NBER WORKING PAPER SERIES

LITERATURE REVIEW AND EVIDENCE AGGREGATION: A TOOLKIT FOR APPLIED MICRO

Peter Ganong
Avik Garg
Maximilian Kasy

Working Paper 35403
http://www.nber.org/papers/w35403

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2026

We thank Anna Dreber Almenberg, Martha Bailey, Jonathan Cohen, Emma Duchini, Alex Frankel, Ryan Kellogg, Jack Mountjoy, Virginia Minni, Matt Notowidigo, Mikkel Plagborg-Møller, Charlie Rafkin, Shanthi Ramanth, Lutz Sager, and Witold Więcek for helpful conversations. We thank Giuseppe Cognata, Shravan Haribalaraman, Ari Jacob, Sedona Jolly, and Min Seo Kim for their excellent research assistance. Ganong thanks Pascal Noel for asking questions which were a source of inspiration to start working on this topic. We declare that we have no relevant or material financial interests that relate to the research described in this paper. No party had the right to review this paper prior to submission. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2026 by Peter Ganong, Avik Garg, and Maximilian Kasy. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Literature Review and Evidence Aggregation: a Toolkit for Applied Micro
Peter Ganong, Avik Garg, and Maximilian Kasy
NBER Working Paper No. 35403
July 2026
JEL No. C11, C53, C54, C87

ABSTRACT

Consider an analyst interested in predicting the size of an effect. She has identified a set of prior published studies of similar effects. We provide a toolkit for (i) summarizing the prior literature, (ii) making predictions of effects in new contexts, and (iii) correcting for the bias from selectivity in the prior literature. We illustrate these methods with empirical examples from labor, public, behavioral, environmental, and development economics. Some of the tools are relevant even when only three prior studies are available. We show how it is possible to use covariates to transparently make predictions for a new context by reweighting prior estimates. The mean effect—after correcting for selectivity—is between 12% and 21% of the simple mean in our empirical examples. We conclude with a cookbook for practitioners producing meta-analyses.

Peter Ganong
The University of Chicago
Harris School of Public Policy
and NBER
ganong@uchicago.edu

Avik Garg
The University of Chicago
avikg@uchicago.edu

Maximilian Kasy
University of Oxford
maximilian.kasy@economics.ox.ac.uk

Contents

1 Introduction **1**
1.1 Literature 4

2 Formal Setup **5**

3 Evidence aggregation **7**
3.1 Theory – Empirical Bayes 7
3.1.1 Parametric Empirical Bayes (normal latent distribution) . 8
3.1.2 Heavy tails and non-parametric empirical Bayes 9
3.2 Empirical examples 10
3.2.1 Average and dispersion – "small" meta-analyses 10
3.2.2 Average and dispersion – "large" meta-analyses 15
3.2.3 Non-parametric empirical Bayes 18

4 Evidence aggregation with covariates **19**
4.1 Theory 19
4.1.1 Meta-regressions 20
4.1.2 Prediction and Gaussian process priors 21
4.1.3 Hierarchical Bayes 23
4.1.4 Binary implementation decisions 24
4.2 Empirical examples 24
4.2.1 Regression interpretation: precision and latent heterogeneity 24
4.2.2 Predictive models, external validity, and the weights on prior studies 26

5 Selectivity **32**
5.1 Identification 32
5.1.1 Distribution of p-values (Elliott et al. (2022)) 33
5.1.2 Meta-studies 34
5.2 Estimation and empirical examples 42
5.2.1 Testing for selectivity 42
5.2.2 Estimating the latent distribution of effects in presence of selectivity 44

6 A model with selectivity and covariates **45**
6.1 Theory 45
6.2 Empirical example of a model with selectivity and covariates . . . 48

7 Cookbook **49**

Appendixes – for online publication only **58**

A Derivation of Tweedie's formula **59**

B "Small" meta-analysis **61**
B.1 Elasticity of Housing Prices with Respect to Air Pollution 61
B.2 Effect of Pay Transparency Laws on the Gender Pay Gap 61
B.3 Effect of Paid Family Leave on Mother's Earnings 61
B.4 Additional settings from AEJ 2025 where small sample meta-analysis is feasible 62

C Binary implementation decisions **66**

D Covariates **68**

E Smoothness and monotonicity of the p-curve **69**

F Density ratio in Cohen and Ganong (2026) **70**

G Density Estimation and Bandwidth Selection **72**

1 Introduction

Consider an analyst interested in estimating or predicting the size of an effect – the estimand θ_0. She has a dataset with estimates from several prior studies of similar causal effects with standard errors. We provide a toolkit for three goals that she might have:

1. She wants to compare a new estimate $\hat{\theta}_0$ to the prior literature.

2. She observes covariates X_i for both prior studies where estimates are available, plus a new context where no estimate is available, and wants to predict θ_0 for this new context.

3. She is worried that published estimates are "missing" estimates due to specification searching, p-hacking or publication bias, and wants to estimate how much selectivity is present. If it is present, she wants to estimate the distribution of effects adjusting for the bias from selectivity.

To develop this toolkit, we alternate between theoretical discussions of methods and practical case-studies. We rely on an empirical Bayes approach to provide a unified theoretical framework. This framework assumes that the estimands θ_i are drawn from some distribution μ, while the observed estimates are conditionally normal, $\hat{\theta}_i \sim N(\theta_i, \sigma_i^2)$. If there is selectivity, then observability of study i might depend on the Z-statistic $Z_i = \hat{\theta}_i/\sigma_i$. The distribution μ might be of independent interest, but more importantly it serves as a device to construct predictions for θ_0 for some new context.

To illustrate these methods, we present both small case studies with three to five prior estimates (drawing on Sager and Singer 2025, Blundell et al. 2025, Bailey et al. 2025), and larger case studies with dozens of estimates. The larger case studies are drawn from several fields of applied economics, studying the effect of active labor market policy (Card et al., 2018), unemployment insurance (Cohen and Ganong, 2026), nudges (DellaVigna and Linos, 2022), and unconditional cash transfers (Crosta et al., 2024).

Question 1: Comparing estimates to the prior literature This question appears in every applied micro paper that estimates a quantity also studied in prior work. The analyst begins by asking whether the estimates in the prior literature differ primarily because of sampling variation in $\hat{\theta}_i$ or because of between-study variation in underlying effects θ_i. Between-study variation can be measured by comparing the dispersion of estimates $\hat{\theta}_i$ to the standard errors σ_i. Then, she can compute the precision-weighted mean of the prior estimates.

An estimate of the mean and distribution of the underlying effects is useful in several scenarios. Suppose an analyst has a new estimate which addresses a bias in

the existing literature; she can test whether her new estimate could have emerged from the distribution of effects in the existing literature. Alternatively, suppose that the analyst has a new better-powered identification strategy and she wants to quantify the improvement in precision. Finally, suppose the prior literature contains meaningful between-study variation and the analyst has a new estimate to add to this literature. She can point to meaningful between-study variation as a motivation to produce estimates for additional contexts and to search for mechanisms that predict heterogeneity.

We illustrate that each of these three scenarios actually arises using papers published in *American Economic Journal: Economic Policy* (AEJ Policy) in 2025. In fact, these scenarios appear to emerge frequently. Of the 61 papers published in 2025 in AEJ Policy, 27 papers discuss prior estimates of similar targets from the literature. A crude lower bound is therefore that meta-analysis can be used—without the author finding any additional related studies—in nearly half of recently-published research in one prominent journal in applied micro.

In settings with dozens of estimates, an analyst can go further. If the underlying distribution of effects is normal, then the empirical Bayes leads to linear shrinkage estimators of effects. However, she can also use richer models which do not assume that the underlying effects are distributed normally. Using maximum likelihood estimation, they can estimate a t distribution to account for fat tails, or even use a nonparametric distribution to account for skewness or multi-modality. Empirically, we find that we can reject the normal model in favor of a richer model in three out of our four larger case studies.

Question 2: Predicting effects for new contexts The second question—what is the predicted effect θ_0 in a new context—can be addressed by meta-regression or a generalization using Gaussian Process priors. The classic linear meta-regression approach to this question involves simply regressing estimates on covariates and using the coefficients from the regression for a new set of covariate values. Meta-regression's prediction for the new context can be rewritten as a weighted average of the prior estimates. Reweighting is attractive because it gives the analyst a transparent understanding of how the existing studies contribute to a prediction for the new θ_0.

Gaussian Process priors provide a way to predict θ_0 which maintains transparency through weighted averages, but does not require the restrictive linear additivity assumption of meta-regression. Instead, the researcher chooses a hyperparameter which corresponds to how much they are willing to import insights from existing studies in distinct contexts. When a large number of prior estimates are available for the same context, meta-regressions and Gaussian Process priors make similar predictions. However, when a researcher wants to extrapolate to a new context with no prior studies or only a few prior studies, the distinction

between the methods becomes meaningful.

We illustrate the predictions from these methods using a dataset on the employment effects of active labor market policies from Card et al. (2018). Consider an analyst who seeks to predict the long-term effect of training for the long-term unemployed (LTU). This prediction exercise necessarily involves extrapolation, because the review does not contain any studies of this exact horizon by program by population cell. Fortunately, there are related studies of: (i) the long-term effects of training for a different population (*unemployment benefit recipients*) and (ii) the *short-term* effects of training for LTUs.

The analyst may rely on these studies interpreted through the lens of a meta-regression or a Gaussian Process prior. Meta-regression or a Gaussian Process prior with a long length scale—meaning that the analyst is willing to draw heavily on studies from contexts that are further away—deliver a precise posterior for the program's forecasted effect. The standard deviation for the posterior mean is 1 p.p. or less, which is quite small relative to the a posterior mean of roughly 10 p.p. In contrast, if the analyst uses a Gaussian Process prior with a short length scale, the distribution for the prediction is far more uncertain, with a posterior standard deviation of 7 p.p. We discuss how the researcher's preferred *economic* model of employment determination can be used to inform the *statistical* choice of the length scale hyperparameter.

Question 3: Detecting and correcting selectivity Turning to the third question, we finally examine selectivity, where reporting and publication decisions can depend on an estimate's Z-statistic. We review two types of approaches to selectivity. We first discuss methods for testing for the presence of selectivity based on the distribution of p-values across studies. We then turn to methods which leverage the richer information contained in the joint distribution of estimates $\hat{\theta}_i$ and standard errors σ_i. Leveraging this richer information allows us to not only test for selectivity, but also to estimate its extent, and to obtain effect distributions that correct for selectivity. We provide graphical intuition for identification in this setting.

We find that selectivity is ubiquitous and that accounting for selectivity can lead to dramatically different estimates of parameters of interest in the context of our four larger case studies. First, insignificant findings are much less likely to be published: the probability of publishing an insignificant estimate is between 5% and 28% as compared to a positive significant estimate. Second, failing to correct for selectivity leads researchers to overstate the mean effect. The mean effect—after correcting for selectivity—is between 12% and 21% of the simple mean across the four applications we study. Ignoring selectivity leads to severely biased conclusions about the underlying effects.

Roadmap The remainder of the article is structured as follows: Section 2 introduces the formal setup, Section 3 discusses aggregation, Section 4 discusses aggregation with covariates, Section 5 discusses selectivity, Section 6 discusses how to combine selectivity and covariates, and Section 7 provides a cookbook for practitioners. Before introducing our formal setup, the remainder of this section provides a brief review of some of the relevant methodological literature.

1.1 Literature

Key theoretical references for our review In the following, we draw on several literatures in statistical theory that provide a foundation for the more specific questions encountered in evidence aggregation. The *empirical Bayes* (or random-effects) framework that organizes our discussion throughout originates with Robbins (1956), Stein (1981), and Morris (1983); We rely on both parametric and non-parametric implementations: kernel deconvolution (Meister, 2009), convex programming for non-parametric maximum likelihood of the prior (Koenker and Mizera, 2014), Gaussian process priors for flexible extrapolation across covariates (Williams and Rasmussen, 2006), and Tweedie-type shrinkage of individual estimates (Efron, 2011). An approach closely related to empirical Bayes imposes priors on hyper-parameters, yielding *hierarchical Bayes* estimators, cf. Gelman et al. (2014, chapter 5); we discuss this approach in subsection 4.1.

A central focus of our review is *selective reporting and publication* of empirical findings. Concerns about selectivity in published economics research date back at least to Stanley and Jarrell (1989). Bias tests via meta-regression were introduced by Egger et al. (1997) and refined by Stanley et al. (2017). Brodeur et al. (2016) document bunching of reported test statistics just above conventional significance thresholds, an approach extended to a broader cross-section of identification strategies by Brodeur et al. (2020); Ioannidis et al. (2017) highlight the prevalence of severely underpowered findings and propose to focus on highly-powered studies. Elliott et al. (2022) provide non-parametric tests for selectivity based on violations of monotonicity of the p-curve. Likelihood-based methods that estimate selection from the joint distribution of estimates and standard errors are developed in Andrews and Kasy (2019); we build on these in Section 5. Christensen and Miguel (2018) survey the broader research-transparency movement concerned with selective reporting.

Meta-analysis in economics Quantitative research synthesis has a long tradition in economics, with early applications such as Card and Krueger (1995)'s synthesis of the minimum-wage literature. Methodological tools for meta-regression and the detection of publication selection are developed by Stanley (2008) and Stanley and Doucouliagos (2014), and surveyed in the practitioner's guide of Irsova

et al. (2023). A more recent wave exploits standardized designs and larger samples of studies, including the comparison of academic and at-scale nudge trials in DellaVigna and Linos (2022), the replication exercise of Camerer et al. (2016), and the generalizability analysis of Vivalt (2019).

Approaches beyond the scope of the present review There are a number of exciting and insightful developments in the recent literature on evidence aggregation which are beyond the scope of the present review:

An increasing set of studies, in particular in development and experimental economics, make micro-data available. In settings where this is the case, meta-studies can directly analyze the *pooled micro-data*, yielding insights beyond those available from aggregate statistics; see, e.g., (Bandiera et al., 2021; Meager, 2022; Kremer et al., 2023; Lund et al., 2024; Mullins, 2025).

Meta-regressions, as discussed in subsection 4.1, are often used to explore possible drivers of effect heterogeneity. Parts of the literature use priors for regression coefficients putting a point-mass at 0. This leads to *Bayesian model averaging* approaches (Irsova et al., 2023). In our review (and in line with the recommendations of Gelman et al. 2014), we instead focus on continuous priors that do not assume true sparsity in a correctly specified linear model.

A central focus of our review is prediction of effects θ_i for specific instances. A decision-theoretic foundation for the proposed approaches can be found in the theory of compound decision problems Zhang (2003). An active literature in theoretical econometrics explores the broader *decision-theoretic foundations* and ramifications of evidence aggregation, e.g., Manski (2020), Christensen et al. (2026), Ishihara and Kitagawa (2024), and Montiel Olea et al. (2026).

2 Formal Setup

Throughout, our review is based on the following framework. This framework allows us to provide a unified discussion of the issues involved in meta-studies, including information aggregation and extrapolation, as well as p-hacking and publication bias.

Suppose we are interested in a set of empirical studies indexed by $i \in 1, \ldots, n$. Corresponding to each of these studies there is an (unknown) estimand θ_i. This might for example be the true average treatment effect, if i is an experimental study, or it might be the elasticity of some behavioral relationship. Each study reports an estimate $\hat{\theta}_i$ of θ_i. Each study is furthermore characterized by study-level covariates X_i, which might include features such as sampling frame and site of the study, study design and identifying variation, treatment or policy characteristics, etc. We will use θ to denote $(\theta)_{i \in 1,\ldots,n}$ and use $\hat{\theta}, X$ analogously.

Not all studies are available to us when we conduct a meta-analysis. Some studies are never published, either because of researcher decisions (possibly reflecting issues such as p-hacking or specification searching), or because of reviewer and editor decisions (which might lead to publication bias). Some studies might also simply be excluded from the sampling frame of a meta-study, for instance based on the journal or year they were published in. We indicate observability and inclusion of a study in a meta-analysis by $D_i \in \{0, 1\}$, where $D_i = 1$ for studies that are included.

We assume that the sampling distribution of $\hat{\theta}_i$ is given by

$$\hat{\theta}_i | \theta_i, \sigma_i^2, X_i \sim N(\theta_i, \sigma_i^2),$$

where the standard error σ_i is known to us whenever a study is published. This assumption implies that $\hat{\theta}_i$ is unbiased for θ_i; alternatively, we can think of this assumption as simply *defining* the estimand θ_i as the expectation of $\hat{\theta}_i$. This assumption also implies that $\hat{\theta}_i$ is normally distributed. Normality is typically justified by asymptotic arguments based on the central limit theorem. The same asymptotic arguments also imply that we can treat σ_i^2 as known.

Under normality, it is useful to define the Z-statistic Z_i, the corresponding normalized parameter ω_i, and the p-value P_i for a one-sided test of the null hypothesis that $\theta_i = 0$, via

$$Z_i = \hat{\theta}_i / \sigma_i, \quad \omega_i = \theta_i / \sigma_i, \quad \text{and } P_i = \Phi(-Z_i),$$

where Φ is the cumulative distribution function of the standard normal distribution.

We furthermore assume that the estimands θ_i are themselves sampled from some unknown distribution μ across studies,

$$\theta_i \sim^{iid} \mu.$$

The distribution μ is the population counterpart to the sample distribution $\{\theta_1, \ldots, \theta_n\}$. The sample distribution is well-defined, regardless of the substantive relationship between the estimands of different studies i. We discuss the interpretation of the distribution μ in more detail below in Section 3.1.

We will at times rely on the stronger assumption $\theta_i | \sigma_i \sim^{iid} \mu$, so that the estimands vary independently of estimator precision across studies. For the most part, this assumption only serves as a convenient simplification for the construction of estimators. The one point where this assumption is truly required is identification of selection models using meta-studies, discussed in section 5.

The observability D_i of studies, which reflects reporting and publication decisions, can in principle depend on $\hat{\theta}_i, \sigma_i$, and X_i in complicated ways. We will

however assume that the probability of being observable only depends on the Z-statistic Z_i,

$$P(D_i = 1|\hat{\theta}_i, \sigma_i) = \bar{d}(Z_i).$$

This is a restrictive assumption, which is maintained in much of the literature on p-hacking and publication bias (cf. Section 5).

3 Evidence aggregation

3.1 Theory – Empirical Bayes

We start our methodological review by considering methods for evidence aggregation. For clarity of exposition, throughout this section we assume that there is no selective reporting, so that $\bar{d} \equiv 1$. The natural framework for discussing evidence aggregation is the empirical Bayes framework, which was first introduced by Robbins (1956). The approach, as applied to our setting, can be summarized as follows:

1. Specify a model for the sampling distribution of $\hat{\theta}_i|\theta_i, \sigma_i^2, X_i$. In our context, we have $\hat{\theta}_i|\theta_i, \sigma_i^2, X_i \sim N(\theta_i, \sigma_i^2)$.

2. Specify a model for μ (the "latent distribution") for the distribution of estimands θ_i across i, possibly conditional on observed covariates X_i. This model might be parametric or non-parametric (unrestricted); we will consider both cases.

3. Based on the observed distribution of $(\hat{\theta}_i, \sigma_i, X_i)$, estimate the distribution μ of θ_i (possibly conditional on X_i and σ_i), by maximizing the marginal likelihood

   $$f(\hat{\theta}_i|\sigma_i^2, X_i) = \int \tfrac{1}{\sigma_i}\varphi\left(\tfrac{\hat{\theta}_i - \theta_i}{\sigma_i}\right) d\mu(\theta_i|X_i, \sigma_i),$$

   where φ is the standard normal density. Alternatively, one might use method of moments estimators that are based on this marginal distribution.

4. Lastly, form study-level estimates of θ_i using the posterior mean of θ_i, $E[\theta_i|\hat{\theta}_i, \sigma_i^2, X_i]$. This posterior mean is calculated using the "prior" μ estimated in step 3.

   Additionally, in line with the goals listed under items 1 and 2 in our introduction, form predictions for a new problem instance $i = 0$ given X_0, $P(\theta_0|X_0)$, which might in turn be used for a downstream decision problem; we will discuss examples later.

Many meta-studies stop at step 3 of this recipe, characterizing properties such as the expectation $E[\theta_i]$ (the average effect), the variance $Var(\theta_i)$ (effect heterogeneity), or the covariance $Cov(\theta_i, X_i)$ (predictors of effect heterogeneity). Step 4 in this recipe is, however, necessary whenever the goal is to solve a specific policy problem, where policies are chosen for a specific site i, rather than across all possible sites.

Viewed through the lens of these four steps, the distribution μ is best thought of as a conceptual device. It is an input to the construction of empirical Bayes estimators, decision procedures in compound decision problems, and as a device to discuss questions of external validity. However, the justification of these procedures does not necessarily rely on the family of priors. In addition, we note that empirical Bayes has provably good properties for such compound decision problems, *conditional* on the θ_i; see for instance Stein (1981).[1] For a general review, see Zhang (2003). The primary object of interest for both analysts and policymakers will typically be the effect θ_0 for new instances, rather than the distribution μ itself.

