Identification of and correction for publication bias*

Isaiah Andrews† Maximilian Kasy‡

March 24, 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 of smaller magnitude or opposite sign than those in the initial studies (cf.

*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. This research was funded in part by the Silverman (1968) Family Career Development Chair at MIT.

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

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

Open Science Collaboration, 2015; Camerer et al., 2016). One leading explanation for replication failure is publication bias (cf. Ioannidis, 2005, 2008; McCrary et al., 2016; Christensen and Miguel, 2016). Journal editors and referees may be more likely to publish results that are statistically significant, 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 variance 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 based on 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 5.1 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.

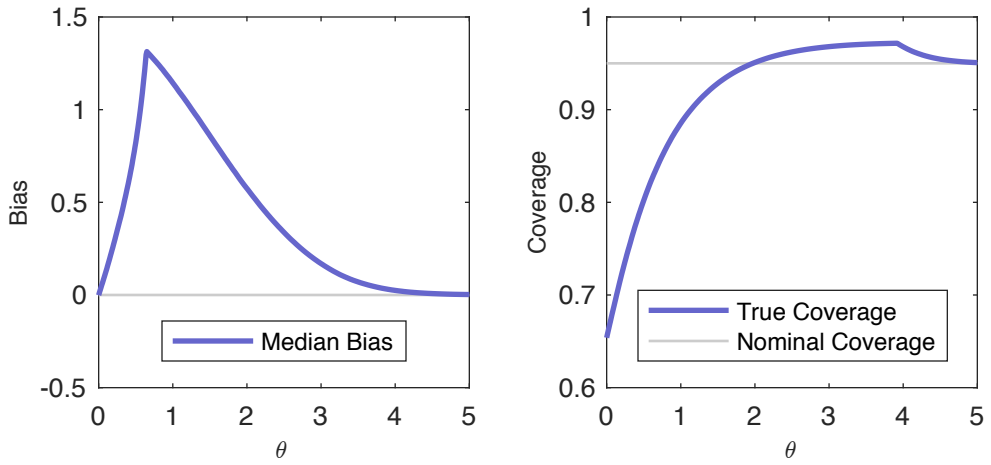

Figure 1: The left panel plots the median bias of the conventional estimator $\hat{\Theta}_j = Z_j$, while the right panel plots the true coverage of the conventional 95% confidence interval, both for $p(z) = .1 + .9 \cdot \mathbf{1}(|Z| > 1.96)$.

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 nonparametric 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

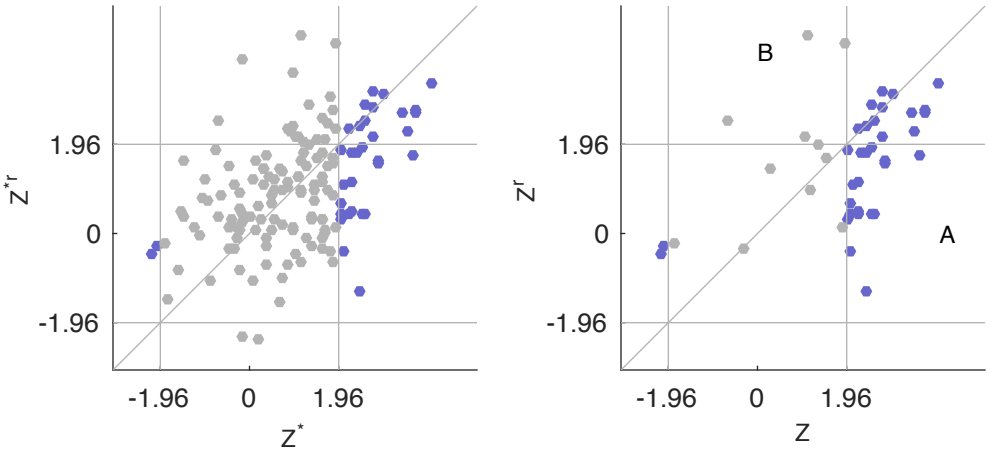

Figure 2: This figure illustrates the effect of selective publication in the replication experiments setting using simulated data, where selection is on statistical significance, as described in the text. The left panel shows the joint distribution of a random sample of latent estimates and replications;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

