that μ is known and the prior distribution of Θ_i^* is i.i.d. across i.*

2. *Common Parameters: π_μ assigns positive probability only to point-measures μ, so that Θ_i^* is constant across i (equal to Θ_0^*) with probability 1.*

The unrelated parameters prior corresponds to the case where each latent study considers a different parameter. Thus, under priors in this class, learning the true parameter value Θ_i^* in latent study i conveys no information about the true parameter value $\Theta_{i'}^*$ in latent study i', and Θ_i^* is iid across i. The common parameters prior, by contrast, assumes that all latent studies attempt to estimate the same parameter Θ_0^*. Thus, priors in this class imply that Θ_i^* is perfectly dependent across i.

For both the unrelated and common parameters classes, the marginal prior π_{Θ^*} for Θ^* is unrestricted. For any π_{Θ^*} there is a unique prior in each class consistent with this marginal distribution.

If we observe a single draw X^*, our posterior for Θ^* depends only on the marginal prior π_{Θ^*}, and so is the same whether we consider the unrelated or common parameters priors. By contrast, when we observe a single draw X from the distribution of published papers, which class of priors we use turns out to be important. The following result is closely related to the findings of Yekutieli (2012).

Lemma 7 (Two posterior distributions)
Based on single observation of X, we obtain the following posteriors:

1. *Under unrelated parameters priors:*

$$f_{\Theta|X}(\theta|x) = f_{X^*|\Theta^*}(x|\theta) \cdot \pi_{\Theta^*}(\theta)/\pi_{X^*}(x)$$

2. *Under common parameters priors:*

$$f_{\Theta|X}(\theta|x) = \frac{p\left(x\right)}{E\left[p\left(X_i^*\right)|\Theta_i^* = \theta\right]} f_{X^*|\Theta^*}(x|\theta) \cdot \pi_{\Theta^*}(\theta)/\pi_{X^*}(x)$$

$$\propto f_{X|\Theta}(x|\theta) \cdot \pi_{\Theta^*}(\theta)$$

Proof of Lemma 7:

1. Unrelated parameters: By construction $D_i \perp \Theta_i | X_i^*$, and thus

$$\begin{aligned} f_{\Theta|X}(\theta|x) &= f_{\Theta_i^*|X_i^*,D_i}(\theta|x, d=1) \\ &= f_{\Theta_i^*|X_i^*}(\theta|x) \\ &= f_{X^*|\Theta^*}(x|\theta) \cdot \pi_{\Theta^*}(\theta)/f_{X^*}(x). \end{aligned}$$

2. Common parameters: This follows immediately from the truncated likelihood derived in Lemma 1 of the main text.

□

Under the unrelated parameters prior, our posterior $f_{\Theta|X}(\theta|x)$ after observing $X = x$ is the same as our posterior had we observed $X^* = x$. The form of $p(\cdot)$ has no effect on our posterior distribution, and inference proceeds exactly as in the case without selection. Under the common parameters prior, by contrast, our posterior $f_{\Theta|X}(\theta|x)$ corresponds to updating our marginal prior π_{Θ^*} using the truncated likelihood $f_{X|\Theta}(x|\theta)$ derived in Lemma 1.

The fact that selection has no effect on our posterior under the common parameters prior may be surprising, but reflects the fact that under this prior, selection changes the marginal prior π_Θ for true effects in published studies. In particular, under this prior we have

$$\pi_\Theta(\theta) = \frac{E\left[p\left(X_i^*\right)|\Theta_i^* = \theta\right]}{E\left[p\left(X_i^*\right)\right]}\pi_{\Theta^*}(\theta),$$

which reflects the fact that the distribution of true effects for published studies differs from that for latent studies under this prior. When we update this prior based on observation of X, however, the adjustment by $E\left[p\left(X_i^*\right)|\Theta_i^* = \theta\right]$ in the prior cancels that in the likelihood, and selection has no net effect on the posterior. Under the common parameters prior, by contrast, $\pi_{\Theta^*} = \pi_\Theta$, so the adjustment term in the prior due to selective inference continues to play a role in the posterior. For related discussion, see Yekutieli (2012).

I Optimal selection in a simple model

In the main text we discuss how to account for selective publication in inference and how to identify selectivity. It is natural to ask, however, whether selective publication is a good idea in the first place or just a misguided application of statistics leading to either publication bias or needlessly complicated inference. The answer to this question depends on the journal's objective function. One possible story views the published estimates as inputs into policy decisions, for instance in development economics, education, public finance, or medicine. If there are constraints on how many studies are published and read, then selectivity of the sort we observe might be justified.

We discuss a stylized version of this story in a development economics context, though our story might also be considered a stylized description of medical publishing and doctors' prescriptions of treatments for patients. Suppose that each i corresponds to a different policy intervention. Suppose the distribution μ of true treatment effects Θ_i^* is known to journal editors and readers, and that the expected effect $E[\Theta_i^*]$ of a randomly chosen treatment on the likelihood of escaping poverty is non-positive. Suppose further that the journal is read by policy makers who aim to minimize poverty. Assume finally that each treatment is relevant for a population of equal size, normalized to 1. A policy maker wishes to implement a given treatment j if the expected impact on the outcome considered is positive, conditional on the observed estimate $X_j = x$. Thus, their optimal treatment assignment rule is

$$t(x) = \mathbf{1}(E[\Theta_j | X_j = x] > 0), \tag{15}$$

resulting in expected outcome

$$v(x) = \max(0, E[\Theta_j | X_j = x]). \tag{16}$$

where $E[\Theta_j|X_j]$ is the policymakers' posterior expectation of Θ_j after observing X_j.[10] Suppose the journal also aims to minimize poverty, but faces a marginal (opportunity) cost of c, in units comparable to treatment outcomes, when publishing a given study. Policymakers update their behavior only for published studies with $E[\Theta_j|X_j] > 0$.

[10]Perhaps surprisingly, truncation is irrelevant for this posterior expectation. This stems from the fact that we assume policy makers have unrelated parameters priors as in Definition 4 above.

This updated behavior results in an expected poverty reduction of $E[\Theta_j|X_j]$ relative to the status quo. It follows that the optimal publication rule for the journal is

$$p(X_i^*) = \mathbf{1}(E[\Theta_i^*|X_i^*] > c). \tag{17}$$

If the conditional expectation is monotonic in X_i^*, this rule is equivalent to

$$p(X_i^*) = \mathbf{1}(X_i^* > x_c),$$

so that results should get published if they are positive "significant" at the critical value x_c, defined via $E[\Theta_i^*|X_i^* = x_c] = c$.

This result rationalizes selectivity in the publication process: the optimal rule derived here corresponds to one-sided testing. A more realistic version of this story allows for variation across i in the variance of X_i^*, the cost of implementing treatment, the size of the populations to be treated, etc. All of these would affect the critical value x_c, which thus should vary across i and need not be equal to conventional critical values of hypothesis tests. What remains true, however, is that publication decisions that are optimal according to the above model are selective in a way which leads to publication bias, and correct inference needs to account for this selectivity.