We start by considering the parametric empirical Bayes approach absent covariates, with a normal family of priors (settings with covariates are discussed in section 4 below). In doing so, we follow Morris (1983); see also Rubin (1981), and DerSimonian and Laird (1986). We next generalize by allowing that μ has heavy tails, considering a family of t-distributions. We can finally leave the distribution of θ_i fully unrestricted.

3.1.1 Parametric Empirical Bayes (normal latent distribution)

Let us first consider the simplest empirical Bayes setting, following Morris (1983), where we ignore covariates and assume that the effects θ_i are normally distributed across studies, independently of σ_i, so that μ is given by

$$\theta_i|\sigma_i \sim N(\bar{\theta}, \tau^2).$$

Under this assumption, the marginal distribution of $\hat{\theta}_i$, integrating out the latent θ_i, is given by

$$\hat{\theta}_i|\sigma_i \sim N(\bar{\theta}, \tau^2 + \sigma_i^2).$$

[1] Compound decision problems involve repeated decisions across the instances i, and loss that is averaged across these instances. A leading example is estimation of the vector $(\theta_1, \ldots, \theta_n)$, with mean squared error $\frac{1}{n}\sum_i(\hat{\theta}_i - \theta_i)^2$.

A simple method-of-moments estimator of $\bar{\theta}$ is given by $\hat{\bar{\theta}} = \frac{1}{n}\sum_i \hat{\theta}_i$. A simple estimator of τ^2 is given by

$$\hat{\tau}^2 = \max\left(\frac{1}{n}\sum_i \left[\left(\hat{\theta}_i - \hat{\bar{\theta}}\right)^2 - \sigma_i^2\right], 0\right) \tag{1}$$

The efficiency of the estimator of $\bar{\theta}$ could then be further improved by using a precision-weighted average,

$$\hat{\bar{\theta}}' = \frac{1}{\sum_i \frac{1}{\hat{\tau}^2+\sigma_i^2}} \cdot \sum_i \frac{\hat{\theta}_i}{\hat{\tau}^2 + \sigma_i^2}. \tag{2}$$

Alternatively, both $\bar{\theta}$ and τ^2 can be estimated simultaneously using maximum likelihood. In the absence of any further information about a new instance θ_0, we can think of $\hat{\bar{\theta}}$ as a predictor for θ_0.

To estimate θ_i for individual instances i where we do have an estimator $\hat{\theta}_i$, we can then apply step 4 of the empirical Bayes recipe, where

$$E[\theta_i|\hat{\theta}_i, \sigma_i] = \bar{\theta} + \frac{\tau^2}{\tau^2 + \sigma_i^2} \cdot (\hat{\theta}_i - \bar{\theta}). \tag{3}$$

Substituting estimates for $\bar{\theta}$ and τ^2 into this expression yields the empirical Bayes estimator of θ_i. It is interesting to note that in the special case where σ_i^2 is constant across i, the empirical Bayes approach essentially recovers the James-Stein shrinkage estimator (up to a small degrees of freedom correction). This shrinkage estimator is known to uniformly dominate $(\hat{\theta}_i)_{i=1}^n$ as estimator of the vector of θ_i's; see Stein (1981) for a proof.

3.1.2 Heavy tails and non-parametric empirical Bayes

The normality assumption $\theta_i \sim N(\bar{\theta}, \tau^2)$ underlying equation (3) is restrictive: in many applications the latent distribution μ has heavy tails, is skewed, or is multi-modal. We can generalize our baseline model to relax this assumption.

First, one can take μ to be a (possibly scaled and shifted) t-distribution with ν degrees of freedom,

$$\theta_i \mid \sigma_i \sim t_{df}(\bar{\theta}, \tau^2)$$

which allows for heavier tails than the normal; smaller ν implies heavier tails.

Second, one can leave μ entirely unrestricted. In this case, μ can be estimated by nonparametric maximum likelihood (Koenker and Mizera, 2014) or by deconvolution (Meister, 2009), and the posterior mean of θ_i admits a model-free

expression known as *Tweedie's formula* (Efron, 2011):

$$E\left[\theta_i \,\middle|\, \hat{\theta}_i, \sigma_i\right] = \hat{\theta}_i + \sigma_i^2 \cdot \frac{\partial}{\partial \hat{\theta}_i} \log f(\hat{\theta}_i \,|\, \sigma_i), \tag{4}$$

where $f(\hat{\theta}_i \,|\, \sigma_i)$ is the marginal density of $\hat{\theta}_i$ given σ_i. Intuitively, the posterior mean adjusts the raw estimate $\hat{\theta}_i$ in the direction of higher marginal density: the magnitude and direction of shrinkage are determined by the local shape of f rather than by a global normality assumption. We illustrate the empirical relevance of this generalization, which allows us to appropriately adjust for heavy tails, skewness, and multi-modality, in Section 3.2.3. The derivation of Tweedie's formula and its connection to deconvolution are given in Appendix A.

3.2 Empirical examples

3.2.1 Average and dispersion – "small" meta-analyses

The primary focus of many applied micro papers is to credibly estimate a quantity of interest and then to interpret the new estimate in part by comparing it to prior work. To illustrate how meta-analysis might be able to add additional insights to this type of exercise, we start from the studies which a paper *already cites* as estimating a similar object and ask what additional insights can be extracted based on applying the equations from Section 3.1.

We distinguish between three uses of meta-analysis. In the first example, an analyst has what they believe to be a superior methodology and wants to know if their estimate is meaningfully different from the prior literature. In the second example, an analyst has a new estimate and they want to construct the most likely estimate for the true effect, taking into account the distribution of effects in the prior literature. In the third example, an analyst has a new estimate and they want to quantify the improvement in precision.

Case Study #1: Correcting Bias Sager and Singer (2025) explore different strategies to estimate the effect of pollution on housing prices, with an eye towards correcting bias in the methods used in the prior literature. They describe their paper as contributing three insights to the literature on pollution damages. After describing the first two insights, they write

> Our third insight is that pollution damages might be even larger than previously thought... The implied elasticity [of house prices] with respect to [pollutants] $PM_{2.5}$ of around -1.4 is around twice that found for PM_{10} (Bento et al., 2015) and up to four times the elasticity for

Total Suspended Particles (TSP or PM_{100}) (Chay and Greenstone, 2005).

Table 1: Elasticity of Housing Prices with Respect to Air Pollution

Study	Estimate	SE
Panel A: Prior estimates (n=3)		
Chay and Greenstone (2005)	-0.27	(0.10)
Bento et al. (2015)	-0.60	(0.23)
Grainger (2012)	-0.37	(0.15)
Panel B: Mean of prior estimates		
Simple mean	-0.41	(0.10)
Precision-weighted mean $\hat{\bar{\theta}}'$ (equation 2)	-0.34	(0.08)
Panel C: Dispersion of prior estimates		
Variance of estimates	0.019	–
Average squared standard error	0.028	–
Between-study std. dev. $\hat{\tau}$ (equation 1)	0.00	–
Panel D: Summary of prior literature		
$\theta_i \sim N(-0.34, 0.00)$		
Panel E: New estimate		
Sager and Singer (2025)	-1.44	(0.35)
p-value that new estimate drawn from $N(-0.34, 0.00)$	<0.01	–

Note: The variance of the estimates is $\frac{1}{n}\sum_i \left(\hat{\theta}_i - \bar{\theta}\right)^2$, the average squared standard error is $\frac{1}{n}\sum_i \sigma_i^2$, and equation (1) is the former minus the latter, subject to a floor of 0. In Panel D, $N(\cdot, \cdot)$ reports the mean and standard deviation of the fitted distribution of treatment effects. See appendix B.1 for details on how estimates in panels A and E are calculated.

A meta-analytic framework offers additional lessons which complement the discussion in Sager and Singer (2025). With the estimates from Table 1, one could therefore have expanded on the "third insight" discussion in that paper, adding

> A meta-analysis of the prior literature indicates that the existing estimates are consistent with a homogeneous treatment effect of pollution with an elasticity of -0.34. Our elasticity estimate of -1.4 is significantly larger (more negative) than the prior literature. Endogeneity issues in identification strategies used in the prior literature lead them

> to understate the importance of pollution for affecting housing prices. However, we note that the standard error for our estimate is larger than much of the prior literature; had our design found an elasticity similar to the prior literature, we would be unable to reject the hypothesis that our estimate is drawn from the distribution of prior estimates.

Case Study #2: Best Estimate of True Effect Blundell et al. (2025) is interested in estimating the effect of pay transparency laws on the gender pay gap. They study a 2018 UK law that requires firms with over 250 employees to publicly disclose "gender equality indicators." When discussing their paper in the context of the broader literature, they mention three other papers that attempt to estimate a similar quantity, writing:

> The closest studies to ours are Bennedsen et al. (2022); Baker et al. (2023); and Gulyas et al. (2023)... Both Bennedsen et al. (2022) and Baker et al. (2023) find that transparency leads to pay compression by slowing down men's wage growth. In contrast, Gulyas et al. (2023) find no impact on individuals' wages or the gender pay gap.

Meta-analysis yields additional insights which help to put this paper's findings in context. Unlike in Sager and Singer (2025), Blundell et al. (2025) believes the prior literature to be relevant to their context, so this new estimate is the most efficient estimate available. With the estimates from Table 2 one could therefore have expanded on the discussion in that paper, adding

> A meta-analysis of the prior literature indicates that of the four disclosure laws studied (including ours), the mean effect is -11.6% and and the standard deviation across studies is 7.6%. This is useful for two purposes. First, the fact that there is substantial variation across studies in whether disclosure laws reduce gender pay gaps motivates an analysis in our paper of the mechanisms which give rise to variation in effectiveness. We focus in particular on a binary covariate:whether the law requires *public* disclosure or simply requires disclosure to employees. Second, combining our new study with what is already known from the prior studies in the literature, we estimate that the effect of the 2018 UK law was -14.9%. This is lower than our econometric estimate because the Empirical Bayes framework shrinks our estimate of -18.8% toward the mean estimate from the literature as a whole.

The three examples in this section are drawn from articles published in *American Economic Journal: Economic Policy* in 2025. To assess the generalizability

Table 2: Effect of Pay Transparency Laws on the Gender Pay Gap

Study	Estimate	SE
Panel A: Prior and new estimates (n=4)		
Baker et al. (2023)	-27.2	(8.3)
Bennedsen et al. (2022)	-13.0	(5.2)
Gulyas et al. (2023)	0.0	(0.6)
Blundell et al. (2025) (new estimate)	-18.8	(8.1)
Panel B: Mean of estimates		
Simple mean	-14.8	(4.9)
Precision-weighted mean $\hat{\bar{\theta}}'$ (equation 2)	-11.6	(4.7)
Panel C: Dispersion of estimates		
Variance of estimates	98.0	–
Average squared standard error	40.5	–
Between-study std. dev. $\hat{\tau}$ (equation 1)	7.6	–
p-value that $\hat{\tau} = 0$	<0.01	–
Panel D: Summary of literature		
$\theta_i \sim N(-11.6, 7.6)$		
Panel E: Forecast for new estimate		
Forecast $E[\theta_0 \mid \hat{\theta}_0, \sigma_0]$ (equation 3)	-14.9	(5.5)

Note: See appendix B.2 for details on how estimates in panel A are calculated.

of the examples listed here, we reviewed all 61 articles published in the journal in 2025. In essentially all of these articles, the primary object of interest is a causal effect; however, many of the articles do not directly compare their estimates to prior papers or only compare their estimates to a single prior paper. Table 8 lists 27 papers which compare their new estimates to at least two prior estimates from the literature. A crude lower bound is therefore that meta-analysis can be used—without identifying any additional related studies—in almost half of high-quality research in one journal in applied micro.

Case Study #3: Improving Precision Bailey et al. (2025) is interested in the effect of paid family leave on mothers' earnings and employment. They use a regression discontinuity design in California to estimate the impact of this policy. Describing the prior literature, they write

> For the United States, empirical evidence regarding the employment [and earnings] effects of these policies has been inconclusive owing to

small sample sizes and incomplete data. While some studies suggest that paid leave improves women's short-term career outcomes (Rossin-Slater, Ruhm, and Waldfogel, 2013; Campbell, Chin, Chyn, and Hastings, 2018), estimates tend to be imprecise... [description of data and research design...] These microdata are around 200 times larger than publicly available surveys, which increases the precision of the estimates...

Table 3: Effect of Paid Family Leave on Mothers' Short-Run Earnings

Study	Estimate	SE
Panel A: Prior estimates (n=4)		
Rossin-Slater, Ruhm, and Waldfogel (2013)	7.5	(4.0)
Campbell, Chin, Chyn, and Hastings (2018)	10.2	(10.7)
Baum and Ruhm (2016)	3.6	(3.4)
Timpe (2024)	-5.6	(2.4)
Panel B: Mean of prior estimates		
Simple mean	3.9	(3.0)
Precision-weighted mean $\hat{\bar{\theta}}'$ (equation 2)	-0.3	(1.7)
Panel C: Dispersion of prior estimates		
Variance of estimates	35.9	–
Average squared standard error	37.2	–
Between-study std. dev. $\hat{\tau}$ (equation 1)	0.0	–
Panel D: Summary of prior literature		
$\theta_i \sim N(-0.3, 0.0)$		
Panel E: New estimate and updated literature (n=5)		
Bailey, Byker, Patel, and Ramnath (2025) (new estimate)	-0.8	(0.8)
Precision-weighted mean incl. new $\hat{\bar{\theta}}'$ (equation 2)	-0.3	(1.3)

Note: Estimates are for the effect of paid leave availability. See appendix B.3 for details on how estimates in panel A are calculated.

Table 3 conducts a meta-analysis for the effect of offering paid family leave on earnings. With such a meta-analysis in hand, the discussion might have continued

Our estimates reduce statistical uncertainty about the effects of paid family leave because our data and research design enable a substantial improvement in precision. The standard error for our research design is 0.8%, while the standard errors for the prior studies ranges from 2.4% to 10.7%. As a result, the standard error for the precision-weighted

mean including our design is 1.3%, whereas it would be 1.7% without our study. Because our estimate of -0.8% is so similar to the mean of the existing estimates, adding our study to the pool leaves the mean estimate unchanged (up to one decimal point).

The simple meta-analyses done in this section are likely to generalize to other research in applied micro. All three examples are drawn from articles published in *American Economic Journal: Economic Policy* in 2025. To assess the generalizability of the examples listed here, we reviewed all 61 articles published in the journal in 2025. In essentially all of these articles, the primary object of interest is a causal effect; however, many of the articles do not directly compare their estimates to prior papers or only compare their estimates to a single prior paper.

Appendix B.4 lists 27 papers which compare their new estimates to at least two prior estimates from the literature. A crude lower bound is therefore that meta-analysis can be used—without identifying any additional related studies—in almost half of high-quality research in one journal in applied micro.

3.2.2 Average and dispersion – "large" meta-analyses

Applications In the remainder of the paper, we re-analyze datasets from four existing meta-analyses: Card et al. (2018), Cohen and Ganong (2026), Crosta et al. (2024), and DellaVigna and Linos (2022). These meta-analyses cover four different subfields of economics – public, development, behavioral, and labor – which allows us to emphasize considerations specific to different context. Unlike the "small" meta-analyses in the previous section, each of these meta-analyses is based on a relatively large number of studies and uses covariates to predict treatment effects. The larger n and the covariates allow us to use richer methodologies to extract additional insights.

Cohen and Ganong (2026) analyzes the effect of unemployment benefit generosity on unemployment duration using 93 independent estimates from 57 studies, most of which are based on natural experiments. There are no randomized controlled trial (RCT) experiments in the sample, and a few older studies rely on a selection-on-observables assumption rather than a natural experiment. The paper finds that insignificant results are about 10% as likely to be published as significant results. The paper also finds that the baseline replacement rate for unemployment benefits is a meaningful predictor of the benefit duration elasticity, a point to which we return in Section 6.2.

Card et al. (2018) analyzes the effect of active labor market policies (ALMPs) on employment. About one-fifth of the studies the paper considers use an RCT design, and more than half of the studies occurred in Europe. The paper has an extensive heterogeneity analysis by covariates, which we revisit in Section 4.2.1.

Standard errors are missing from many of the studies in the original meta-analysis. We drop those studies for a final sample of 169 estimates from 45 studies.

DellaVigna and Linos (2022) analyzes the effect of nudges on a wide range of outcomes. The paper considers the effect of RCTs from two different samples: 74 estimates that were published in academic journals and 241 estimates from studies conducted by two government-run "nudge units." DellaVigna and Linos (2022) find substantial publication bias in the sample of academic studies, relative to the nudge unit sample. We will combine both the nudge unit and academic journal estimates. In the selectivity section, we will assume no selectivity, $\bar{d} \equiv 1$, for estimates from the nudge unit subsample.

Crosta et al. (2024) analyzes the effect of unconditional cash transfer programs in low- and middle-income countries on several different outcomes. The paper exclusively considers RCT studies. Crosta et al. (2024) concludes, based on 75 estimates, that for each additional $100 of transfer, recipients have monthly consumption that is $3.30 higher.

Evidence aggregation Table 4 illustrates four lessons about evidence aggregation. First, for all four applications, there is substantial dispersion in estimated treatment effects beyond what would be expected based only on standard errors. This is reflected in the fact that we estimate $\tau^2 > 0$ both when using a normal distribution or a t distribution for the latent distribution of estimates.

Second, using the MLE normal (or equation 2) provides a principled way to reduce weight on noisy outlier estimates. This is most easily seen in the context of Crosta et al. (2024), where the simple mean and MLE normal deliver meaningfully different conclusions. The simple mean implies that a $1 transfer raises consumption by $0.13 while MLE normal implies that $1 transfer raises consumption by $0.03. This pattern arises because Crosta et al. (2024) is a setting where a few estimates have unusually large treatment effects and standard errors. The simple mean weights these outliers equally to the other estimates. Indeed, when the 10 percent of estimates with the largest standard errors are excluded, the method of moments estimator shrinks to $0.04, in line with the MLE normal estimate.

Third, MLE estimates of the mean can be much more precise than the method of moments estimates. The precision advantages of the MLE are greatest when there is (i) variation in sampling uncertainty (σ_i^2) and (ii) between-study variation is much smaller than sampling uncertainty ($\tau^2 << \sigma_i^2$). This is illustrated in the estimates for Crosta et al. (2024), which meets both of these conditions. At the opposite extreme, when $\tau^2 >> \sigma_i^2$, as in DellaVigna and Linos (2022), the weights on each study will be similar, leading MLE normal to yield a similar standard error to the method of moments. Standard errors for MLE estimators are calculated using a sandwich estimator. We include standard errors for $\hat{\tau}$ in the "large" meta-analyses in Table 4 but not in the "small" meta-analyses in Tables 1 and 2 since

Table 4: Parameter Estimates of Distribution of Treatment Effects

Treatment Effect & Dataset	Method	$\bar{\theta}$	τ	df
Elasticity of Unemp Duration w.r.t. Unemp Benefits Cohen and Ganong (2026)	Simple Mean	0.513 (0.050)	–	–
	Prec-wtd. Mean (eqn 2)	0.448 (0.044)	0.345 (0.059)	–
	MLE (normal)	0.447 (0.050)	0.338 (0.053)	–
	MLE (t)	0.382 (0.046)	0.240 (0.032)	4.04 (1.69)
Effect of Active Labor Market Program on Employment Card et al. (2018)	Simple Mean	0.083 (0.008)	–	–
	Prec-wtd. Mean (eqn 2)	0.040 (0.007)	0.000 (2.615)	–
	MLE (normal)	0.074 (0.008)	0.086 (0.007)	–
	MLE (t)	0.057 (0.020)	0.064 (0.023)	4.11 (4.13)
Effect of $1 Transfer on Monthly Consumption Crosta et al. (2024)	Simple Mean	0.127 (0.049)	–	–
	Prec-wtd. Mean (eqn 2)	0.011 (0.004)	0.000 (2.212)	–
	MLE (normal)	0.028 (0.004)	0.020 (0.003)	–
	MLE (t)	0.028	0.020	∞
Effect of Nudge on Targeted Outcome DellaVigna and Linos (2022)	Simple Mean	0.031 (0.004)	–	–
	Prec-wtd. Mean (eqn 2)	0.026 (0.003)	0.070 (0.002)	–
	MLE (normal)	0.024 (0.004)	0.050 (0.008)	–
	MLE (t)	0.004 (0.001)	0.005 (0.001)	1.00 (0.084)

Notes: This table shows estimates of the distribution of treatment effects for four empirical applications. $\bar{\theta}$ is the median for all three distributions. Standard errors are shown in parentheses. For Precision Weighted Mean and Maximum Likelihood Estimate (MLE) normal, the variance is τ^2. For MLE t, the variance is $\tau^2 \frac{df}{df-2}$ if $df > 2$ and ∞ otherwise. For Precision Weighted Mean, we use the τ^2 formula from equation 1. We do not show standard errors with method MLE t for Crosta et al. (2024) since they are numerically unstable.