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

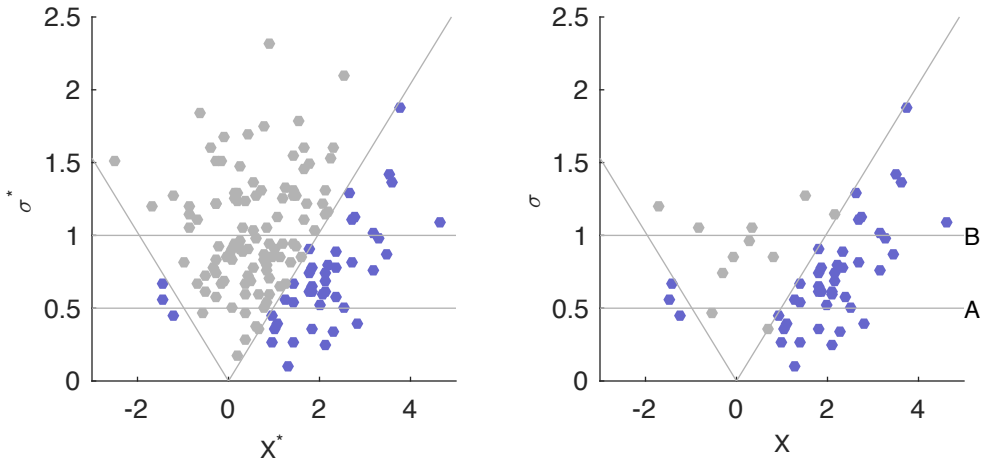

Figure 3: This figure illustrates the effect of selective publication in the meta-studies setting using simulated data, where selection is on statistical significance, as described in the text. The left panel shows a random sample of latent estimates; the right panel shows the subset of estimates which 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 nonparametrically 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 cannot 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

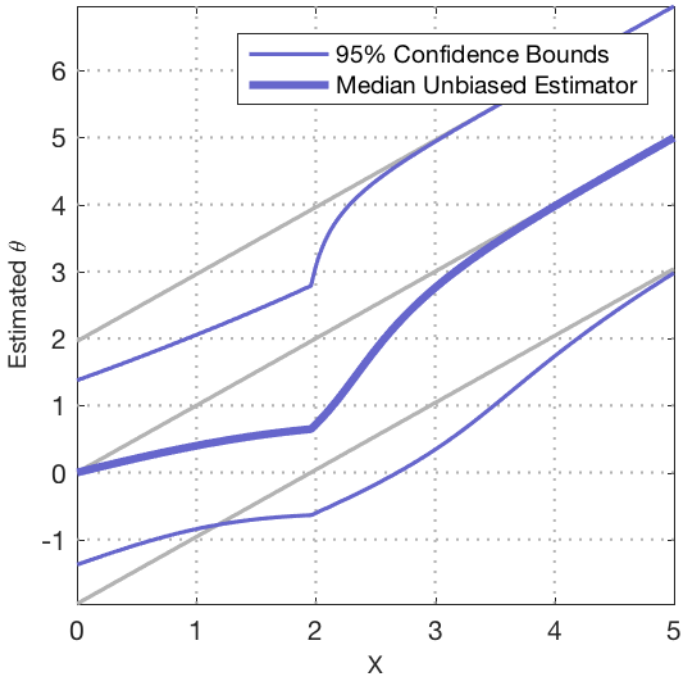

Figure 4: This figure plots frequentist 95% confidence bounds and the median unbiased estimator for the normal model where results that are significant at the 5% level are published with probability one, while insignificant results are published with probability 10%. The usual (uncorrected) estimator 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 need to be adjusted. 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 J 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 results imply nonparametric 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 consider step function models for $p(\cdot)$, with jumps at conventional critical values, and possibly at zero. We assume the latent effects Θ^* are normally distributed. In our first two applications, the sign of the original estimates is normalized to be positive. We denote these normalized estimates by $W = |Z|$, and in these settings we impose that $p(\cdot)$ is symmetric, and that the mean of Θ^* is zero.[2] Details and further motivation for these specifications, as well as a specification for the model of Section 3.1.3, are discussed in Section C of the supplement.