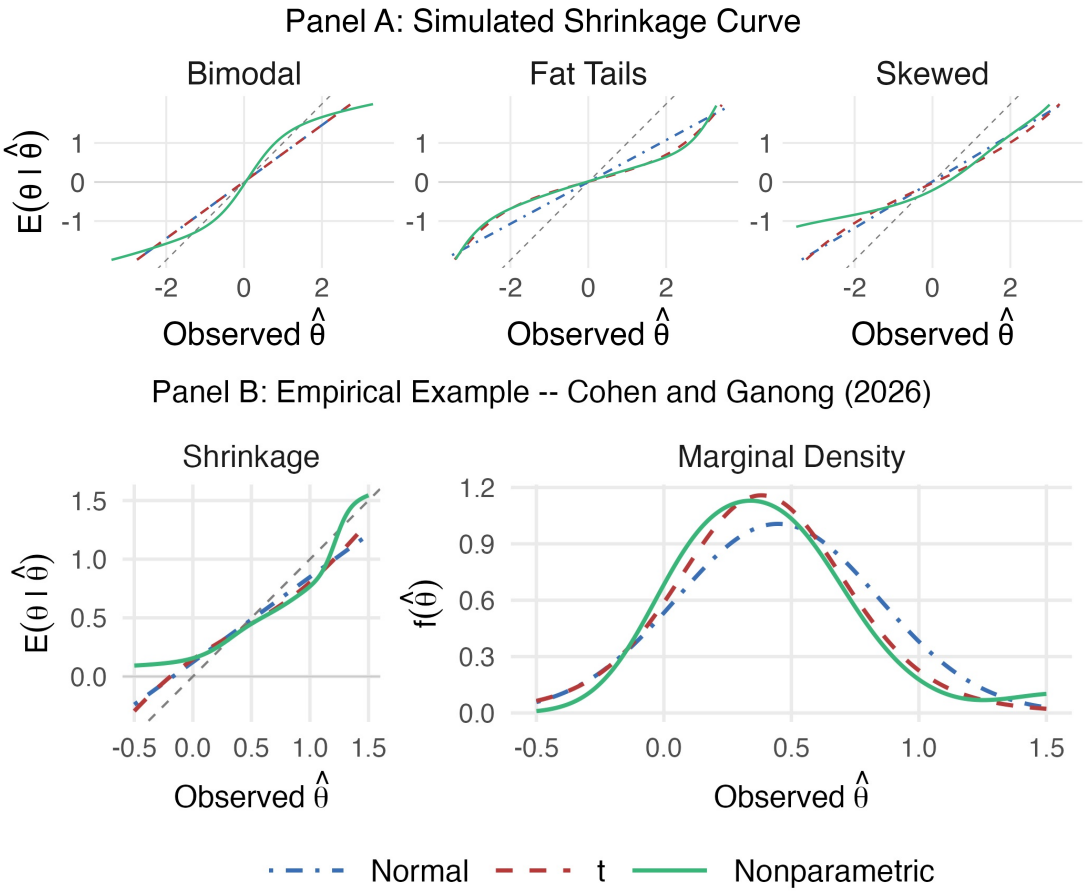
these standard errors are imprecisely estimated in small samples.

Fourth, a non-normal model often provides a better fit to the data. For three of the applications (Cohen and Ganong 2026, Card et al. 2018, and DellaVigna and Linos 2022), we reject that the latent distribution is normal. This is reflected in the finite degrees of freedom parameter df. In these three applications, assuming that the latent distribution is t leads to a much lower estimate for the latent median $\bar{\theta}$ (as compared to assuming that the latent distribution is normal). This finding will serve as a justification for why we prefer to use a t distribution in the context of analyzing selectivity in Section 5.2. For Crosta et al. (2024), assuming the latent distribution is t leads to pathological results since the normal model is the best fit.

3.2.3 Non-parametric empirical Bayes

Until now, we have studied models which assume that the latent distribution of effects can be modeled using a normal or a t distribution. When the latent distribution is in fact skewed or multi-modal, these parametric assumptions will lead to biased conclusions. A more flexible alternative—the nonparametric MLE (Koenker and Mizera, 2014)—combined with Tweedie's formula (equation (4)) allow the shape of the posterior mean to adapt to the data.

Figure 1: Empirical Bayes shrinkage under skewness, fat tails and multimodality



Notes: Panel A shows posterior means $E[\theta_i \mid \hat{\theta}_i, \sigma_i]$ vs. $\hat{\theta}_i$. The left plot uses draws from a mixture of $N(-1.5, 0.6^2)$ and $N(1.5, 0.6^2)$, the center plot uses draws from $0.5t_{2.5}$, and the right plot uses draws from $0.5(\chi^2_3 - 3)$. For all three, sampling noise σ^2 is drawn from Uniform$(0.4, 1.6)$. The reference sampling variance for the new estimate $\hat{\theta}_i$ is $\sigma_i^2 = 1$ Panel B shows the shrinkage curve (left) and estimated marginal density of $\hat{\theta}_i$ (right) for data from Cohen and Ganong (2026). The reference sampling variance for the new estimate $\hat{\theta}_i$ is $\sigma_i^2 = 0.04$, the 75th percentile sampling variance in Cohen and Ganong (2026). On all shrinkage plots, the dashed 45-degree line represents no shrinkage.

Figure 1, Panel A, shows shrinkage curves under three simulated latent distributions that violate normality: a bimodal mixture of two normal distributions, a heavy-tailed t_3, and a right-skewed (centered) χ^2_3. In each case the NPMLE curve departs sharply from the normal benchmark in the tails. In the bimodal case it is strongly non-monotone in slope, pulling observations toward the closer mode rather than a single global average. The t-model captures heavy tails but cannot recover skewness or multimodality. In the skewness case, right-tail estimates far away from zero are shrunk less than closer estimates. Panel B shows the analogous comparison for the meta-analysis of unemployment elasticities from Cohen and Ganong (2026): the marginal density of $\hat{\theta}_i$ is visibly right-skewed with a heavier right tail than a matching normal, and the NPMLE shrinkage curve is correspondingly asymmetric—it shrinks negative estimates aggressively but flattens out for large positive $\hat{\theta}_i$ rather than pulling them back to the global mean.

The choice of latent distribution is most consequential for *study-level* estimates of individual θ_i—ranking treatments, identifying "best" effects, or cutoff-based decisions. For aggregate quantities like $\bar{\theta}$, the differences across models are typically modest. As a practical guideline: when $n \gtrsim 50$ and the empirical distribution of $\hat{\theta}_i$ shows visible deviations from normality, the NPMLE (or at least a t-model) should be preferred; with small n, parametric models might remain more reliable.

4 Evidence aggregation with covariates

Thus far, we have studied what a meta-analysis can discover based on estimates and standard errors alone. Dispersion in the distribution of effects across studies ($\tau^2 > 0$) naturally motivates a search for predictors of study-level heterogeneity. In each of our four applications, readers and researchers will naturally wonder *why* treatments are more effective in some instances than others. In this section, we consider what more can be learned when study-level covariates are available.

4.1 Theory

The basic empirical Bayes approach can be generalized by allowing for covariates, which enter as linear predictors of θ_i; see for example Chetty et al. (2014). A further generalization, using tools that might be less familiar to readers in economics, allows for a non-linear relationship between the covariates X_i and the conditional mean of θ_i given X_i, and does not restrict the functional form of this relationship. This can be implemented using Gaussian process priors (Williams and Rasmussen, 2006), which allow for closed form predictions, despite the unrestricted functional form.

This approach is particularly useful when we want to think about the external

validity of a meta-study, and the extrapolation of estimates to new problem instances. This approach allows for an assessment of (posterior) uncertainty, when extrapolating outside the support of previously encountered instances. By contrast, extrapolations based on (linear) functional forms might give a false sense of certainty.

4.1.1 Meta-regressions

To make predictions of θ_0 for new instances, one needs to learn the relation of effect heterogeneity to contextual covariates X_i, to make predictions for new instances characterized by X_0. This motivates meta-regressions of the form

$$\hat{\theta}_i = X_i \cdot \beta + \epsilon_i,$$

as estimated by Stanley and Jarrell (1989) and by many authors since. We emphasize that the variation in X_i is not exogenous and so this regression identifies a *correlation* between the covariate and the treatment effect rather than a *causal* relationship.

To rationalize such meta-regressions, and generalizing the basic parametric empirical Bayes model, assume that

$$\theta_i | X_i = x \sim N(x \cdot \beta, \tau^2). \tag{5}$$

Here the coefficient vector β takes the role previously held by $\bar{\theta}$. For this model, we then get

$$\hat{\theta}_i | X_i = x, \sigma_i \sim N(x \cdot \beta, \tau^2 + \sigma_i^2).$$

Generalizing the method-of-moments approach of Section 3.1.1, β can be estimated using either OLS, or a weighted least squares (WLS) regression where the weights are $\frac{1}{\sigma_i^2}$. As a further generalization, we can run a regression of $\hat{\theta}_i$ on x, and τ^2 can then be estimated as the average of $e_i^2 - \hat{\sigma}_i^2$, where e_i are the estimated residuals of the WLS regression. We could then re-run the regression, weighting each observation by the inverse of its variance, $\frac{1}{\tau^2 + \sigma_i^2}$, which yields a weighted least squares random effects (WLS RE) predictor.

Alternatively we could estimate the hyper-parameters τ^2 and β jointly by maximizing the marginal likelihood, as in the empirical Bayes approach. Or we could impose a prior on both τ^2 and β to obtain a hierarchical Bayes model, as discussed further below.

4.1.2 Prediction and Gaussian process priors

It is common to think of predictions based on regressions as a two-step process: First, the regression coefficients β are estimated. Then the effect θ_0 for a new instance is predicted as $X_0 \cdot \widehat{\beta}$. There is an alternative, numerically equivalent way to obtain the same result: we can leap-frog estimation of β and directly predict θ_0.

This alternative perspective, which we will explain next, has two advantages for our purposes. First, it makes explicit that predictions are weighted averages of the available estimates in a meta-study, and clarifies the weights used. Second, it lends itself naturally to generalizations from the linear model to non-linear (but smooth) relationships between θ and X. Using such more general models in turn is helpful because it avoids undue reliance on functional form for extrapolation.

Prediction in the linear model (Ridge) In this spirit, let us now characterize the posterior for θ_0 as a function of $(\hat{\theta}_1, \ldots, \hat{\theta}_n)$. In addition to the linear model of Equation (5), assume that τ^2 is known (possibly from first-stage maximum likelihood estimation), and suppose that we have a prior for the coefficient vector β of the form

$$\beta \sim N(0, \rho^2 \cdot I),$$

where ρ^2 is known. Then the prior distribution of $(\theta_0, \hat{\theta}_1, \ldots, \hat{\theta}_n)$ is jointly normal conditional on τ^2 and conditional on the observed covariates $(X_0, X_1, \ldots, X_n)$; we will leave the conditioning implicit in our notation. The prior moments of this joint normal distribution are given as follows, for $i, j = 0, 1, \ldots, n$:

$$\begin{aligned} E[\theta_i] &= E[X_i \cdot \beta] = X_i \cdot E[\beta] = 0, \\ C_{i,i} &= Var(\theta_i) = Var(X_i \cdot \beta) + Var(\theta_i|\beta) = \rho^2 \cdot \|X_i\|^2 + \tau^2, \\ C_{i,j} &= Cov(\theta_i, \theta_j) = Cov(X_i \cdot \beta, X_j \cdot \beta) = \rho^2 \cdot X_i X_j', \text{ for } i \neq j. \end{aligned}$$

where $\rho^2 + \tau^2$ governs the prior variance of θ_i. The ratio $\rho^2/(\rho^2 + \tau^2)$ determines the expected share of variation of θ_i that is explained by X_i (analogous to an R^2 statistic). We collect these terms in the n-vector of prior covariances c and the $n \times n$ prior variance matrix of the estimands C,

$$c = (C_{0,1}, \ldots, C_{0,n}), \qquad C = \begin{pmatrix} C_{1,1} & \ldots & C_{1,n} \\ \vdots & \ddots & \vdots \\ C_{n,1} & \ldots & C_{n,n} \end{pmatrix},$$

to get

$$(\theta_0, \hat{\theta}_1, \ldots, \hat{\theta}_n) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} C_{0,0} & c, \\ c' & C + diag(\sigma_i^2) \end{pmatrix}\right). \tag{6}$$

Given joint normality of $(\theta_0, \hat{\theta}_1, \ldots, \hat{\theta}_n)$, the posterior mean of θ_0 is equal to the posterior best linear predictor,

$$E[\theta_0|(\hat{\theta}_1, \ldots, \hat{\theta}_n)] = \underbrace{Cov(\theta_0, (\hat{\theta}_1, \ldots, \hat{\theta}_n))}_{c} \cdot \underbrace{Var((\hat{\theta}_1, \ldots, \hat{\theta}_n))^{-1}}_{(C+diag(\sigma^2))^{-1}} \cdot (\hat{\theta}_1, \ldots, \hat{\theta}_n)'. \quad (7)$$

For intuition, note that this formula for the best linear predictor is the multivariate counterpart of the familiar formula for the univariate OLS slope, which is equal to the covariance of prediction target and predictor, divided by predictor variance. The posterior distribution of θ_0 is thus given by

$$\theta_0|(\hat{\theta}_1, \ldots, \hat{\theta}_n) \sim N\Big(c \cdot (C + \text{diag}\,(\sigma_i^2))^{-1} \cdot (\hat{\theta}_1, \ldots, \hat{\theta}_n)', \\ C_{0,0} - c \cdot (C + \text{diag}\,(\sigma_i^2))^{-1} \cdot c'\Big). \quad (8)$$

Note that the posterior mean of θ_0 is a weighted average of the $\hat{\theta}_i$, with larger weights put on studies i that have covariate values X_i closer to X_0 (so that $C_{0,i} = \rho^2 \cdot X_0 X_i'$ is larger), and on studies that have smaller values of σ_i. The covariances $C_{0,i}$ measure how similar (close) instance 0 is to instance i. The coefficients $c \cdot (C + diag(\sigma_i^2))^{-1}$ then give the optimal weights for a posterior prediction. The variance $C_{0,0} - c \cdot (C + diag(\sigma_i^2))^{-1} \cdot c'$ measures the posterior uncertainty for θ_0.

Nonlinear generalization (Gaussian Process) The linear model that we just discussed specifies the prior co-variance of θ_i and θ_j as $C_{ij} = K(X_i, X_j)$ for $i \neq j$ and $C_{ii} = K(X_i, X_i) + \tau^2$, where $K(x, x') = \rho^2 \cdot xx'$. The function $K(x, x')$ is called a covariance kernel. As we saw in equation (7), the predicted value of θ_0 can be calculated using just this covariance kernel, without any explicit reference to the regression slopes β.

We have assumed thus far that the prediction function $\bar{\theta}(x) = E[\theta_i|X_i = x]$ is linear; this is reflected in the prior covariances $K(x, x') = Cov(\bar{\theta}(x), \bar{\theta}(x'))$. This approach can be generalized beyond linear models. If we assume that the θ_i are jointly normal (conditional on covariates), with prior covariances specified by a general covariance kernel $K(X_i, X_j)$, then we obtain a so-called Gaussian process prior for $\bar{\theta}(x)$ (Williams and Rasmussen, 2006). Denoting again $C_{ij} = K(X_i, X_j)$ for $i \neq j$ and $C_{ii} = K(X_i, X_i) + \tau^2$, the exact same expression as before (Equation (6)) describes the joint prior distribution of θ_0 and the observed estimates $(\hat{\theta}_1, \ldots, \hat{\theta}_n)$. Correspondingly, the same expression (Equation (7)) continues to describe the posterior distribution of θ_0 given the observed estimates.

The linear model specifies $K(x, x') = \rho^2 \cdot xx'$. A popular alternative covariance

kernel, which does not impose linearity, is the squared exponential kernel,

$$K(x, x') = \rho^2 \cdot \exp\left(-\frac{\|x - x'\|^2}{l^2}\right), \tag{9}$$

The "length scale" l^2 governs the smoothness of the relationship between covariates X and effects θ, where larger l imply smoother functions. This length scale thus place a role analogous to the bandwidth in kernel regressions. These hyper-parameters ρ^2, τ^2 and l^2 could be chosen a priori, for a fully Bayesian specification, or in a data-dependent manner, as in the empirical Bayes approach.

There are many other kernels which have been used in the literature. Spline regression, for example, corresponds to a Gaussian Process prior where $\bar{\theta}(\cdot)$ follows an integrated Brownian motion. While the squared exponential kernel leads to infinitely differentiable predictions, spline regressions are only twice differentiable, in general.

4.1.3 Hierarchical Bayes

The empirical Bayes approach estimates hyper-parameters, such as τ^2 and β for the linear model, by looking at the data. Alternatively, one might put a prior on hyper-parameters (i.e. $\beta \sim \pi_\beta, \quad \tau \sim \pi_\tau$) and obtain a posterior distribution for hyper-parameters and study-level estimates in one step. The posterior based on the meta-regression model from Equation 5 can be written as

$$P(\theta, \beta, \tau | \hat{\theta}, \sigma_i, X) \propto \pi_\beta \pi_\tau \prod \frac{1}{\sigma_i} \varphi\left(\frac{\hat{\theta}_i - \theta_i}{\sigma_i}\right) \frac{1}{\tau} \varphi\left(\frac{\theta_i - X_i\beta}{\tau}\right).$$

This yields hierarchical Bayes estimators (see e.g. Meager 2019).

If we put a normal prior $\beta \sim N(0, \rho^2 I)$ on the coefficients and a normal prior with very large variance on τ, we recover the same estimated β as the linear GP model (Ridge) discussed above. In particular

$$E[\beta | \hat{\theta}_1, \ldots, \hat{\theta}_n, \tau^2] = \underset{b}{\text{argmin}} \sum_i \frac{(\hat{\theta}_i - X_i \cdot b)^2}{\tau^2 + \sigma_i^2} + \frac{\|b\|^2}{\rho^2}.$$

We use Hierarchical Bayes in subsubsection 4.2.1 and subsection 6.2.[2] One advantage of the hierarchical Bayes approach is that it facilitates uncertainty quantification for predictions, taking into account hyper-parameter uncertainty. This

[2]We also use a variant of Hierarchical Bayes in Section 4.2.2 when we do estimation with Gaussian Process priors. In that case, we choose the hyper-parameters manually. This can be thought of as a prior which is a point mass (rather than a distribution).

matters especially in the context of smaller meta-studies, where hyper-parameter uncertainty is non-negligible.[3]

4.1.4 Binary implementation decisions

A leading downstream use of the apparatus developed above is the choice of binary implementation decisions $A_i \in \{0, 1\}$ – for example, whether to scale up a treatment evaluated in study i, or whether to implement a treatment in a new instance characterized by covariates X_0. If θ_i captures all relevant costs and benefits of the treatment, so that welfare is $U(A_i, \theta_i) = A_i \cdot \theta_i$, a hypothetical decision-maker who observed θ_i might choose $A_i = \mathbf{1}(\theta_i \geq 0)$. In practice we observe only $(\hat{\theta}_i, \sigma_i, X_i)$, and the optimal Bayesian decision becomes

$$A_i \;=\; \mathbf{1}\big(E[\theta_i \mid \hat{\theta}_i, \sigma_i, X_i] \geq 0\big), \tag{10}$$

where the posterior mean is computed using the empirical Bayes or Gaussian process expressions above. Note that the sign of this posterior mean may differ from that of $\hat{\theta}_i$ when $\hat{\theta}_i$ is small relative to its standard error or inconsistent with the effects predicted by the covariates. For a new problem instance for which $\hat{\theta}_0$ is not observed (but covariates X_0 are observed), the analogous rule is $A_0 = \mathbf{1}\big(E[\theta_0 \mid X_0, \hat{\theta}_1, \ldots, \hat{\theta}_n] \geq 0\big)$, using the Gaussian process posterior mean for θ_0 from Equation (7). We discuss other implementation decisions in more detail in Appendix C.

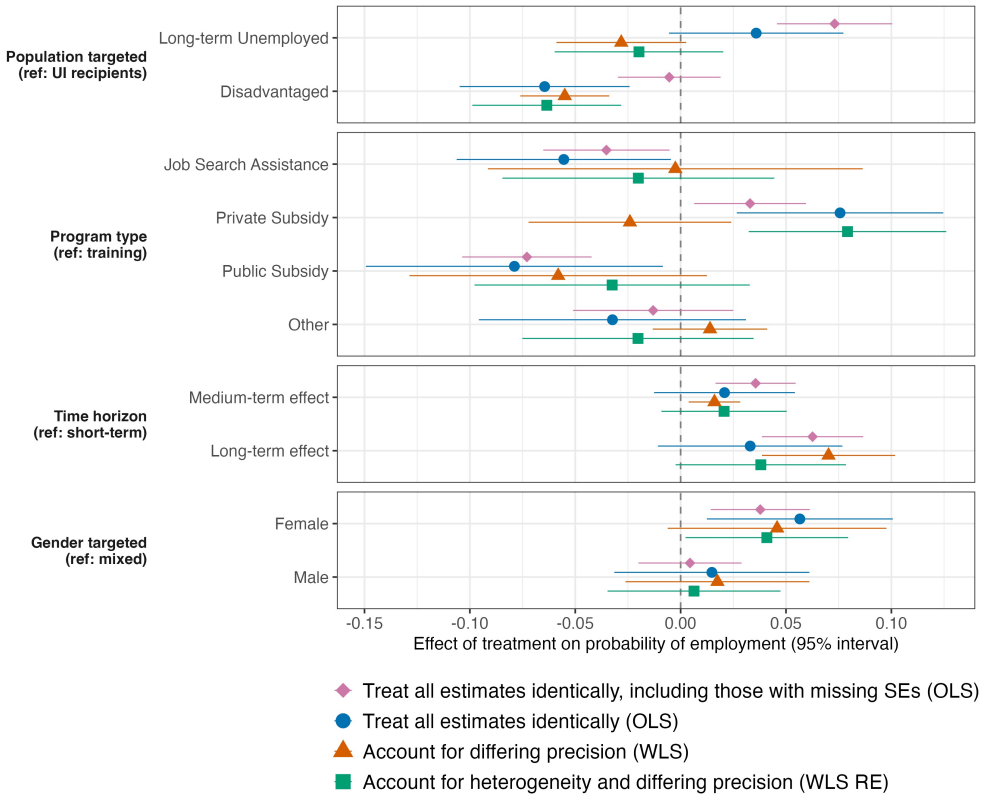
4.2 Empirical examples

4.2.1 Regression interpretation: precision and latent heterogeneity

Card et al. (2018) studies how the effectiveness of active labor market programs (ALMPs) varies with the program type (training, job search assistance, subsidized private sector employment, subsidized public sector employment), with the gender targeted (mixed-gender, female, male), with the time horizon (short-term, medium-term, long-term), and with the population targeted (UI recipients, long-term unemployed, or disadvantaged). The paper considers models where the first group in each parenthetical constitutes the reference group and asks how changes in program type, gender targeted, time horizon, and population targeted affect the probability of employment. The models used in the paper give each estimate

[3] An alternative route to uncertainty quantification, which stays within the empirical Bayes framework rather than imposing priors on hyper-parameters, is provided by the robust empirical Bayes confidence intervals of Armstrong et al. (2022).

Figure 2: Covariate heterogeneity in effect of active labor market programs (Card et al., 2018)



equal weight when assessing the role of covariates.[4] We capture several of the research questions examined in their paper by estimating

$$\hat{\theta}_i = \alpha + \beta_1 \, TimeH_i + \beta_2 \, ProgramType_i + \beta_3 \, PopTargeted_i + \beta_4 \, Gender_i + \epsilon_i. \quad (11)$$

Figure 2 shows how the effectiveness of ALMPs varies with these characteristics, comparing estimates from four types of models. The first regression weights every estimate equally and includes studies with missing standard errors, as in the original Card et al. (2018) analysis.[5] Using this model, the paper concludes that

[4]This estimation approach is quite common. Card et al. (2018) and DellaVigna and Linos (2022) do this via Ordinary Least Squares, while Cohen and Ganong (2026) does this via Bayesian Model Averaging.

[5]In our analysis below, we treat the θ_i as independent draws. As noted in Section 3.2.2, in the studies reviewed by Card et al. (2018), there are often multiple estimates i per study j. Generalizing our approach, these might be modeled as $\theta_{ij} = \bar{\theta}_i + \eta_{ij}$, with each of the $\bar{\theta}_i$ and

programs which target the long-term unemployed are more effective at raising employment. We replicate this finding (shown in purple diamonds): targeting the long-term unemployed raises employment effects by about 7 percentage points. The second limits the sample to studies with non-missing standard errors; the finding holds in this subsample as well.

Applying the frameworks from Section 3.1.1, we explore alternative models that account for differences in precision across estimates and allow for a latent distribution of treatment effects. The third ("WLS") weights each estimate by the inverse of its variance, $w_i = 1/\sigma_i^2$. The fourth ("WLS with random effects") additionally allows for a latent distribution of treatment effects, weighting each estimate by $w_i = 1/(\tau^2 + \sigma_i^2)$, with $\hat{\tau^2} = 0.074$ obtained by a hierarchical Bayes model as discussed in subsubsection 4.1.3. This is reminiscent of the method-of-moments weights in Equation (2).

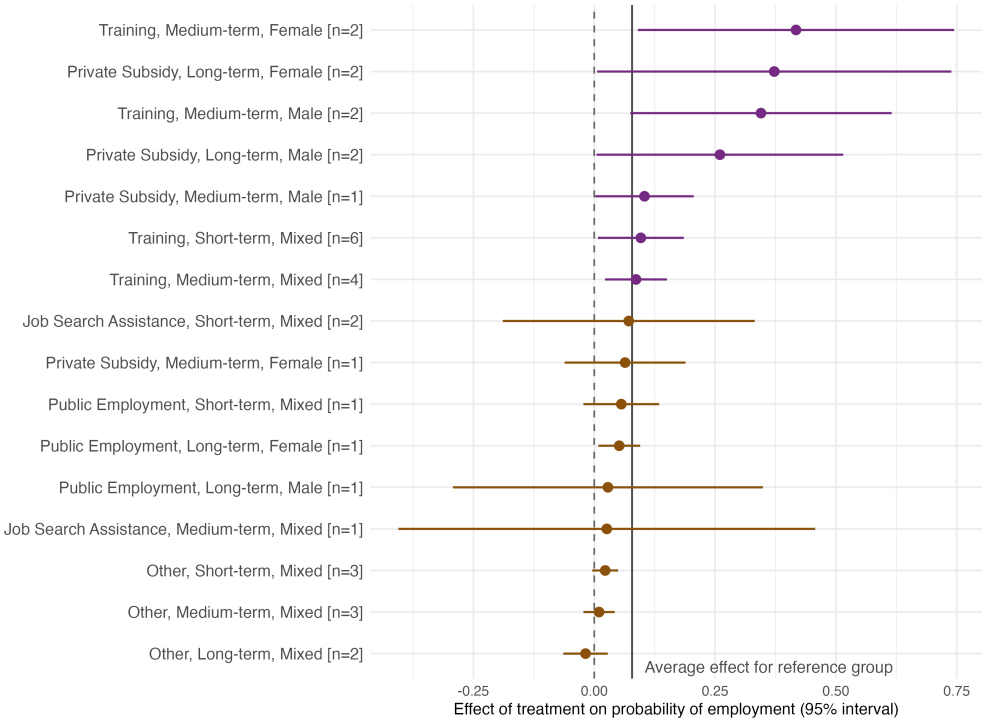
Accounting for differences in precision across estimates undoes – and in fact reverses – the estimated positive effect of targeting the long-term unemployed. The intuition is most easily seen by looking at the study-level estimates and standard errors for programs targeting this group. Figure 3 shows the estimated treatment effect and 95% confidence interval for every study targeting the long-term unemployed. A vertical dark line marks the precision-weighted average treatment effect for all studies targeting the reference group (UI recipients). Two studies report treatment effects of at least 30 percentage points – an enormously large effect relative to much of the literature – but their estimates are imprecise, with standard errors larger than for many of the other studies. The studies with small or negative effects, by contrast, have small standard errors (or no standard errors at all). A precision-weighted regression gives little weight to the first group and substantial weight to the second; this is why, in Figure 2, the coefficient for targeting the long-term unemployed becomes negative under WLS. The same qualitative pattern emerges when we additionally account for latent heterogeneity using $w_i = 1/(\tau^2 + \sigma_i^2)$.

4.2.2 Predictive models, external validity, and the weights on prior studies

Regressions such as equation (11) should be interpreted as predictive models for the effect θ_0 in a new problem instance characterized by covariates X_0, with posterior given by Equation (7). The predicted effect is a weighted average of the previously reported estimates, with weights determined by the prior covariance kernel $K(\cdot, \cdot)$ and by the standard errors σ_i. This connects directly to the question of *external validity*: how confident can a policymaker or researcher be that the conclusions of the existing literature carry over to a new setting?

η_{ij} independent. We do not pursue this here.

Figure 3: Study-level treatment effects for programs targeting the long-term unemployed



Notes: Each point is the mean treatment effect across the estimates from a single study targeting the long-term unemployed, with [n=k] giving the number of such estimates from that study in the Card et al. (2018) data; the whiskers are 95% intervals based on the average standard error across those estimates. The dark vertical line marks the precision-weighted average treatment effect for the reference group of studies targeting UI recipients (= 0.036). Studies with a mean effect above this average are purple and those below are brown. Vodopivec (1999) is excluded because of its extreme standard errors.

To illustrate, consider two hypothetical new contexts in which a policymaker is contemplating a training program for recipients of mixed gender evaluated at the long-term horizon, again using the Card et al. (2018) data:

1. **Context A (target UI recipients, within support).** A *training* program targeted to UI recipients of mixed gender, . Ten of the 169 estimates in the sample fall in this cell, so the data speak directly to this context.

2. **Context B (target long-term unemployed, outside support).** A *training* program targeted to long-term unemployed (LTU) of mixed gender, evaluated at the *long-term* horizon. No estimate in the sample falls in this cell, so prediction is fully extrapolated along the recipient dimension.

We focus on the long-term unemployed for continuity with the meta-regression

discussion above, and compare to UI recipients because they are the most commonly observed recipient group in the meta-analysis sample.

For each context, we compute the posterior mean prediction for θ_0, the posterior standard deviation, and the implicit weights w_i on each prior estimate, under three models that differ only in the prior covariance kernel: (i) weighted least squares with random effects; (ii) a squared-exponential GPP with a long length scale ℓ, smoothing more aggressively across the covariate space; and (iii) a squared-exponential GPP with a short length scale ℓ, putting most weight on very close neighbours.

Predictions Prior hyperparameters are set to sensible defaults based on the data. Covariates are pre-normalised (mean 0, variance 1) before kernels are evaluated. The sample variance of $\hat{\theta}_i$ is $\widehat{\mathrm{Var}}(\hat{\theta}) \approx 0.012$ and the median sampling variance is $\widetilde{\sigma}^2 \approx 0.001$, giving a signal variance $V \equiv \widehat{\mathrm{Var}}(\hat{\theta}) - \widetilde{\sigma}^2 \approx 0.011$. For the linear model, we set $\tau^2 = \rho^2 = V/(q+2)$, where $q = \dim(X_i)$.[6] For the squared exponential kernel, we set $\tau^2 = \rho^2 = V/2$. These choices ensure that the unconditional prior variance of θ_i matches the observed variance. The squared-exponential length scales are $\ell = \sqrt{q}/2 \approx 1.58$ (short) and $\ell = 2\sqrt{q} \approx 6.32$ (long).

Table 5 reports posterior means and standard deviations under each method, and Figure 4 displays the weights graphically for the most heavily-weighted estimates (positive or negative). The per-study estimates, standard errors, distances to each test point, citations, and sample sizes for every row displayed in Figure 4 are reported in Appendix D. Across the three specifications, the posterior means for the employment effects range from about 12 to 15 percentage points for Context A and from about 6 to 10 percentage points for Context B. Context A exceeding Context B is a faint echo of the negative LTU coefficient seen in the regression analysis above.

Table 5: Posterior distribution of employment effects of training

		WLS	GPP long l	GPP short l
UI recipients	Posterior mean	0.122	0.136	0.149
	Posterior SD	0.003	0.003	0.004
Long-term unemployed	Posterior mean	0.094	0.096	0.058
	Posterior SD	0.005	0.010	0.067

Policy Recommendations From predicted effects, it is then a short step to policy recommendations using the results from equation 4.1.4. Suppose that

[6]The number of covariates q is 10, because the equation 11 is estimated using the ten dummy variables shown in Figure 2.

the decisionmaker is willing to pay \$10,000 for each additional employed person and the program costs \$1,000 to administer per-participant, regardless of who is served. The estimates then imply that they would fund the program for UI recipients for whom the expected value is positive ($\$10,000 \times 0.122 - \$1,000 > 0$) but not for the long term unemployed for whom the expected value is negative ($\$10,000 \times 0.094 - \$1,000 < 0$) . After receiving such a recommendation, it would be natural for the policymaker (or the researcher generating the recommendations) to want to understand which estimates are most influential for generating the policy recommendation.

Mechanisms A detailed review of Figure 4 clarifies the mechanisms underlying the predictions. The left-most column of Figure 4 shows a reference case that ignores covariates entirely: a fixed-effect inverse-variance-weighted meta-analysis ($\tau^2 = 0$) where each estimate's weight is proportional to $1/\sigma_i^2$. Estimates #14, #15, and #16 receive the most weight because their reported standard errors are roughly an order of magnitude smaller than every other estimate in the sample. Assigning these estimates such a high weight is of course nonsense because they describe programs which are very different from the context of interest.

Focusing first on context A (in sample prediction for UI recipients), Figure 4 shows that the three models do essentially the same thing, which is to construct an inverse-variance weighted mean of the ten estimates whose covariates exactly match this context. Estimates #1–#7 have small standard errors and receive meaningful weight; the other three estimates are sufficiently noisy that they receive little weight.

The differences between the models are most consequential and interesting in context B (out of sample prediction for the long-term unemployed); two patterns are noteworthy. First, meta regression and GP with long ℓ place substantial negative weights on a wide set of estimates. With little or no data at the target cell, the linear predictor effectively interpolates from neighbouring cells using contrasts of the form

$$E[\theta_0 \mid \text{LTU, L}] = \underbrace{E[\theta_0 \mid \text{UI, L}]}_{\text{estimates \#1–\#7}} + \underbrace{E[\theta_0 \mid \text{LTU, S}]}_{\text{estimates \#9, \#10}} - \underbrace{E[\theta_0 \mid \text{UI, S}]}_{\text{estimate \#8}}. \tag{12}$$

This assumption is analogous to the parallel trends assumption in a difference-in-difference design.

This re-weighting pattern is visible in Figure 4 for Context B under meta regression: positive weight on UI training at the long horizon which we used in Context A (Estimates #1–#7), positive weight on LTU training at the short horizon (Estimates #9–#10), and negative weight on UI training at the short horizon (Estimate #8). The same mechanism also generates the other negative

meta-regression weights visible in the figure.[7] Negative weights aren't a problem in this setting: they are how a linear predictor (or long ℓ) extrapolates by subtracting off distant test points.

Second, the posterior standard deviations differ dramatically across methods for the out-of-support Context B, even though the point predictions agree. For meta regression, the posterior standard deviation in Context B is essentially the same as in Context A. For the long length scale it grows roughly threefold. For the short length scale it grows by a factor of about seventeen, reflecting that the data are silent about a recipient $\times$ horizon cell with zero observations under a strictly local kernel.

To understand the mechanism by which the posterior standard deviation is so tight in context A (unconditionally) and in context B (when using a long length scale), but is so large in context B when using a short length scale, it is helpful to review Table 9, which reports the kernel covariances from equation 9 between the prior studies and the new context. With long ℓ, the model can confidently predict behavior in the new context. With short ℓ, the model assumes little covariance between the existing studies and the new context. Instead, the model's prediction is much closer to the precision of the prior before any data is observed (which is the unconditional prior variance equal discussed above under "Predictions").

How much to extrapolate across the covariate space is ultimately a question about the economic model the researcher is willing to entertain, and the choice of length scale makes that model explicit. A researcher who views employment impacts of the training program as an *additive* function of the population served and the time horizon will want to borrow heavily across cells. Even if a researcher doesn't view the relationships as literally additive, so long as they are comfortable with extrapolation from far away test points, they can use long ℓ GP. Concretely, in the context of this application, these models are suitable if the researcher believes that impacts for UI and LTU differ by one constant, and likewise that long- and short-term impacts differ by a second constant.[8] Under such additivity, the long-term effect of training on the long-term unemployed can be reconstructed by adding a population adjustment to the training effect for UI recipients, and a long length scale is appropriate.

[7]For example, negative weight is consistently assigned to Estimate #16 (Dis/Tr/S, whose unusually small standard error pins down the disadvantaged-inflow indicator and is then subtracted off because the test point is not disadvantaged), To a lesser extent, there are also negative weights on Estimates #14, and #15. This occurs for the same underlying reason: because the test point is neither LTU/Other nor Disadvantaged, the model nets out their identifying contributions.

[8]This is reminiscent of the literature on surrogates, where short-term impacts are used to predict long-term outcomes (Athey et al., 2025).

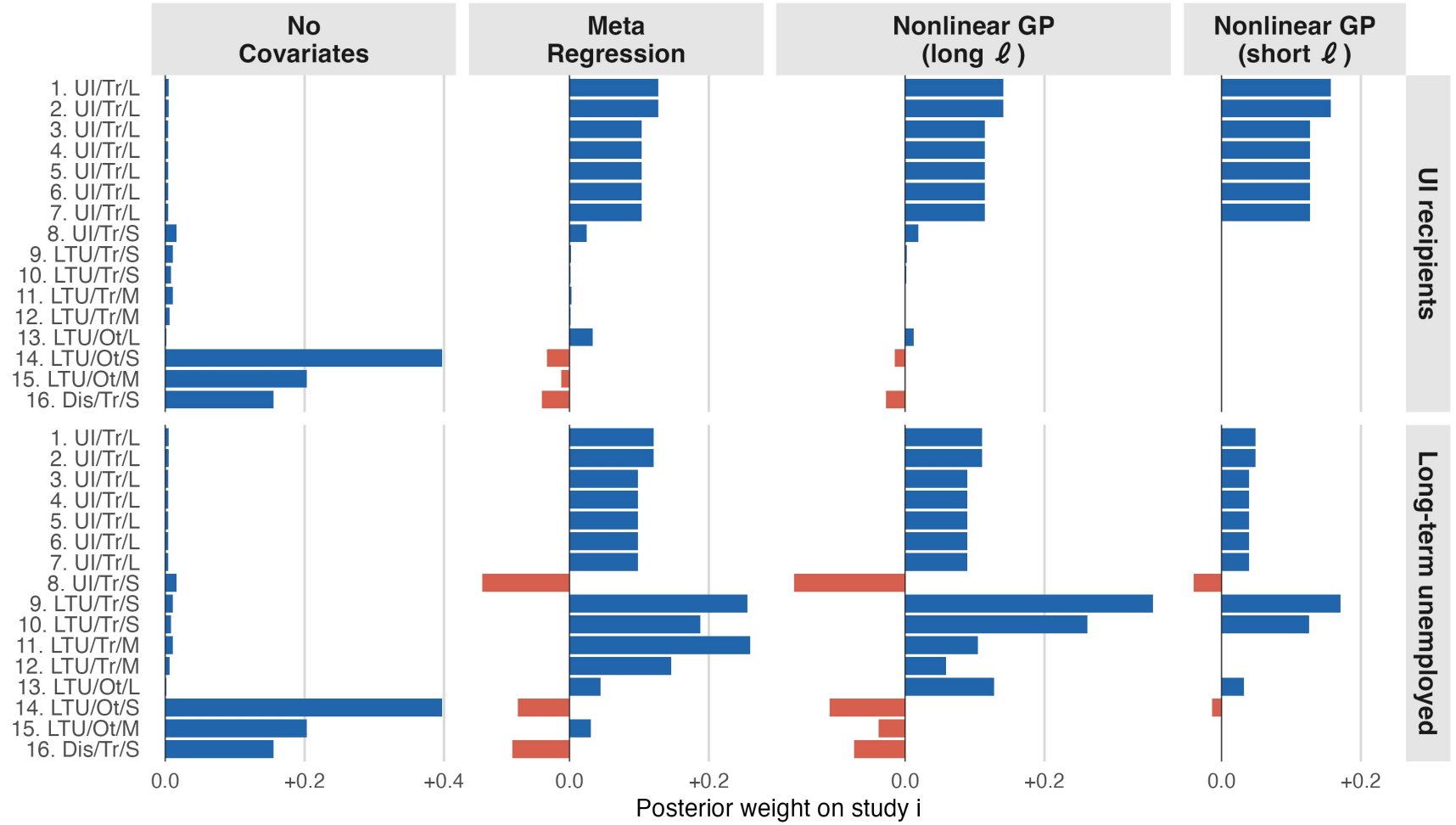


Figure 4: Predictive weights on prior estimates for two hypothetical CKW contexts

Notes: Each column is a prediction method (the leftmost, "No Covariates," is a precision-only baseline that ignores covariates); the top row of panels is Context A (UI recipients, in support) and the bottom row is Context B (long-term unemployed, out of support), both for mixed-gender Training programs at the long-term horizon. Bars show the posterior weight each study receives in the predictor of θ_0, coloured by sign (blue positive, red negative). The y-axis lists the sixteen studies that appear among the ten highest-weight estimates in either context, labelled by target/program/horizon: target — UI = registered unemployed, LTU = long-term unemployed, Dis = disadvantaged; program — Tr = training, Ot = other; horizon — S = short, M = medium, L = long. Hyperparameters and per-study estimates, standard errors, and kernel distances are reported in Table 9.

A researcher who instead believes that the long-term unemployed respond to training *differently* — in either direction — will prefer a short length scale that declines to make this extrapolation. The structural literature supplies models of both signs. Training may be *more* effective for the long-term unemployed: in the equilibrium search model of Kospentaris (2021), skill loss and reduced search effort account for much of the decline in job-finding among the long-term unemployed, so that human-capital programs do the most for precisely the workers who have lost the most. Or, training may be *less* effective for them: if employers screen on unemployment duration (Kroft et al., 2013; Jarosch and Pilossoph, 2019), so that callback rates fall sharply as a spell lengthens, then even retrained long-term unemployed workers may be unable to convert their new skills into offers. Either mechanism implies a training-by-population interaction that the additive predictor cannot represent, and a researcher who finds either one credible should shorten the length scale accordingly.