5.1 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] Further details on the

[2]Identification of the mean of Θ^* would be irregular in this setting, in the sense that the Fisher information for this parameter can be zero, yielding nonstandard asymptotic behavior for the maximum likelihood estimator. If we instead estimate this parameter, the MLE is zero in all specifications.

[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.

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.

A caveat to the interpretation of our results is that Camerer et al. (2016) select the most important statistically significant finding from each paper, as emphasized by the original authors, for replication. This selection changes the interpretation of $p(\cdot)$, which has to be interpreted as the probability that a result was published *and* selected for replication.

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 is confirmed by both our replication and meta-study approaches.

Results from replication specifications The middle panel of Figure 5 plots the joint distribution of W, W^r 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

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

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%

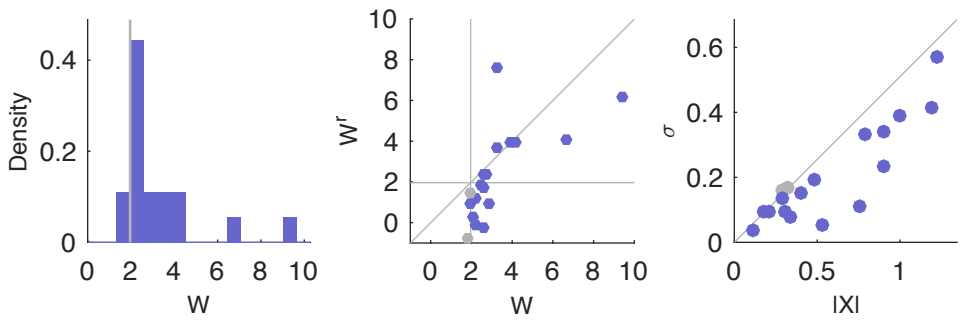

Figure 5: The left panel shows a binned density plot for the normalized z-statistics $W = |X|/\sigma$ using data from Camerer et al. (2016). The grey line marks $W = 1.96$. The middle panel plots the z-statistics W from the initial study against the estimate W^r from the replication study. The grey lines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

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 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.

level based on a two-sided z-test, relative to a result that is significant at the 5% level. These estimates therefore imply that significant results are 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 $p(z, \theta) = p(z)$, based on a model discussed in Section C.1 of the supplement, 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]

[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.