5 Selectivity

5.1 Identification

In Section 4, we focused on evidence aggregation and extrapolation of findings to new contexts, in the absence of selectivity in the publication process. In practice, however, selectivity is very much a problem for inference based on published evidence: the probability $\bar{d}$ that an empirical finding is published might be a function of that finding itself.

In this chapter, we discuss identification in the presence of selectivity. Such selectivity might be due to decisions by various parties, including researchers (who might engage in p-hacking or specification searching), reviewers and editors (who selectively accept papers, leading to publication bias), as well as due to those conducting a meta-study (who always face the choice of which studies to include). Selectivity does not necessarily reflect improper research methods, but it complicates the interpretation of published findings.[9]

Different identification methods differ in terms of their intended purpose and ambition. They might be designed to just test for the presence of selectivity ($\bar{d} \not\equiv 1$). More ambitiously, they might estimate the degree of selectivity and the shape of the selection function $\bar{d}(\cdot)$. They might also provide selection-corrected

[9] As discussed formally in Frankel and Kasy (2022), there are conflicting objectives: validity of conventional inference requires the *absence of selectivity*. If published findings are to inform subsequent decision-making, however, then *surprising findings* should be selected, because those are the ones which will have a meaningful impact on decisions. But if there is uncertainty over the validity of some identification strategy or data-source, then surprising findings might be less plausible, and plausibility might require the selection of *unsurprising findings*.

estimates of average effects $\bar{\theta}$, or more broadly of the distribution μ of θ_i across studies. Finally, some methods might provide study-level corrected estimates of the θ_i themselves.

Different methods for identification furthermore leverage data of varying degrees of richness. First, they might only use the (marginal) distribution of published Z-statistics or p-values, and seek to test for jumps or non-monotonicity of the p-curve (density of published p-values). Second, using richer information, they might use the joint distribution of estimates and standard errors (or sample sizes), and seek to leverage a possible dependence between estimates and standard errors to test for the presence of selectivity, or – more ambitiously – to identify the degree and shape of selectivity.[10]

Lastly, methods for identifying selectivity might additionally leverage unselected samples of ground-truth estimates. Such estimates might come from research efforts that were pre-registered, for instance by grant agencies, non-academic research labs, or systematic replication studies. Because such non-selected ground truth samples are only rarely available, we focus our review here on the first two approaches, using either just the p-curve, or alternatively the joint distribution of estimates and standard errors.

The results in this section focus on non-parametric identification, and avoid reliance on functional form assumptions. We then discuss both parametric and non-parametric estimation based on these non-parametric identification results in the context of the empirical examples below.

5.1.1 Distribution of p-values (Elliott et al. (2022))

Testing for selectivity: The derivation in the preceding section assumed that publication was non-selective, so that D_i is independent of Z_i across studies. In the presence of selectivity, however, we have in general that $\bar{d}(z) = P(D_i = 1|Z_i = z)$ is not constant as a function of z, and thus, denoting $\tilde{d}(p) = \bar{d}(z)$,

$$g(p) \cdot \frac{\tilde{d}(p)}{E[D_i]} \neq g(p).$$

In words, the density of published p-values deviates from the density $g(p)$ of latent p-values by a factor proportional to the publication probability $\tilde{d}(p)$.

The latent density of Z-statistics is given by the convolution $f(z) = \int \varphi(z - \omega) d\nu(\omega)$, where ν is the distribution of $\omega_i = \theta_i/\sigma_i$. The corresponding density of

[10]The function $\bar{d}(\cdot)$ might for instance be U-shaped, if there is a preference for significant results, or inverse U-shaped, if there is a preference for insignificant results, or monotonically increasing/decreasing, if there is a preference for positive/negative effects, etc. See also Kasy (2021).

p-values equals $g(p) = f(z)/\varphi(z)$ for $z = \Phi^{-1}(1-p)$. Elliott et al. (2022) characterize this density, showing that it is smooth and non-increasing. In Appendix E we provide a short proof of these properties.

This suggest the following testable hypotheses, which are implied by the null hypothesis of no selectivity:

1. H_0': $f(p)$ is continuous at p, for conventional significance levels such as $p = .05$.

2. H_0'': $f(p)$ is non-increasing in p, across all values of p.

Tests based of the first of these derived hypotheses (i.e. continuity/smoothness) have power if $\tilde{d}(p)$ is discontinuous at conventional levels. Tests based on the second of these derived hypotheses do not have power if $\tilde{d}(p)$ is also non-increasing: In that case $f(p) = g(p) \cdot \tilde{d}(p)$ is non-increasing, since both $g(p)$ and $\tilde{d}(p)$ are. Typically, researchers prefer smaller p-values, and "significant" results. Whenever that is the case, then $\tilde{d}(p)$ is indeed non-increasing.

There are, however, scenarios where the opposite is true. This includes notably cases where researchers may wish to find an insignificant result, for instance when testing for pre-trends in a difference-in-differences context, and in the context placebo tests for the validity of some research design more broadly.

5.1.2 Meta-studies

The approaches that we have just discussed are based solely on the distribution of p-values P_i or, equivalently, the distribution of Z-statistics $Z_i = \hat{\theta}_i/\sigma_i$. These approaches only allow us to test for the presence of selectivity, but not to identify its shape or magnitude, nor to correct published estimates. In meta-studies, however, we typically have not only P-values at our disposal, but instead observe both point estimates $\hat{\theta}_i$ and standard errors σ_i. We can therefore leverage their *joint* distribution for identification, if we are willing to impose some additional assumptions. This also allows us to be more ambitious, going beyond mere tests of the presence of selectivity, to estimate the magnitude and shape of selectivity, and to implement selection corrections.

Approaches that leverage this joint distribution typically rely on two assumptions, both of which are substantively restrictive. The first of these is the assumption that estimands and standard errors vary independently across studies, so that

$$\theta_i \perp \sigma_i. \tag{13}$$

The second is the assumption that selection is only based on p-values,

$$E[D_i = 1|\widehat{\theta}_i = \theta, \sigma_i = \sigma] = \bar{d}(Z_i). \tag{14}$$

These two assumption are easily generalized to allow for independence conditional on observed covariates X_i; we do so in subsection 6.1 below. The first of these two assumptions might be violated if sample sizes in experiments are chosen based on power calculations, and if the researchers preparing these experiments have correct priors regarding how the magnitude of treatment effects varies across studies (Allcott, 2015; Gechter et al., 2024). The second of these assumptions might be violated if selectivity depends on the magnitude of estimates, in addition to p-values, for instance because only certain magnitudes are considered substantively (economically) meaningful by authors or reviewers, regardless of statistical significance.

We will discuss three versions of identification based on the joint distribution of $\hat{\theta}_i$ and σ_i: (1) meta-regressions (Egger et al., 1997), (2) the "weighted average of adequately powered" estimates (Stanley et al., 2017), and (3) simultaneous estimation of selection function and distribution of estimands (Andrews and Kasy, 2019).

All three approaches are based on the same two assumptions: independence and selection based on p-values. The theoretical arguments made below to justify the approaches of Egger et al. (1997) and of Stanley et al. (2017) do not appear explicitly in the original references, but they rationalize the methods proposed in these papers.

Testing for selectivity: Meta-regressions (Egger et al., 1997) Consider the null-hypothesis that there is no selectivity, so that $\bar{d} \equiv 1$. Under this null hypothesis, the independence assumption $\theta_i \perp \sigma_i | X_i$ implies conditional mean independence $E[\hat{\theta}_i|\sigma_i, X_i] = E[\theta_i|\sigma_i, X_i] = E[\theta_i|X_i]$. As proposed by Egger et al. (1997), this motivates linear regressions of $\hat{\theta}_i$ on σ_i (and possibly X_i, which we omit for notational simplicity in the following),

$$\hat{\theta}_i = \alpha + \beta \cdot \sigma_i + \epsilon_i,$$

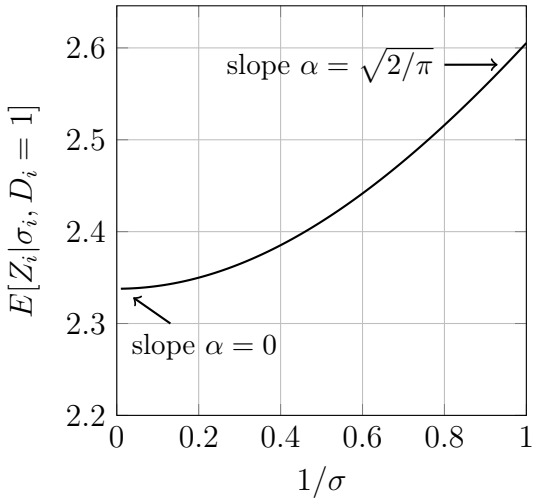
or equivalently (this is the version originally proposed by Egger et al. 1997), after dividing both sides by σ_i,

$$Z_i = \beta + \alpha \cdot \tfrac{1}{\sigma_i} + \tfrac{\epsilon_i}{\sigma_i}. \tag{15}$$

This latter form of the regression can be motivated as a weighted least-squares estimator that down-weights noisier observations. Note that α and β trade places as intercept and slope, after this transformation.

Under the null hypothesis, we obtain $\beta = 0$: There should be no systematic relation between the magnitude of estimates $\hat{\theta}_i$ and the size of standard errors σ_i. A conventional t-test can be used to test this implication, based on either version of the regression. Rejection of the null that $\beta = 0$ can then be interpreted as evidence of selectivity.

Figure 5: Nonlinearity of meta-regressions.



Notes: This figure plots $E[Z_i|\sigma_i, D_i = 1]$ for $D_i = \mathbf{1}(Z_i \geq \bar{z})$ with $\bar{z} = 1.96$ and $\theta \sim N(0, 1)$. For this example, $E[Z_i|\sigma_i, D_i = 1] = \rho \cdot \frac{\varphi(\bar{z}/\rho)}{1-\Phi(\bar{z}/\rho)}$, where $\rho = \sqrt{1 + \frac{1}{\sigma^2}}$, and correspondingly $E[\hat{\theta}_i|\sigma_i, D_i = 1] = \sqrt{1 + \sigma^2} \cdot \frac{\varphi(\bar{z}/\rho)}{1-\Phi(\bar{z}/\rho)}$; this is a special case of equation (16).
The meta-regression approach, as in Egger et al. (1997), fits a linear approximation to the non-linear function $E[Z_i|\sigma_i, D_i = 1]$.

Note that linearity of the relationship between $\hat{\theta}_i$ and σ_i is trivially satisfied *under the null*, since they are mean-independent. Linearity is however generically violated *under the alternative*, when the published evidence is selected, so that $\bar{d} \not\equiv 1$: The function

$$E[Z_i|\sigma_i, D_i = 1] = \int \int \frac{\bar{d}(z)}{E[\bar{d}(Z)|\sigma_i]} \cdot z \cdot \varphi\left(z - \tfrac{\theta}{\sigma_i}\right) dz \; d\mu(\theta) \tag{16}$$

depends nonlinearly on $\frac{1}{\sigma}$. This is illustrated by Figure 5, which plots $E[Z_i|\sigma_i, D_i = 1]$, for $D_i = \mathbf{1}(Z_i \geq \bar{z})$ with $\bar{z} = 1.96$ and $\theta \sim N(0, 1)$, against $1/\sigma$.

This observation is important because many papers aim to interpret α in regression (15) as an estimate of $\lim_{\sigma \to 0} E[\hat{\theta}_i|\sigma_i = \sigma, D_i = 1]$ (which in turn is

taken to identify the average effect $E[\theta_i]$). Such an interpretation is *not* valid in general. In the example of Figure 5, *any* coefficient α between 0 and $\sqrt{2/\pi}$ could be obtained, depending on the distribution of σ_i. Linear extrapolation will therefore necessarily lead to biased estimates of the average effect $\bar{\theta}$.

Estimating average effects: Highly powered studies (Stanley et al., 2017) Meta-regressions allow us to construct a test of the null that there is no selectivity (as was the case for our discussion of the distribution of p-values). An alternative goal, in line with traditional meta-studies, is to obtain an estimate of the expectation $\bar{\theta}$ of θ_i across studies, that is, of the average effect. Assume that $\bar{d}(z) \to 1$ as $|z| \to \infty$, which means that findings with large z-statistics are always published. By definition, and recalling our assumption that selection depends only on the Z-statistic,

$$E[\hat{\theta}_i|\sigma_i = \sigma, D_i = 1] = \frac{E[\hat{\theta}_i \cdot \bar{d}(\hat{\theta}_i/\sigma)|\sigma_i = \sigma]}{E[\bar{d}(\hat{\theta}_i/\sigma)|\sigma_i = \sigma]}.$$

By assumption, $\bar{d}(\theta/\sigma) \to 1$ as $\sigma \to 0$, whenever $\theta \neq 0$. By the dominated convergence theorem (assuming that $E[|\theta_i|] < \infty$ and $P(\theta_i = 0) = 0$), this implies

$$\lim_{\sigma \to 0} E[\hat{\theta}_i|\sigma_i = \sigma, D_i = 1] = E[\theta_i].$$

It follows that the average estimate for "highly powered" studies, with small σ_i, is approximately equal to the average estimand $E[\theta_i]$. Put differently, highly powered studies are not subject to selection, because they always yield large z-statistics.

This argument justifies the proposed estimator of $E[\theta_i]$ in Stanley et al. (2017) (see also Ioannidis et al. 2017), who suggest to focus on a weighted average of highly powered studies, in order to to estimate $\bar{\theta} = E[\theta_i]$. To be empirically feasible, this approach requires the meta-study to include observations in the vicinity of $\sigma_i = 0$. Put differently, there need to be studies that are highly powered in the sense that $E[\bar{d}(Z_i)|\sigma_i] \approx 1$. In practice, it might well be the case that the latent distribution of θ_i has mass around 0, so that even studies with small σ_i might not be highly powered. In this case, no reliable estimate of $\bar{\theta}$ can be formed by using the approach of Stanley et al. (2017). We return to this point in Section 5.2.

Note also that the assumption that $\bar{d}(z) \to 1$ as $|z| \to \infty$ requires that both positive and negative estimates are published, whenever $|z|$ is large enough. This excludes one-sided selection, where only positive significant estimates are published, for instance. In the example in Figure 5 this is not the case. For this example, $\lim_{\sigma \to 0} E[\hat{\theta}_i|\sigma_i = \sigma, D_i = 1] = 2 \cdot \varphi(0) = \sqrt{2/\pi} \neq 0 = E[\theta_i]$.

Estimating selectivity and effect distributions (Andrews and Kasy, 2019) Meta-regressions allow us to test for the presence of selectivity. Highly powered studies allow us to estimate average effects. As it turns out, however, we can achieve considerably more under our assumptions. We can not only test for the presence of selectivity, but we can also identify *how much* more likely significant estimates are to be published. And we can not only identify the average effect $\bar{\theta}$ when there are highly powered studies, but we can non-parametrically identify the distribution μ of θ_i across studies – for instance the *variance* of effects across studies, or the presence of *fat tails*. This was proven in Andrews and Kasy (2019): both the distribution μ of θ and the selection function $\bar{d}(z)$ are non-parametrically identified under the assumptions that $\hat{\theta}_i|\theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2)$, $\theta_i \sim \mu$, $\theta_i \perp \sigma_i$, and $E[D_i = 1|\widehat{\theta}_i, \sigma_i] = \bar{d}(Z_i)$.

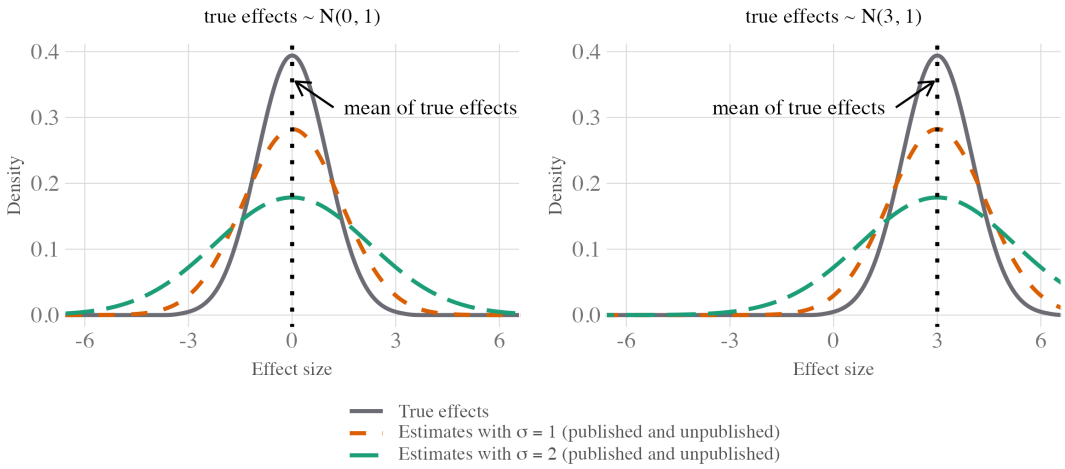
To provide some intuition for this identification result, we provide both algebra and graphical examples. We first describe how to construct functions of the data which are purged of selectivity, and then describe how to recover the distribution μ of θ. Algebra and graphical examples are used to *illustrate* the source of non-parametric identification in the model. Construction of the actual estimators uses maximum likelihood estimation.

Suppose that studies can have only two levels of the standard error: σ_1 and σ_2. The distributions of estimates, conditional on either standard error, draw from the same latent distribution of true effects and are affected by the same selection process. They differ only in the fact that σ_2 is noisier than σ_1. Assume further (again just for illustration) that the latent distribution μ is normal.[11]

Figure 6 illustrates several distributions in this simplified environment. Two latent distributions are shown in black: in the left panel the mean is zero and in the right panel the mean is three. Both latent distributions have variance, $\tau^2 = 1$. In addition, the figure shows the distribution of estimates drawn from the latent distribution with noise $\sigma = 1$ (published and unpublished) and the distribution of estimates with noise $\sigma = 2$ (published and unpublished).

[11] Andrews and Kasy (2019) prove identification in a setting where the distributions of true effects and standard errors are left unrestricted, and selectivity can based on an arbitrary function of the Z-statistic.

Figure 6: Distribution of true effects and all estimates – published and unpublished

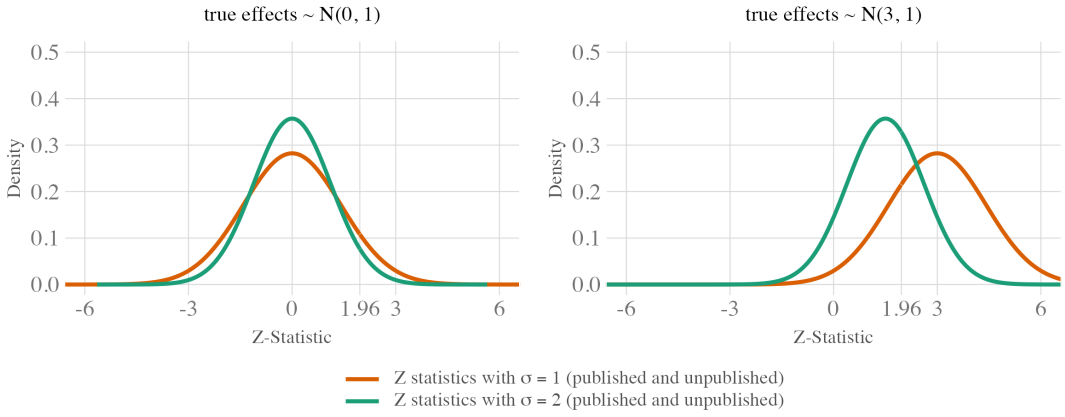


Consider the density of the Z-statistic across studies (both published and unpublished), conditional on the standard error:

$$f(z|\sigma) = \int \varphi\left(z - \tfrac{\theta}{\sigma}\right) d\mu(\theta)$$

Figure 7 shows the distributions of the Z-statistics for the illustrative example. Larger standard errors are associated with *smaller* Z-statistics so the distribution for $\sigma = 1$ is more dispersed than the distribution for $\sigma = 2$.

Figure 7: Distribution of Z-statistics – published and unpublished



The density of the Z-statistic conditional on the standard error and conditional

on observability ($D_i = 1$) is then given by

$$f(z|\sigma, D = 1) = \frac{\bar{d}(z)}{E[\bar{d}(Z)|\sigma]} \cdot f(z|\sigma).$$

Figure 8: Distribution of Z-statistics – published only

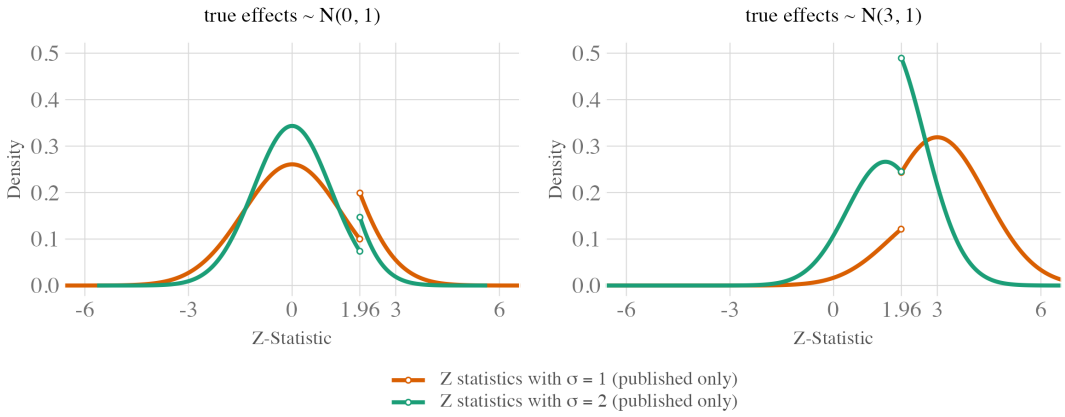


Figure 8 shows the distributions of the Z-statistics in the presence of selectivity. We assume that selectivity can be characterized by two values: the publication rate for significant estimates ($Z > 1.96$), which is normalized to 1, and the publication rate for insignificant estimates ($Z < 1.96$), which is assumed in this example to be 0.5. Both distributions therefore show a discontinuity with a doubling of the density at $Z = 1.96$.[12]

The key identification idea behind the meta-studies model in Andrews and Kasy (2019) is that if we take the ratio of the density $f(z|\sigma, D = 1)$ across these two values of σ, then the selection probability $\bar{d}(z)$ appears in both the numerator and the denominator so it drops out:

$$\frac{f(z|\sigma = 1, D = 1)}{f(z|\sigma = 2, D = 1)} = \frac{\frac{\bar{d}(z)}{E[\bar{d}(Z)|\sigma=1]} \cdot f(z|\sigma = 1)}{\frac{\bar{d}(z)}{E[\bar{d}(Z)|\sigma=2]} \cdot f(z|\sigma = 2)} \tag{17}$$

$$= const \cdot \frac{\int \varphi\left(z - \frac{\theta}{\sigma=1}\right) d\mu(\theta)}{\int \varphi\left(z - \frac{\theta}{\sigma=2}\right) d\mu(\theta)}, \tag{18}$$

[12]This plot makes apparent the connection between the meta-studies approach method and the approach based on discontinuities in the distribution of p-values from Section 5.1.1. The latter effectively zooms in on the density close to the discontinuity. The meta-studies method uses all of the information in the density but requires the independence assumption; it also yields much richer results.

where *const.* is a constant that does not depend on z.

Figure 9: Ratio of observed $\sigma = 1$ distribution to $\sigma = 2$ distribution

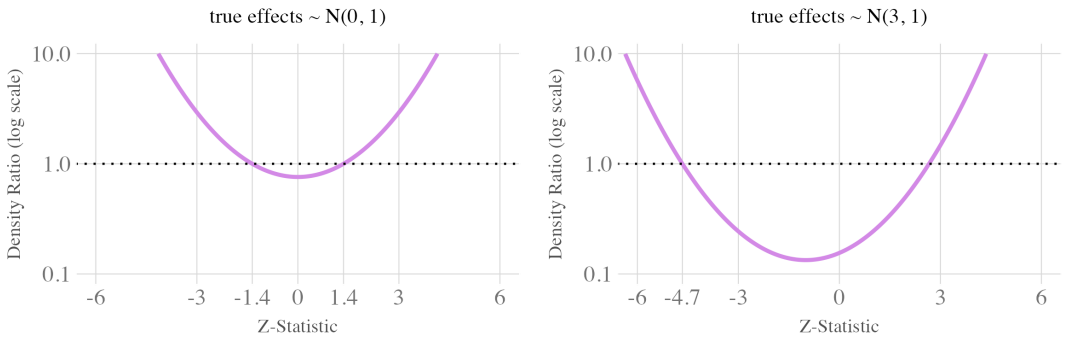


Figure 9 shows the ratio of the two densities. Notice the ratio of densities in Figure 7 is the same as the ratio of densities in Figure 8. Although the densities in Figure 8 are distorted by selectivity, the *ratio* of the densities is not distorted. This is because the density function for $\sigma = 1$ jumps by a factor of $\frac{1}{0.5} = 2$ at the cutoff, and the density function for $\sigma = 2$ also jumps by a factor of two at the cutoff. Having established how to construct functions of the data which are purged of selectivity, we next turn to how parameters of the distribution can be identified using these functions of the data.

The relative distribution of Z-statistics—encapsulated by the density ratio—provides a "signature" which can be uses to identify the mean of the latent distribution. In Figure 9, when the latent mean is zero, the density ratio of the $\sigma = 1$ distribution to the $\sigma = 2$ distribution is greater than 1 when $z < -1.4$ or $z > 1.4$. In contrast, when the latent mean is three, the density ratio of the $\sigma = 1$ distribution to the $\sigma = 2$ distribution is less than 1 for a much larger region (from $z > -4.7$ to $z < 2.7$). Put simply, when the latent mean is positive, estimates from the precise distribution will more often be associated with large positive Z-statistics than estimates from the imprecise distribution. The larger the region where the density ratio is less than 1, the greater is the latent mean. This is the "reduced form" signature of a positive latent mean in the presence of selectivity. We show this same signature exists in data from Cohen and Ganong (2026) in Appendix F.

The result of Andrews and Kasy (2019) formalizes this intuition, showing the distribution μ of θ is uniquely pinned down by the ratio shown in Equation 17.

Once μ is identified, we can recover $\bar{d}$ (up to a multiplicative constant) from

$$\bar{d}(z) = const. \cdot \frac{f(z|\sigma, D = 1)}{f(z|\sigma)}.$$

Since identification of both μ and $\bar{d}$ is non-parametric, they could in principle be estimated non-parametrically (e.g. using GMM, as discussed in the appendix of Andrews and Kasy 2019). For moderate sample sizes (number of estimates in the meta-study), however, parametric estimation tends to be better behaved.

5.2 Estimation and empirical examples

5.2.1 Testing for selectivity

We investigate the extent of selectivity across all of our running applications. The nature of selectivity depends upon researchers' priors and preferences about the size of the causal effect of interest. In all four of our applications, economic theory or nudge design predicts a *positive* effect of the treatment (e.g. higher re-employment, longer unemployment, higher consumption, increase in outcome targeted by the nudge). The concern regarding selectivity is therefore that researchers and the publication process will select for positive significant effects (rather than positive and negative significant effects).

We report results from three types of tests for selectivity discussed above. For Elliott et al. (2022), we test whether there is a discontinuity in the density of one-sided p-values and Z-statistics (a monotonic transformation of p-values) at $p = 0.025$ (corresponds to $Z = 1.96$) using a local polynomial density estimator (Cattaneo et al., 2020). Estimates are shown in columns 4 and 5 of Table 6.[13] Second, for Egger et al. (1997), we estimate equation 15 and report β, where $\beta \neq 0$ indicates the presence of selectivity. Third, we implement estimators using the model in Andrews and Kasy (2019), both assuming a latent t distribution and making no assumption about the shape of the latent distribution. Given the discussion above about selectivity being specifically likely to generate positive significant effects, we report estimates for $P(\text{publish}|Z < 1.96)/P(\text{publish}|Z > 1.96)$.

We find widespread evidence of selectivity. We find evidence of selectivity—in the sense that we can reject the null hypothesis of no selectivity—for three out of four applions using the tests from Andrews and Kasy (2019) and the test fromEgger et al. (1997).[14] The evidence using the density discontinuity approach

[13] We show histograms of p-values and discuss sensitivity of our results to bandwidth selection in Appendix G.

[14] Card et al. (2018) reports a finding of no selectivity based on the test in Egger et al. (1997); we replicate that finding,

Table 6: Testing for Selectivity in Four Empirical Applications

Study	Regression Coefficient	$P(\text{publish}\|Z < 1.96)/P(\text{publish}\|Z \geq 1.96)$			
	Eggers	MLE Latent t	MLE Nonparametric	Density Discont. (p-value)	Density Discont. (z-statistic)
Card, Kluve, & Weber (2018)	0.01	0.25	0.19	0.07	0.53
	[-0.01, 0.04]	[0.00, 0.57]	[0.04, 0.34]	[0.00, 0.89]	[0.00, 1.38]
Cohen & Ganong (2025)	1.35	0.05	0.13	0.08	0.46
	[0.92, 1.77]	[0.00, 0.11]	[0.01, 0.25]	[0.00, 1.27]	[0.00, 2.50]
Crosta et al. (2024)[1]	0.47	0.28	0.12	0.27	0.8
	[0.40, 0.54]	[0.01, 0.54]	[0.03, 0.21]	[0.00, 1.25]	[0.00, 24.03]
DellaVigna and Linos (2022)[2]	2.71	0.12		0.63	0.73
	[2.29, 3.13]	[0.00, 1.37]		[0.00, 1.93]	[0.00, 2.67]

Notes: This table re-analyzes four different meta-analyses: effect of active labor market policies, the effect of benefits on unemployment duration, the effect of nudges, the effect of cash transfers on consumption. Top row for each study is the estimate; the bottom row is a 95% confidence interval. The first column implements the test proposed in Egger et al. (1997). The second and third columns show the relative publication probability using the test from Andrews and Kasy (2019) for a t distribution of latent estimates and a non-parametric distribution of latent estimates. The 4th column and 5th column show the ratio of density of Z-statistics (p-values) at the $z = 1.96$ ($p = 0.025$) cutoff; we implement suggested methods from Cattaneo et al. (2022) using quadratic fit. [1]For Crosta et al. (2024), MLE latent normal is used, not MLE latent t. We use a linear fit, not quadratic fit for density ratios. [2]For DellaVigna and Linos (2022), we report relative publication probability for the academic journal subsample. We omit the MLE nonparametric estimate due to numerical issues.

from Elliott et al. (2022) is nuanced. In every case, the estimate implies that there is substantial selectivity at the cutoff but we are only once able to reject the null that the density ratio at the cutoff is 1. The fact that we both estimate substantial selectivity and are unable to reject the null indicates the test for discontinuity is low-powered in our contexts.[15] Using the MLE method with a latent t distribution, we find that the probability of reporting insignificant results relative to significant ones varies from 5% to 28% depending on the application.

A finding of selectivity may be surprising in the context of RCTs (nudges and cash transfers) because trial pre-registration should in principle limit the selective reporting of findings. One possibility is that selectivity emerges not in terms of which studies report results, but which outcomes are reported or emphasized by study authors. DellaVigna and Linos (2022) find evidence of selective reporting in terms of the most significant Z-statistic (as opposed to all Z-statistics). Although Crosta et al. (2024) works to standardize outcomes as much as possible, there is still a range of consumption definitions which researchers can report. Sometimes

[15]Elliott et al. (2025) show that in some contexts the power of discontinuity tests can be quite low. So a failure to reject the null of no selectivity should be interpreted as dispositive. We also find that testing for $f(p)$ increasing in p is low-powered (another concern raised by Elliott et al. (2025)).

food is excluded and sometimes it is included, sometimes durables are excluded and sometimes they are included, sometimes the time horizon is one month and sometimes it is one year. Another possibility is that the identifying assumption of orthogonal estimates and standard errors is violated through power calculations, where there are some studies which are projected to have bigger effects for that subpopulation, but estimates for that subpopulation rely on smaller trials.[16]

5.2.2 Estimating the latent distribution of effects in presence of selectivity

We next ask whether selectivity affects the distribution of the latent estimates across our applications in Table 7.

We use two approaches to correct for selectivity.[17] Our first approach analyzes highly powered studies using the approach proposed in Ioannidis et al. (2017). We estimate $\hat{\alpha}$ from equation (15) and then take the subset of estimates with $\sigma_i < |\hat{\alpha}|/2.8$. Ioannidis et al. (2017) say that this approach filters to the studies which will capture the true latent mean (of $\hat{\alpha}$) with 80% power. They then calculate a precision-weighted mean among this subset of "highly-powered" studies. The highly-powered studies method (at least as implemented following the procedure in Ioannidis et al. 2017) only delivers a meaningful subset of studies to analyze in two out of four meta-analyses. The fact that there are very few studies which meet this criteria motivates alternative selection correction methods that use all the available data. In our second approach, we use the model from Andrews and Kasy (2019).

Correcting for selectivity using this method reduces the mean of the latent distribution $\bar{\theta}$ in our four applications. More specifically, the ratio of the corrected mean to the simple mean ranges from 12% to 21% depending on the application.[18] In the context of selection oriented towards positive significant findings (which we believe is a good description of the four meta-analyses that we study), this is precisely what one would expect to find. The distribution of published findings—which features many positive significant findings but has "deleted" some of the positive insignificant and negative insignifcant findings—will have too high of

[16] As one concrete example of this, many of the studies with the highest standard errors have framing which is tied to specific life circumstances (e.g. small business entrepreneurship or children). One might plausibly think that there are larger effects of cash in this subpopulation (indeed this is the motivation for targeting in the first place).

[17] We do not use the correction approach from equation (15) because, as shown in Section 5.1.2, it is not identified; we do not use the correction approach from MLE nonparametric since we do not believe it is robust in our settings.

[18] Our estimates imply that the latent distribution for DellaVigna and Linos (2022) is Cauchy. The mean is undefined for a Cauchy distribution. In this case, $\bar{\theta}$ can be interpreted as the median.

a mean. Selectivity can therefore lead an unsophisticated literature review to overstate the average causal effects of the policies studied.

Table 7 also reveals that the precision weighting and incorporating fat tails from Section 3 provide a partial back door approach to selection correction. If the true data-generating process involves selectivity in favor of positive significant findings, then estimates with large standard errors will on average be more biased than estimates with small standard errors. Precision-weighting reduces the influence of these more-biased estimates. Our findings also reflect this pattern: the precision-weighted means (shown earlier in Table 4 and repeated for convenience in Table 7) are smaller than the simple means in all four applications.[19] This is what we would expect to see in the presence of selectivity; in the absence of selectivity, the precision-weighted means could plausibly be larger or smaller than the simple means.

Last but not least, it is useful to note that the empirical estimates from the MLE latent t model falsify an identifying assumption from the highly-powered approach. The problem, which we describe in the theory section, is that if there are some latent effects θ_i which are zero, then no standard error will be sufficiently small as to qualify as "highly powered." Across all applications, the distributions implied by the estimates for $\bar{\theta}$ and τ imply that there exists a meaningful distribution of latent effects in the neighborhood of zero.

6 A model with selectivity and covariates

6.1 Theory

Many meta-analyses want to combine covariates and selectivity – that is, they want a predictive model for new problem instances characterized by covariates that also accounts for selectivity.

Our discussion in subsection 5.1 sidestepped study level covariates. Covariates can, however, be incorporated either by partitioning the observed studies and then separately implementing any of the selectivity methods[20] or by directly building

[19]This has the interesting implication that selection correction will *not* always change the conclusion of a meta-analysis. As one concrete example, Crosta et al. (2024) focus on MLE estimates which precision weight each estimate (but do not correct for selectivity). In that case, the simple mean for the estimated average effect of $1 on monthly consumption is $0.12, the estimated effect when using MLE without correcting for selectivity is $0.03, and the estimated effect of using the MLE with correcting for selectivity of $0.02. So even without correcting for selectivity, precision weighting generates similar conclusions about the mean of the effect distribution to a selection correction.

[20]See for example Brodeur et al. (2020), who consider p-curves separately depending on the methods employed in different empirical studies.

Table 7: Estimating Mean Average Effects and Dispersion in Presence of Selectivity

Hyper-parameter	Simple Mean	Prec-wtd. Mean (eqn 2)	MLE Latent t	Model latent distribution with selectivity: Highly Powered	Model latent distribution with selectivity: MLE Latent t
Card, Kluve, & Weber (2018) – Effect of Active Labor Market Program on Employment					
$\bar{\theta}$ (Mean)	0.083	0.040	0.057	0.040	0.010
	(0.008)	(0.007)	(0.020)	(0.001)	(0.016)
% of Simple Mean		48%	69%	48%	12%
τ (Between-Study Dispersion)		0.000	0.064		0.030
		(2.615)	(0.023)		(0.014)
ν (Degrees of Freedom)			4.11		1.49
			(4.13)		(0.37)
Cohen & Ganong (2025) – Elasticity of Unemp Duration w.r.t. Unemp Benefits					
$\bar{\theta}$ (Mean)	0.513	0.448	0.382	0.350	0.107
	(0.050)	(0.044)	(0.046)	(0.003)	(0.089)
% of Simple Mean		87%	74%	68%	21%
τ (Between-Study Dispersion)		0.345	0.240		0.244
		(0.059)	(0.032)		(0.053)
ν (Degrees of Freedom)			4.04		2.89
			(1.69)		(1.14)
Crosta et al. (2024)[1] *– Effect of $1 Transfer on Monthly Consumption*					
$\bar{\theta}$ (Mean)	0.127	0.011	0.028		0.018
	(0.049)	(0.004)	(0.004)		(0.005)
% of Simple Mean		9%	22%		14%
τ (Between-Study Dispersion)		0.000	0.020		0.019
		(2.212)	(0.003)		(0.003)
DellaVigna and Linos (2022) – Effect of Nudge on Targeted Outcome					
$\bar{\theta}$ (Mean)	0.031	0.026	0.004		0.005
	(0.004)	(0.003)	(0.001)		(0.026)
% of Simple Mean		84%	13%		16%
τ (Between-Study Dispersion)		0.070	0.005		0.010
		(0.002)	(0.001)		(0.122)
ν (Degrees of Freedom)			1.00		1.28
			(0.08)		(6.84)

Notes: This table re-analyzes four different meta-analyses. The first column shows the simple mean. The second column shows estimates from equation 2 (which weights observations by $\frac{1}{\tau^2+\sigma_i^2}$). The third column assumes a latent t distribution and uses Maximum Likelihood Estimation. The fourth column implements Stanley et al. (2017) (which weights observations by $\frac{1}{\sigma_i^2}$). The number of highly-powered estimates is, in descending order of rows, 31/169, 45/71, 2/75, and 6/315. We omit the highly-powered estimates for Crosta et al. (2024) and DellaVigna and Linos (2022) since there are too few studies. The fifth column implementing Andrews and Kasy (2019) with a latent t distribution is a direct counterpart to the third column.
[1]For Crosta et al. (2024), MLE latent normal is used, not MLE latent t. MLE latent t estimates are similar with estimated *df* very large.

covariates into the models used. The former approach requires sample sizes larger than those present in our empirical examples so we will focus on the latter.

Likelihood Recall the assumptions stated in section 2, where we assumed a sampling distribution of $\hat{\theta}_i|\theta_i, \sigma_i^2, X_i \sim N(\theta_i, \sigma_i^2)$, and selection based solely on Z-statistics $Z_i = \hat{\theta}_i/\sigma_i$ according to $P(D_i = 1|\hat{\theta}_i, \sigma_i, X_i) = \bar{d}(Z_i)$.

Suppose now additionally, as in the meta-regression model (equation 5), that

$$\theta_i|X_i = x \sim N(x \cdot \beta, \tau^2).$$

Then, again as before, $\hat{\theta}_i|X_i, \sigma_i \sim N(x \cdot \beta, \tau^2 + \sigma_i^2)$, and thus

$$Z_i|X_i, \sigma_i \sim N(x \cdot \beta/\sigma_i, \tau^2/\sigma_i^2 + 1).$$

We get the conditional likelihood of published Z-statistics by Bayes rule (this implicitly also conditions on model parameters β, τ^2 and $\bar{d}$),

$$f(Z_i|X_i, \sigma_i, D_i = 1) = \underbrace{\frac{1}{\sqrt{\frac{\tau^2}{\sigma_i^2} + 1}} \varphi\left(\frac{Z_i - X_i \cdot \beta}{\sqrt{\frac{\tau^2}{\sigma_i^2} + 1}}\right)}_{f(Z_i|X_i, \sigma_i)} \cdot \underbrace{\frac{\bar{d}(Z_i)}{E[\bar{d}(Z)|X_i, \sigma_i]}}_{\frac{P(D_i=1|Z_i, \sigma_i, X_i)}{P(D_i=1|X_i, \sigma_i)}}.$$

If we consider a parametric specification for $\bar{d}$ of the form $\bar{d}(z) = \mathbf{1}(z > 1.96) + \gamma \cdot \mathbf{1}(z \leq 1.96)$ (selection for positive significant effects), then the denominator in the last fraction equals

$$E[\bar{d}(Z)|X_i, \sigma_i] = 1 + (\gamma - 1) \cdot \Phi\left(\frac{1.96 - X_i \cdot \beta}{\sqrt{\frac{\tau^2}{\sigma_i^2} + 1}}\right).$$

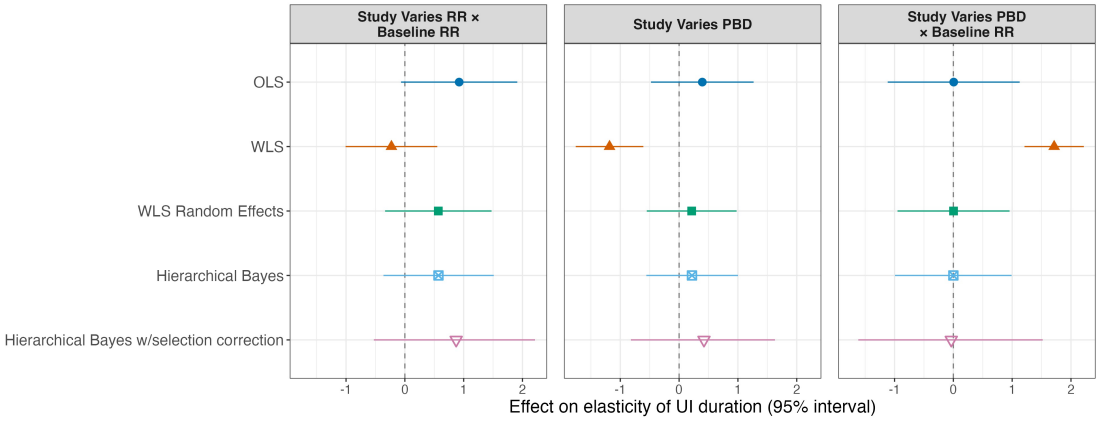
Prior and posterior We can combine this likelihood with a prior for the hyper-parameters β, τ^2 and γ to get a complete hierarchical Bayes specification. We can then use Hamiltonian Monte Carlo (as implemented in Stan) to sample from the posterior. Using the posterior mean $\bar{\beta}$ of β, we can furthermore make predictions for new instances, via $X_0 \cdot \bar{\beta}$.