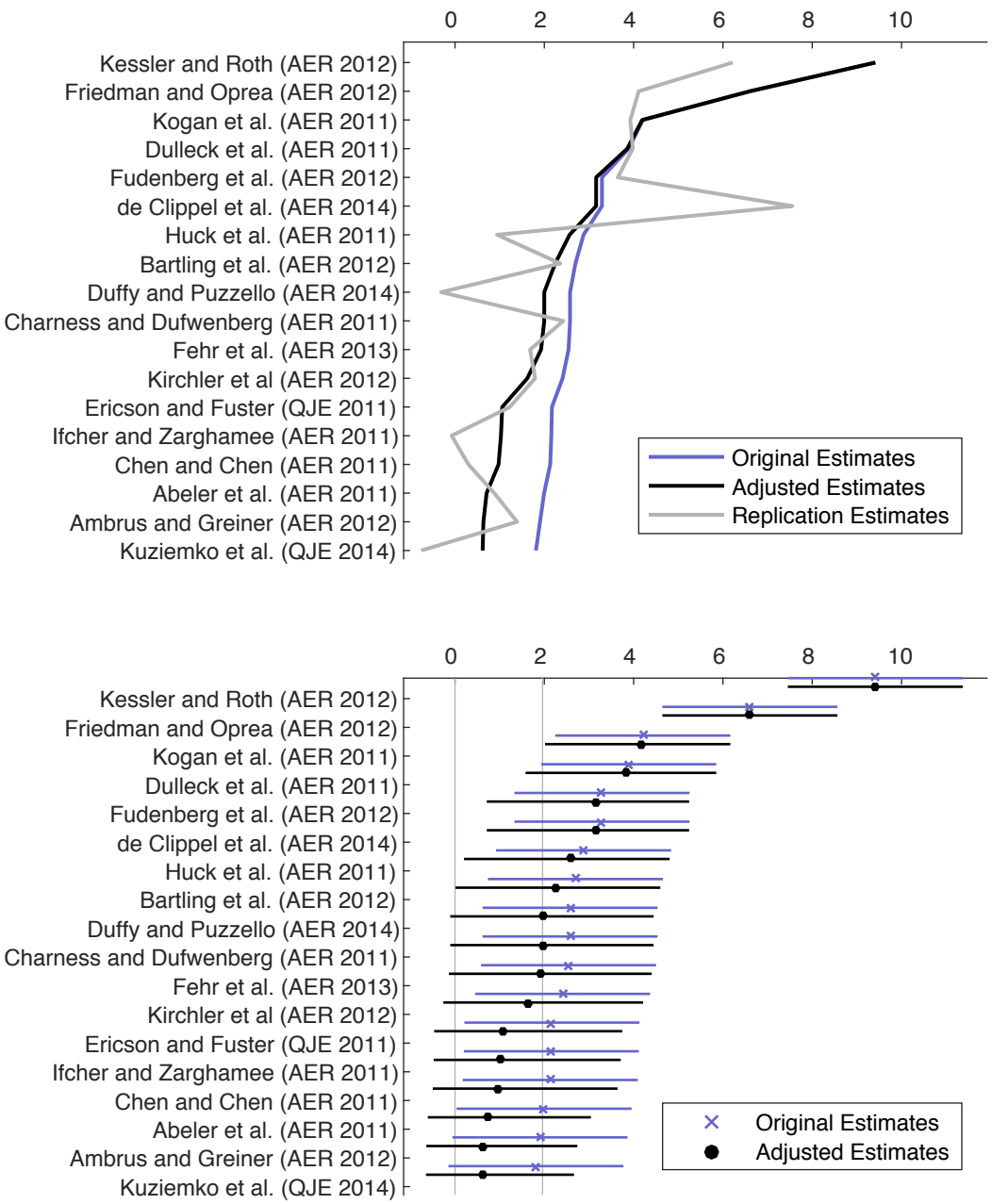

Figure 6: The top panel plots the estimates W and W^r from the original and replication studies in Camerer et al. (2016), along with the median unbiased estimate $\hat{\theta}_{\frac{1}{2}}$ based on the estimated selection model and the original estimate. The bottom panel plots the original estimate 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.

5.2 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.

As with the economics replications above, the systematic selection of results for replication in Open Science Collaboration (2015) is an advantage from our perspective. A complication in this setting 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.

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.