For our empirical application, we assume a normal prior for β, $\beta \sim N(0, \Sigma)$, as in the Ridge regression models discussed earlier, a half-normal prior for τ^2 with large variance, and a lognormal prior for γ, $\log(\gamma) \sim N(0, 1)$.

6.2 Empirical example of a model with selectivity and covariates

The meta-analyses that want to combine covariates and selectivity include three out of four of our applications: Cohen and Ganong (2026), DellaVigna and Linos (2022), Card et al. (2018).[21] Cohen and Ganong (2026) in particular is interested in the optimal replacement rate for unemployment insurance. It finds that higher baseline replacement rates (X_i) are associated with a higher elasticity of unemployment duration with respect to benefits ($\hat{\theta}_i$). In the model they study, solving for the optimal replacement rate requires predicting how the elasticity varies with baseline replacement rate.

Figure 10: Correcting Covariate Coefficients for Selectivity (Cohen and Ganong, 2026)



Notes: This figure shows the coefficients and the 90% confidence interval of the coefficients for $\hat{\theta}_i = \alpha + \beta_1$ Baseline RR $\times$ **1**(Study Varies RR) $+ \beta_2$ Baseline RR $\times$ **1**(Study Varies PBD) $+ \beta_3$ **1**(Study Varies PBD) $+ \epsilon_i$. RR stands for replacement rate; PBD stands for potential benefit duration. The data is from Cohen and Ganong (2026) and the circles represent regression models; squares represent hierarchical Bayes models. From the top, the first model does an equal weight regression; the second model does WLS with $w_i = \frac{1}{\sigma_i^2}$; the third model does WLS with $w_i = \frac{1}{\sigma_i^2+\tau^2}$ where τ^2 is the posterior mean from the fourth model; the fourth model estimates a hierarchical Bayes model with diffuse priors and no selection correction; the fifth model estimates a hierarchical Bayes model with diffuse priors and a selection correction. When hierarchical Bayes is used, the posterior mean of the coefficients is used as the estimate.

Figure 10 builds up to the model described in subsection 6.1. We start with a multivariate OLS regression. Cohen and Ganong (2026) find that estimates for the elasticity of unemployment duration with respect to replacement rate are larger in contexts with a larger baseline replacement rate. We replicate this finding by

[21] The first article studies the combination using Bayesian Model Averaging, the second article studies the combination by using a complex simulation procedure (Table VI), and the third article is a priori interested in both questions, but did not find evidence of selectivity and therefore views covariate-only models as sufficient.

running the minimal regression needed to investigate this interaction effect:

$$\begin{aligned}\hat{\theta}_i =& \alpha + \beta_1 \,\text{Baseline RR} \times \mathbf{1}(\text{Study Varies RR}) \\ &+ \beta_2 \,\text{Baseline RR} \times \mathbf{1}(\text{Study Varies PBD}) + \beta_3 \,\mathbf{1}(\text{Study Varies PBD}) + \epsilon_i\end{aligned}$$

where "RR" is replacement rate and "PBD" is potential benefit duration.

Then, we account for differing precision by switching to weighted least squares with weights $w_i = \frac{1}{\sigma^2}$. This leads to large changes in the coefficients and in particular shows no difference in elasticity with respect to RR for changes in the baseline RR. Accounting further for heterogeneity through weighting by $w_i = \frac{1}{\sigma_i^2+\tau^2}$ (with the τ^2 obtained by estimating the hierarchical Bayes model without a selection correction) leads to more equal weights across studies; generally this will lead to coefficent estimates that fall between those of OLS and WLS with $w_i = \frac{1}{\sigma^2}$.[22] We get identical results by running the hierarchical Bayes model without a selection correction.

The final model accounts for selection. We find that when the replacement rate increases by 10%, the elasticity increases by 8.6%. The hierarchical Bayes model without a selection correction would estimate an elasticity increase of 5.6%. The selection-corrected model implies an optimal replacement rate of 32% in the US context studied by Cohen and Ganong (2026) if all other factors except the slope of the elasticity vs. baseline replacement rate are held constant. The uncorrected model implies an optimal replacement rate of 29%.

7 Cookbook

This section collects our recommendations into a single pipeline. The pipeline runs in six stages, summarized in Figure 11: (0) scoping and data, (1) aggregation, (2) prediction with covariates, (3) selectivity, (4) the combined model, and (5) decisions.

Stage 0 – Scope, collect, assess credibility. Fix the estimand θ_0, the search procedure, and the coding rules and write down the proposed procedure. This matters because apparent selectivity can be an artifact of which *outcome* was coded rather than which study was published, as in the cash-transfer application (subsection 5.2). AI tools make extraction cheap; see Cook et al. (2026) for

[22] In a multivariate regression this is not a rule – as demonstrated by this relationship failing with covariate "PBD x baseline RR" – because changes in the values of other coefficients will change the coefficients of collinear covariates. The hierarchical Bayes model without a selection correction almost exactly replicates this WLS with $w_i = \frac{1}{\sigma_i^2+\tau^2}$.

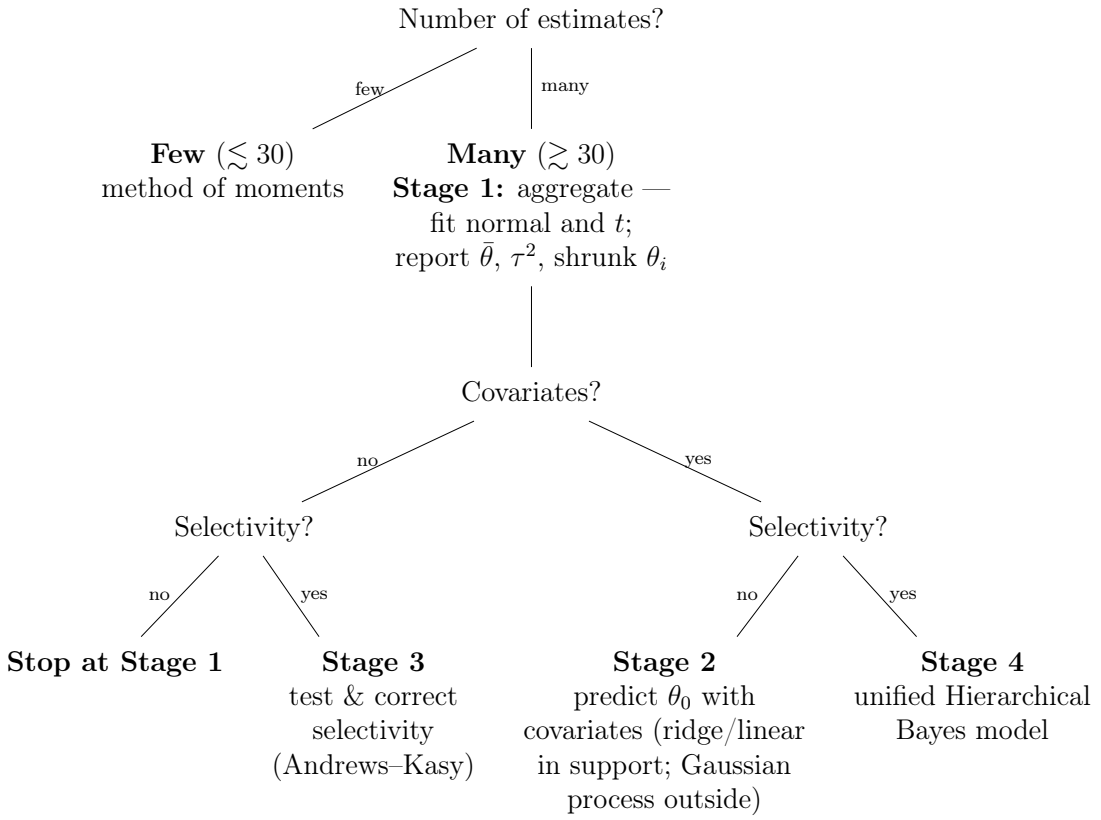


Figure 11: Decision tree for the cookbook pipeline.

pitfalls. The methods in this paper make sense when there are at least three prior estimates, so you should stop if you cannot find three prior estimates.

At this stage, also decide whether to collect covariates and which covariates to collect, since it determines the model in Stage 3. Do not discard lower-quality studies. Instead, collect measures of study quality as part of the covariates (and return to this question in step 3). This could be something relatively objective like the research design (e.g. DiD, RD, RCT), a subjective assessment of the credibility of the paper's identifying assumptions, or both.[23]

Stage 1 – Aggregate without covariates or selection. Ask whether the dispersion of the $\hat{\theta}_i$ exceeds what their standard errors imply (Equation 1). Report

[23]Although we recommend this practice, we did not implement it as part of our empirical examples in this article. That is because the notion of study quality is difficult to treat in a consistent way across very different applications. We advocate for using the meta-analyst's assessment of quality (rather than say journal quality, which could be related to selectivity Kasy (2021).

a precision-weighted mean (Equation 2), not a simple mean, and form shrunken study-level estimates (Equation 3). Whether you proceed to the subsequent steps depends on the number of estimates you have collected. The subsequent steps need a soft floor of roughly thirty estimates (Irsova et al. 2024).[24]

Next, if you have enough studies, fit and compare three latent distributions (Table 4): normal , t for heavy tails, and nonparametric via nonparametric maximum with Tweedie posterior means (Equation 4). Read off whether tails are heavy (small ν) and whether μ is skewed – these choices materially move the estimated mean and shrink τ^2.

Stage 2 – Predict with covariates. Standardize covariates by their empirical standard deviation. For a context *inside* the support of the covariates, use a meta-regression estimated as random effects WLS (Equation 5), equivalently the Gaussian-process posterior mean (Equation 7); the equivalence makes explicit that the prediction is a transparent weighted average of the $\hat{\theta}_i$. Although we emphasize the importance of distributions that allow for fat tails in stage 1, this becomes less important once covariates are available; this is because it is likely that the distribution is assumed to be normal *after conditioning on the prediction* $X_i\beta$ *or* $\bar{\theta}(X_i)$, where the covariates x can explain why a specific estimate $\hat{\theta}_i$ is an outlier.

For extrapolation *outside* the support, use a Gaussian process prior with a squared-exponential kernel (equation 9). Make an active decision about whether to use a long or short length scale. The short length scale will lead to a much larger posterior standard deviation. This choice should be informed by your economic model of how outcomes are determined (see discussion in Section 4.2.2 as an example).

A second use of covariates is to incorporate study quality into the prediction problem. If you are interested in the likely effects of a policy, do the prediction exercise where θ_0 is defined as the highest quality strata of study.[25] The researcher's choice of prior will decide how much consideration to give to the studies from the lower-quality strata. Another option is to do the pipeline for just the high-quality studies; this is implicitly what is already done in the context of the small-sample meta-analysis (Tables 1, 2, and 3).

For any such prediction model for θ_i given X_i, it is important to keep in mind that (i) the resulting coefficients are not causal and (ii) any coefficients in a multivariate regression capture variation *holding all other regressors constant.*

[24]We take the upper bound from Irsova et al. (2024) who suggest at least 10 primary studies with 30 total estimates.

[25]As one example, Cohen and Ganong (2026) report predictions for a regression discontinuity design when doing policy analysis.

Stage 3 – Test and correct for selectivity. Assess whether you believe that equations (13) and (14) are satisfied.[26] If they are not satisfied, then you are limited to the a p-curve density-discontinuity test (Elliott et al., 2022). However, it is important to know that discontinuity tests can have low power (Elliott et al., 2025), so non-rejection may not imply the absence of selectivity. If equations (13) and (14) are satisfied, you can also use the Egger et al. (1997) meta-regression, and the relative publication probability from Andrews and Kasy (2019).

For *correction*, we emphasize that two commonly used methods are invalid. The Egger intercept is a valid test but not a valid bias correction – under selection, $E[Z_i \mid \sigma_i]$ is nonlinear in $1/\sigma_i$ and linear extrapolation to $\sigma_i = 0$ can return almost any value (Figure 5). The highly-powered-studies estimator (Stanley et al., 2017; Ioannidis et al., 2017) is valid only when such studies exist; when μ has mass near zero they may not, as in two of our four applications (Table 7). Our workhorse is the Andrews and Kasy (2019) model, which jointly estimates μ and the selection function from the joint distribution of $\hat{\theta}_i$ and σ_i (section 5.1.2).

One note of caution regarding the correction method of Andrews and Kasy (2019) is that it can deliver results which are fragile to assumptions about the distribution of latent effects as well as the chosen thresholds for selective reporting. Although we report only one selectivity-corrected MLE model here for brevity, we recommend that authors assess sensitivity to both assumptions about the distribution of latent effects as well as chosen thresholds. See Cohen and Ganong (2026) for an example of such sensitivity tests.

Stage 4 – Combine covariates and selectivity. Build covariates into the selection model: the linear meta-regression $\theta_i \mid X_i \sim N(X_i \cdot \beta, \tau^2)$ combined with a step-function $\bar{d}$, estimated by Hamiltonian Monte Carlo in Stan. We recommend doing this with the `baggr` package. Build it up in visible steps – OLS, WLS, random effects WLS, hierarchical Bayes without and then with selection (Figure 10) – so the reader sees what each ingredient does.

Optional Stage 5 – From estimates to decisions. For a binary choice, carry the prediction through to the rule rather than stopping at a coefficient. Under welfare $U(A_i, \theta_i) = A_i\theta_i$, choose $A_i = \mathbf{1}\Big(E[\theta_i \mid \hat{\theta}_i, \sigma_i, X_i] \geq 0\Big)$ for an evaluated site and $A_0 = \mathbf{1}\Big(E[\theta_0 \mid X_0, \hat{\theta}_{1:n}] \geq 0\Big)$ for a new context (equation Equation 4.1.4).

[26] As an example of where they may not be satisfied, Dube and Zipperer (2024) find in the context of minimum wage own-wage elasticities, "credible designs may use less minimum wage variation, resulting in lower precision, but they may also have less bias". In randomized control trials, researchers have priors about the size of the causal effect they use in power calculations to choose sample sizes, leading to larger expected causal effects being systematically less precise. See Chen (2026) for an example of a model with co-dependence (but is not about meta-analysis)

References

Allcott, H. (2015). Site selection bias in program evaluation. *Quarterly Journal of Economics*, 130(3):1117–1165.

Andrews, I. and Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–94.

Armstrong, T. B., Kolesár, M., and Plagborg-Møller, M. (2022). Robust empirical bayes confidence intervals. *Econometrica*, 90(6):2567–2602.

Athey, S., Chetty, R., Imbens, G. W., and Kang, H. (2025). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *The Review of Economic Studies*, page rdaf087.

Bailey, M., Byker, T., Patel, E., and Ramnath, S. (2025). The long-run effects of california's paid family leave act on women's careers and childbearing: New evidence from a regression discontinuity design and US tax data. *American Economic Journal: Economic Policy*, 17(1):401–431.

Baker, M., Halberstam, Y., Kroft, K., Mas, A., and Messacar, D. (2023). Pay transparency and the gender gap. *American Economic Journal: Applied Economics*, 15(2):157–183.

Bandiera, O., Parekh, N., Petrongolo, B., and Rao, M. (2021). Men are from mars, and women too: A Bayesian meta-analysis of overconfidence experiments. *IZA Discussion Paper*, (14950). Later published in Economica 89(355), 678–704, 2022.

Bennedsen, M., Simintzi, E., Tsoutsoura, M., and Wolfenzon, D. (2022). Do firms respond to gender pay gap transparency? *The Journal of Finance*, 77(4):2051–2091.

Bento, A., Freedman, M., and Lang, C. (2015). Who benefits from environmental regulation? evidence from the clean air act amendments. *The Review of Economics and Statistics*, 97(3):610–622.

Blundell, J., Duchini, E., Simion, Ş., and Turrell, A. (2025). Pay transparency and gender equality. *American Economic Journal: Economic Policy*, 17(2):418–445.

Brodeur, A., Cook, N., and Heyes, A. (2020). Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11):3634–60.

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., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Card, D., Kluve, J., and Weber, A. (2018). What works? a meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, 16(3):894–931.

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

Cattaneo, M. D., Jansson, M., and Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531):1449–1455.

Cattaneo, M. D., Jansson, M., and Ma, X. (2022). lpdensity: Local polynomial density estimation and inference. *Journal of Statistical Software*, 101(2):1–25.

Chay, K. Y. and Greenstone, M. (2005). Does Air Quality Matter? Evidence from the Housing Market. *Journal of Political Economy*, 113(2):376–424.

Chen, J. (2026). Empirical Bayes when estimation precision predicts parameters. *Econometrica*, 94(2):305–340. Working paper version: arXiv:2212.14444, 2025.

Chetty, R., Hendren, N., Kline, P., and Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *Quarterly Journal of Economics*, 129(4):1553–1623.

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

Christensen, T., Moon, H. R., and Schorfheide, F. (2026). Optimal decision rules when payoffs are partially identified. *The Review of Economic Studies*. Manuscript MS33570.

Cohen, J. P. and Ganong, P. (2026). Disemployment effects of unemployment insurance: A meta-analysis. *American Economic Review: Insights*, 8(1):1–18.

Cook, N., Bartoš, F., Bom, P. R. D., Gechert, S., Kantová, K., Geyer-Klingeberg, J., Havránek, T., Iršová, Z., Lušková, M., Opatrný, M., Prante, F., Rachinger, H. J., and Stanley, T. D. (2026). Reporting guidelines for meta-analysis in economics—updated for ai. *Journal of Economic Surveys*. Forthcoming.

Crosta, T., Karlan, D., Ong, F., Rüschenpöhler, J., and Udry, C. R. (2024). Unconditional cash transfers: A bayesian meta-analysis of randomized evaluations in low and middle income countries. Working Paper 32779, National Bureau of Economic Research.

DellaVigna, S. and Linos, E. (2022). Rcts to scale: Comprehensive evidence from two nudge units. *Econometrica*, 90(1):81–116.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188.

Dube, A. and Zipperer, B. (2024). Own-wage elasticity: Quantifying the impact of minimum wages on employment. Working Paper 32925, National Bureau of Economic Research.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614.

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

Elliott, G., Kudrin, N., and Wüthrich, K. (2022). Detecting p-hacking. *Econometrica*, 90(2):887–906.

Elliott, G., Kudrin, N., and Wüthrich, K. (2025). The Power of Tests for Detecting p-Hacking. *The Review of Economics and Statistics*, pages 1–44. _eprint: https://direct.mit.edu/rest/article-pdf/doi/10.1162/REST.a.1688/2572274/rest.a.1688.pdf.

Frankel, A. and Kasy, M. (2022). Which findings should be published? *American Economic Journal: Microeconomics*, 14(1).

Gechter, M., Hirano, K., Jiang, J., and Samii, C. (2024). Selecting experimental sites for external validity. Working paper, Penn State and NYU, September 2024.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Taylor & Francis.

Gulyas, A., Seitz, S., and Sinha, S. (2023). Does pay transparency affect the gender wage gap? evidence from austria. *American Economic Journal: Economic Policy*, 15(2):236–255.

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

Irsova, Z., Doucouliagos, H., Havranek, T., and Stanley, T. (2023). Meta-analysis of social science research: A practitioner's guide. *Journal of Economic Surveys*.

Irsova, Z., Doucouliagos, H., Havranek, T., and Stanley, T. D. (2024). Meta-analysis of social science research: A practitioner's guide. *Journal of Economic Surveys*, 38(5):1547–1566.

Ishihara, T. and Kitagawa, T. (2024). Evidence aggregation for treatment choice. *arXiv preprint arXiv:2108.06473*. Revised July 2024.