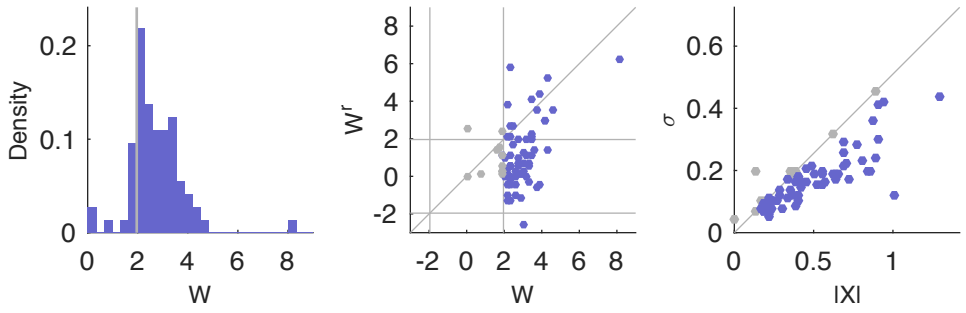

Figure 7: The left panel shows a binned density plot for the normalized z-statistics $W = |X|/\sigma$ using data from Open Science Collaboration (2015). The grey line marks $W = 1.96$. The middle panel plots the z-statistics W from the initial study against the estimate W^r from the replication study.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

Caveats The fact that not all available studies were selected for replication by Open Science Collaboration (2015) raises the possibility of selection of which studies to replicate, though 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 further complication in this setting 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

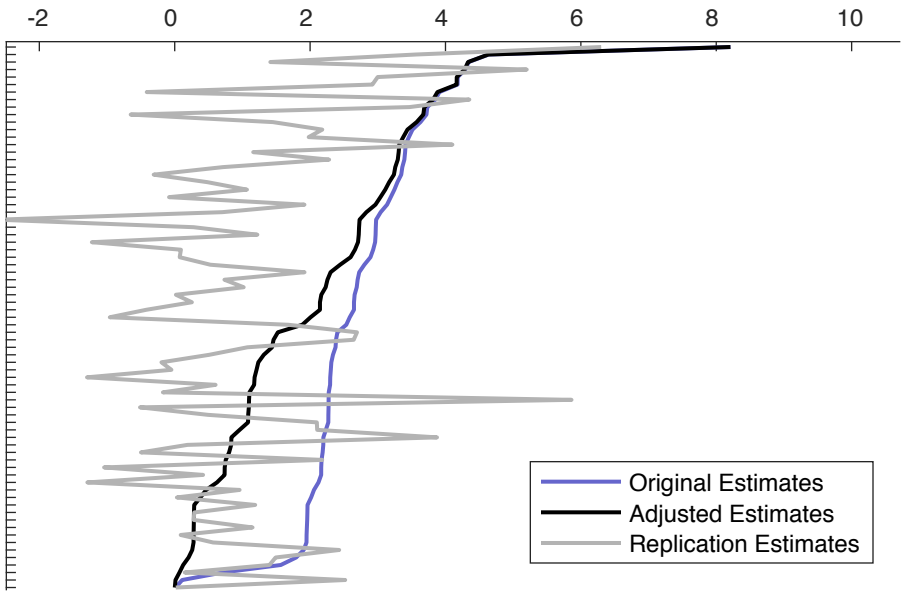

Figure 8: This figure plots the estimates W and W^r from the original and replication studies in Open Science Collaboration (2015), along with the median unbiased estimate $\hat{\theta}_{\frac{1}{2}}$ based on the estimated selection model and the original estimate.

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.

5.3 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 that use US data and were published or circulated as working papers after the year 2000. They collect estimates from all studies fitting their criteria that report both estimated elasticities of employment with respect to the minimum wage and standard errors, resulting in a sample of a thousand estimates drawn from 37 studies, and we use these estimates as the basis of our analysis. For further discussion of these data, see Wolfson and Belman (2015).

Since the Wolfson and Belman (2015) sample includes both published and un-

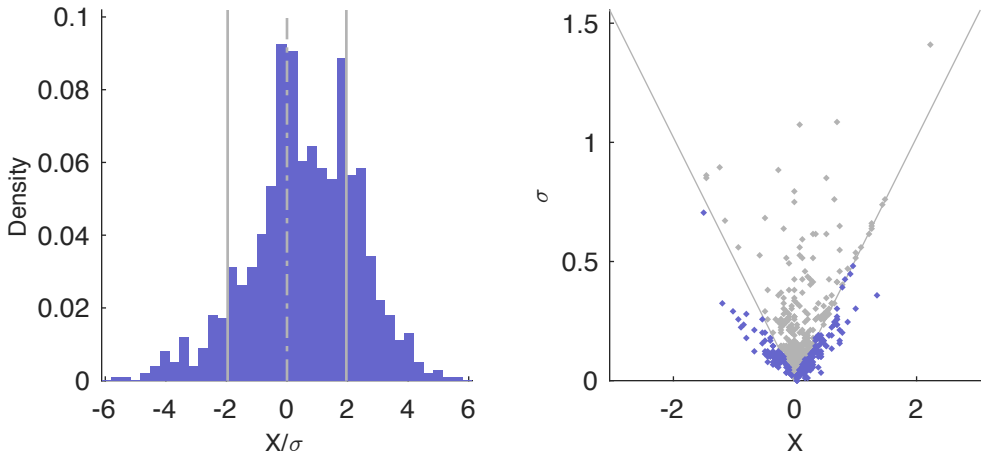

Figure 9: The left panel shows a binned density plot for the z-statistics X/σ in the Wolfson and Belman (2015) data. The solid grey lines mark $|X|/\sigma = 1.96$, while the dash-dotted grey line marks $X/\sigma = 0$. The right panel plots the estimate X against its standard error σ. The grey lines mark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

published 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.

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 finding 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.

$\bar{\theta}$	$\tilde{\tau}$	$\beta_{p,1}$	$\beta_{p,2}$	$\beta_{p,3}$
-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 standard errors clustered by study in parentheses. Publication probabilities β_p measured relative to omitted category of estimates positive and significant at 5% level.

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 interesting 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 $\bar{\theta}$ that we would estimate by ignoring selectivity, $\bar{\theta} = 0.038$ with a standard error of .025, suggesting a negative average estimate of the effect of minimum wages on employment.

Caveats 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 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.

5.4 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 estimates, 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.

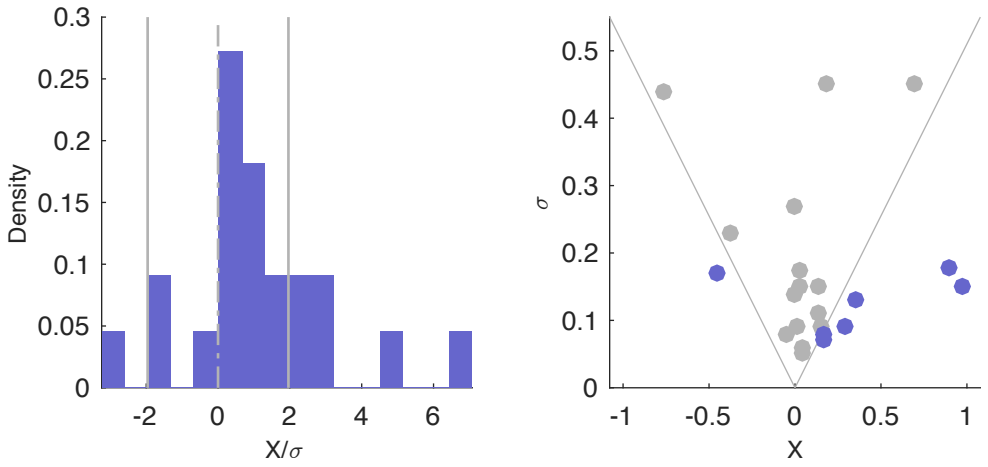

Figure 10: The left panel shows a binned density plot for the z-statistics X/σ in the Croke et al. (2016) data. The solid grey lines mark $|X|/\sigma = 1.96$, while the dash-dotted grey line marks $X/\sigma = 0$. The right panel plots the estimate X against its standard error σ. The grey lines mark $|X|/\sigma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

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.

$\bar{\theta}$	$\tilde{\tau}$	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.

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 suggest selection against negative estimates.

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.

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

[5] In progress.

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 Section J of 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 prior beliefs. The corresponding publication rule favors the publication of surprising findings, thus violating non-selectivity. 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.

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.