Jarosch, G. and Pilossoph, L. (2019). Statistical discrimination and duration dependence in the job finding rate. *Review of Economic Studies*, 86(4):1631–1665.

Kasy, M. (2021). Of forking paths and tied hands: Selective publication of findings, and what economists should do about it. *Journal of Economic Perspectives*, 35(2).

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109(506):674–685.

Kospentaris, I. (2021). Unobserved heterogeneity and skill loss in a structural model of duration dependence. *Review of Economic Dynamics*, 39:280–303.

Kremer, M., Luby, S. P., Maertens, R., Tan, B., and Wik ecek, W. (2023). Water treatment and child mortality: A meta-analysis and cost-effectiveness analysis. Working Paper 30835, National Bureau of Economic Research.

Kroft, K., Lange, F., and Notowidigdo, M. J. (2013). Duration dependence and labor market conditions: Evidence from a field experiment. *Quarterly Journal of Economics*, 128(3):1123–1167.

Lund, C., Orkin, K., Witte, M., Walker, J. H., Davies, T., Haushofer, J., Murray, S., Bass, J., Murray, L., Tol, W., and Patel, V. H. (2024). The effects of mental health interventions on labor market outcomes in low- and middle-income countries. Working Paper 32423, National Bureau of Economic Research.

Manski, C. F. (2020). Toward credible patient-centered meta-analysis. *Epidemiology*, 31(3):345–352.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91.

Meager, R. (2022). Aggregating distributional treatment effects: A bayesian hierarchical analysis of the microcredit literature. *American Economic Review*, 112(6):1818–47.

Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. Springer.

Montiel Olea, J. L., Qiu, C., and Stoye, J. (2026). Decision theory for treatment choice problems with partial identification. *Review of Economic Studies*. Forthcoming.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):pp. 47–55.

Mullins, J. (2025). A structural meta-analysis of welfare reform experiments and their impacts on children. *Journal of Political Economy*, 134(1).

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4):377–401.

Sager, L. and Singer, G. (2025). Clean identification? the effects of the clean air act on air pollution, exposure disparities, and house prices. *American Economic Journal: Economic Policy*, 17(1):1–36.

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

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

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

Stanley, T. D. and Jarrell, S. B. (1989). Meta-regression analysis: A quantitative method of literature surveys. *Journal of Economic Surveys*, 3(2):161–170.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Vivalt, E. (2019). How much can we generalize from impact evaluations? *Journal of the European Economic Association*, forthcoming.

Vodopivec, M. (1999). Does the slovenian public work program increase participants' chances to find a job? *Journal of Comparative Economics*, 27(1):113–130.

Williams, C. and Rasmussen, C. (2006). *Gaussian processes for machine learning*. MIT Press.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods: invited paper. *The Annals of Statistics*, 31(2):379–390.

Appendixes – for online publication only

A Derivation of Tweedie's formula **59**

B "Small" meta-analysis **61**
B.1 Elasticity of Housing Prices with Respect to Air Pollution 61
B.2 Effect of Pay Transparency Laws on the Gender Pay Gap 61
B.3 Effect of Paid Family Leave on Mother's Earnings 61
B.4 Additional settings from AEJ 2025 where small sample meta-analysis is feasible 62

C Binary implementation decisions **66**

D Covariates **68**

E Smoothness and monotonicity of the p-curve **69**

F Density ratio in Cohen and Ganong (2026) **70**

G Density Estimation and Bandwidth Selection **72**

A Derivation of Tweedie's formula

This appendix gives the derivation of Tweedie's formula, equation (4), summarized in Section 3.1.2. To describe the argument without excessive notation, let us consider the normalized parameters $\omega_i = \theta_i/\sigma_i$. Denote the distribution of the normalized parameters ω_i by ν. Then the marginal density of the Z-statistics $Z_i = \hat{\theta}_i/\sigma_i$ (absent p-hacking or publication bias) is given as follows:

$$\begin{aligned} Z_i|\omega_i &\sim N(\omega_i, 1) \\ \omega_i &\sim \nu \\ F(z) = P(Z_i \leq z) &= \int \Phi(z-\omega) d\nu(\omega) \\ f(z) = \frac{\partial P(Z_i \leq z)}{\partial z} &= \int \varphi(z-\omega) d\nu(\omega) \end{aligned} \tag{19}$$

Mathematically, we say that the distribution of Z is given by the convolution of ν with the standard normal density φ. Note that the distribution of Z is directly identified, and we can estimate the density $f(z)$ using any of a number of methods. We can correspondingly also estimate the distribution ν of ω. This is known as deconvolution, cf. Meister (2009). Perhaps surprisingly, non-parametric maximum likelihood for ν also yields a well behaved estimator in this setting (Koenker and Mizera, 2014).

An interesting property of the normal distribution allows us to relate the density $f(z)$ to the empirical Bayes estimate $E[\omega_i|Z_i = z]$ of ω_i. Since $\varphi(z) = (2\pi)^{-1/2} \cdot \exp\left(-\frac{z^2}{2}\right)$, we have that $\varphi'(z) = -z \cdot \varphi(z)$. This property allows us to derive Tweedie's formula (Efron, 2011):

$$\frac{\partial \log(f(z))}{\partial z} = \frac{\int \varphi'(z-\omega) d\nu(\omega)}{\int \varphi(z-\omega) d\nu(\omega)} = -\frac{\int (z-\omega)\varphi(z-\omega) d\nu(\omega)}{\int \varphi(z-\omega) d\nu(\omega)} = -E[Z_i - \omega_i | Z_i = z],$$

and thus

$$E[\omega_i|Z_i = z] = z + \frac{\partial \log(f(z))}{\partial z}. \tag{20}$$

We thus obtain an expression for the empirical Bayes estimator which applies for any distribution of ω.

Generalizing this argument to the setting with covariates and to the unnormalized estimates $\hat{\theta}_i$, we recover equation (4), and more generally

$$E\left[\theta_i|\hat{\theta}_i, \sigma_i, X\right] = \hat{\theta}_i + \sigma_i^2 \cdot \frac{\partial}{\partial \hat{\theta}} \log(f(\hat{\theta}_i|\sigma_i, X_i)).$$

Here $f(\hat{\theta}|\sigma, X)$ is the conditional density. Both the density $f(z)$ and the conditional density $f(\hat{\theta}|\sigma, X)$ can, in principle, be estimated parametrically or non-parametrically, from the observed empirical distribution of $(\hat{\theta}_i, \sigma_i, X_i)$. In practice, however, estimation in the tails and estimation of conditional densities might be poorly behaved.

B "Small" meta-analysis

B.1 Elasticity of Housing Prices with Respect to Air Pollution

Panel A: estimates use nonattainment designation as an instrumental variable for pollution, with change in log house prices as the outcome. Chay and Greenstone (2005) regressor is Total Suspended Particles (PM_{100}): flexible-controls specification, Table 5 Panel A col. 3. Bento et al. (2015) regressor is PM_{10}: elasticity stated in Section V.C, derived from the overidentified IV coefficient of 1.33 (SE = 0.50) on ΔPM_{10} in Table 3. Grainger (2012) regressor is PM_{10}: elasticity derived from the IV coefficient of -0.0145 (SE = (0.0058) in Table 5 Panel A col. 1. Converted to elasticity using the 1990 PM_{10} mean reported in Table 1.

Panel D reports Sager and Singer (2025) regressor is $PM_{2.5}$: preferred M1DiD-IV specification, Table 3 Panel A. SE of implied elasticity computed as $\widehat{\text{SE}}_{\varepsilon} = \widehat{\text{SE}}_{\hat{\theta}} \times \bar{c}$, where $\bar{c}$ is the mean pollutant concentration implied by the reported elasticity.

B.2 Effect of Pay Transparency Laws on the Gender Pay Gap

Panel A: estimates are percent effect relative to conditional pre-policy gender pay gap. Baker et al. (2023) average over four specification with gender pay gap as dependent variable: Table 4 col. 1-4. Converted to percent effect using the conditional gender pay gap of 6 percent at the time of the reform (Figure 2). Bennedsen et al. (2022) underlying coefficient comes from Table 4 col. 2. Converted to percent effect from their statement that the estimate represented a 13% reduction in the conditional gender pay gap. Gulyas et al. (2023) estimate derived from their statement that they can rule of 0.4pp effect at the 95% percent confidence level.

Panel D reports Blundell et al. (2025) treated times post regressor: Table 3 col. 1.

B.3 Effect of Paid Family Leave on Mother's Earnings

All estimates are the effect of paid-leave *availability* (the intent-to-treat or reduced-form effect of the policy, not of individual take-up of leave), expressed as a percent of the control-group mean. Studies that report an availability or difference-in-differences coefficient are used directly; studies that report a take-up effect (two-stage least squares or a local average treatment effect) are converted to the reduced form by multiplying by the paper's own first-stage take-up.

Panel A: Rossin-Slater et al. (2013) is the intent-to-treat effect on log annual wage income for mothers of one-year-olds, Table 7. Campbell et al. (2017) is the two-stage least squares effect on annual wages in quarters 4–7 for the claimants sample, Table 3, converted to a reduced form by multiplying by the first-stage take-up of 0.817 (crossing the eligibility threshold raises the probability of receiving paid leave by 81.7 percentage points). Baum and Ruhm (2016) is the difference-in-differences effect on log hourly wages one year after birth for mothers employed at all during pregnancy, Table 5 of the published version. Timpe (2024) is the difference-in-differences effect on log hourly wages over the first five years for women aged 18–45, Table 3 column 1. The annual-earnings studies (Rossin-Slater et al., Campbell et al., and Bailey et al.) and the hourly-wage studies (Baum and Ruhm, and Timpe) share the percent-effect-on-pay interpretation but are not identical objects.

Panel E reports Bailey et al. (2025), Table 3, Panel B (cumulative real wage earnings in the short run, years 1–3). To match the all-mothers populations of the prior studies, we pool Bailey's first-birth and higher-order-birth estimates, weighting each by its sample size: the first-birth take-up effect is -12.6% (SE 5.7) on $n = 283{,}594$ and the higher-order-birth take-up effect is -0.5% (SE 7.3) on $n = 441{,}589$, giving a pooled take-up effect of -5.2%. The standard error of the pool follows from the delta method. The pooled take-up effect is then converted to the availability scale via Bailey's first stage — women who could take California's paid family leave consecutively after disability leave were 16 percentage points more likely to take it up — giving a reduced-form effect of $-5.2\% \times 0.16 = -0.8\%$ (SE 0.8).

B.4 Additional settings from AEJ 2025 where small sample meta-analysis is feasible

Table 8: Meta-analysis Candidates in the 2025 American Economic Journal: Economic Policy

Authors and year	Paper title	Quantity of interest	No. of other papers	Other estimates of quantity of interest among already-cited papers (first four shown)
Bailey et al. (2025a)	The Long-Run Effects of California's Paid Family Leave Act on Women's Careers and Childbearing: New Evidence from a Regression Discontinuity Design and US Tax Data	Effects of CA paid leave act on labor market outcomes (employment, childbearing)	4	Rossin-Slater et al. (2013); Campbell et al. (2018); Baum & Ruhm (2016); Timpe (2024)
Bailey et al. (2025b)	Show Me the Money! A Field Experiment on Electric Vehicle Charge Timing	Price elasticity of electric vehicle charging time	3	Burkhardt, Gillingham, Kopalle (2023); Qiu et al (2022); Ito, Ida, Tanaka (2018)
Bhuller et al. (2025)	Mental Health Consequences of Correctional Sentencing	Effect of rehabilitative sentencing on inmate mental health	5	Binswanger et al (2007); Weidner, Schultz (2019); Haglund et al (2014); Sailas et al (2006); ...
Bilicka and Güçeri (2025)	Dividend Taxation and Firm Performance with Heterogeneous Payout Responses	Dividend reduction after introduction of a tax	4	Alstadsæter, Fjærli (2009); Bach, et al. (2023) - led to retraction of BM 2022; Alstadsæter, Kopczuk, Telle (2014); Hanlon, Hoopes (2014)
Bilicka and Güçeri (2025)	Dividend Taxation and Firm Performance with Heterogeneous Payout Responses	Elasticity of dividend response	3	Yagan (2015); Chetty, Saez (2006); Poterba (1987)
Blundell et al. (2025)	Pay Transparency and Gender Equality	Decrease in pay gap after transparency act	3	Baker et al. (2023); Bennedsen et al. (2022); Gulyas et al. (2023)
Brough et al. (2025)	Eliminating Fares to Expand Opportunities: Experimental Evidence on the Impacts of Free Public Transportation on Economic and Social Disparities	Employment effects of free fares	3	Phillips (2014); Franklin (2017); Abebe et al (2021)
Brülhart et al. (2025)	Who Bears the Burden of Local Taxes?	Incidence of local taxes on high-income population	3	Suárez-Serrato, Zidar (2016); Fuest, Peichl, Siegloch (2018); Löffler, Siegloch (2021)
Bukovina et al. (2025)	Corporate Minimum Tax and the Elasticity of Taxable Income: Evidence from Administrative Tax Records	Elasticity of taxable income	3	Pomeranz (2015); Almunia, Lopez-Rodriguez (2018); Naritomi (2019)

Continued on next page

Authors and year	Paper title	Quantity of interest	No. of other papers	Other estimates of quantity of interest among already-cited papers (first four shown)
Cabral et al. (2025)	The Impact of Provider Payments on Health Care Utilization of Low-Income Individuals: Evidence from Medicare and Medicaid	Demand elasticity for Medicare	5	Roberts, Desai (2021); Brown, Kowalski, Lurie (2019); Fung et al (2021); Saez (2010); ...
Castell et al. (2025)	Take-up of Social Benefits: Experimental Evidence from France	Effects of various treatments on program take-up	11	Duflo, Saez (2003); Kling et al (2012); Bhargava, Manoli (2015); Bettinger et al (2012); ...
Daysal et al. (2025)	Do Medical Treatments Work for Work? Evidence from Breast Cancer Patients	Mortality effects of radiotherapy	4	Overgaard et al (1997); Ragaz et al (1997); Clarke et al (2005); Darby et al (2011)
Daysal et al. (2025)	Do Medical Treatments Work for Work? Evidence from Breast Cancer Patients	Labor market outcomes of breast cancer survivors	3	Bradley, Bednarek, Neumark (2002); Bradley et al (2005); Heinesen, Kolodziejczyk (2013)
Daysal et al. (2025)	Do Medical Treatments Work for Work? Evidence from Breast Cancer Patients	% of breast cancer survivors that return to work	6	Drolet et al (2005); Balak et al (2008); Johnsson et al (2009); Damkjær et al (2011); ...
Dupas et al. (2025)	Informing Mothers about the Benefits of Conversing with Infants: Experimental Evidence from Ghana	Effect of program on # of mothers conversing with infants	102*	Meta-analysis by Jeong et al. (2021)
Edwards et al. (2025)	HBCU Enrollment and Longer-Term Outcomes	Effects of attending an HBCU on degree completion	6	Fryer and Greenstone (2010); Price, Spriggs, Swinton (2011); Elu et al (2019); Ehrenberg, Rothstein (1999); ...
Getik and Meier (2025)	The Long-Run Effects of Peer Gender on Occupational Sorting and the Wage Gap	Effect of occupational sorting on the female wage	3	Black, Devereux, Salvanes (2010); Annelli, Peri (2019); Brenøe, Zölitz (2020)
Guttman-Kenney et al. (2025)	The Semblance of Success in Nudging Consumers to Pay Down Credit Card Debt	Effects of various nudges on credit card debt levels	4	Agarwal et al (2015); Seira, Elizondo, Laguna-Müggenberg (2017); Adams et al (2022); Medina (2021)
Harris and Mills (2025)	Should College Be "Free"? Evidence on Free College, Early Commitment, and Merit Aid from an Eight-Year Randomized Trial	Effect of financial aid programs on college attainment & college outcomes conditional on attainment	5	Angrist, Autor, Pallais (2021); Barrow et al (2014); Goldrick-Rab et al (2016); Carlson et al (2019); ...
Ito and Zhang (2025)	Do Consumers Distinguish Fixed Cost from Variable Cost? "Schmeduling" in Two-Part Tariffs in Energy	Percentage decrease in energy use	6	Wolak (2011); Ito (2014); Ito (2015); Ito, Ida, Tanaka (2018); ...

Continued on next page

Authors and year	Paper title	Quantity of interest	No. of other papers	Other estimates of quantity of interest among already-cited papers (first four shown)
Kessler and Roth (2025)	Increasing Organ Donor Registration as a Means to Increase Transplantation: An Experiment with Actual Organ Donor Registrations	Effect of "yes-no" framing vs "opt-in" framing on organ donor registrations	2	Johnson, Goldstein (2004); van Dalen, Henkens (2014)
Li et al. (2025)	The Effects of a Multifaceted Poverty Alleviation Program on Rural Income and Household Behavior in China	% increase in rural/poor incomes X years after program implementation	2	Meng (2013); Park, Wang (2010)
Lichter et al. (2025)	Profit Taxation, R&D Spending, and Innovation	Elasticity of R&D spending	2	Agrawal, Rosell, Simcoe (2020); Dechezleprêtre et al (2023)
Mehmood et al. (2025)	Transmitting Rights: Effective Cooperation, Inter-gender Contact, and Student Achievement	Effects of curricular reforms on teacher/student attitudes	2	Alan et al (2021); Dhar et al (2022)
Moscona (2025)	The Management of Aid and Conflict in Africa	Effects of foreign aid on rate of violent conflict	7	Crost, Felter, Johnston (2014); Nunn, Qian (2014); Dube, Naidu (2015); Collier, Hoeffler (2010); ...
Sabet et al. (2025)	Terrorism and Voting: The Rise of Right-Wing Populism in Germany	Effect of terror shocks on right-wing party vote shares	9	Gould, Klor (2010); Berrebi, Klor (2011); Getmansky, Zeitzoff (2014); Rehman, Vanin (2017); ...
Sager and Singer (2025)	Clean Identification? The Effects of the Clean Air Act on Air Pollution, Exposure Disparities, and House Prices	Elasticity of house prices with respect to pollution levels	3	Chay and Greenstone (2005); Bento et al. (2015); Grainger (2012)
Shastry and Tortorice (2025)	Effective Health Aid: Evidence from Gavi's Vaccine Program	Effects of Gavi vaccine program on child mortality	5	Lu et al (2006); Jaupart, Dipple, Dercon (2019); Dykstra et al (2019); Ikilezi et al (2020); ...
Siegloch et al. (2025)	Spillover, Efficiency, and Equity Effects of Regional Firm Subsidies	Employment effects of GRW	7	Brachert, Dettmann, Titze (2019); Kline, Moretti (2014); Alder, Shao, Zilibotti (2016); Criscuolo et al (2019); ...
Sigurdsson (2025)	Labor Supply Responses and Adjustment Frictions: A Tax-Free Year in Iceland	Frisch elasticity	5	Martínez, Saez, Siegenthaler (2021); Fehr, Goette (2007); Farber (2014); Chetty, Friedman, Saez (2013); ...

Notes: The "Other estimates of quantity of interest among already-cited papers" column lists the first four cited studies when a paper cites more than four (the remainder are indicated by "..."). An asterisk (*) in the "No. of other papers" column denotes that the count is the number of studies in a cited meta-analysis rather than individually listed papers.

C Binary implementation decisions

Binary decisions A_i might correspond to hypothesis tests, to decisions about implementing different treatments, or decisions about the same treatment for different sub-populations, etc. For binary decisions, we take θ_i to be the return to a positive decisions ($A_i = 1$) in problem-instance i.

Suppose that each of the studies i corresponds to an (experimental or quasi-experimental) evaluation of a different treatment. Suppose further that for each of these treatments, we wish to make a decision $A_i \in \{0, 1\}$ whether to implement them on a larger scale, with social return θ_i, so that $U(A_i, \theta_i) = A_i \cdot \theta_i$, $A_i \in \{0, 1\}$. For this interpretation, θ_i should be defined such that it captures all relevant costs and benefits of the treatment.

Ideally, if we could observe θ_i, we would choose $A_i = \mathbf{1}(\theta_i \geq 0)$. In practice, we only observe $(\hat{\theta}_i, \sigma_i, X_i)$. Denote the posterior expected return to A_i by $t(\hat{\theta}, \sigma) = E\left[\theta|\hat{\theta}, \sigma\right]$. We derived expressions for this posterior expected return, under various assumptions, in Section 3.1 above. The optimal decision for a Bayesian decision-maker who maximizes expected welfare equals

$$A_i = \mathbf{1}(t(\hat{\theta}_i, \sigma_i, X_i) \geq 0),$$

To estimate the posterior expected return, we recall Tweedie's formula from Equation (4), $t(\hat{\theta}, \sigma, x) = \hat{\theta} + \sigma^2 \cdot \frac{\partial}{\partial \hat{\theta}} \log(f(\hat{\theta}|\sigma, x))$, where $f(\hat{\theta}|\sigma, x)$ is the conditional density. Under the parametric model where $\theta_i|X_i = x \sim N(x \cdot \beta, \tau^2)$, we can specialize this to $E[\theta_i|\hat{\theta}_i, \sigma_i, X_i] = (1 - \kappa_i) \cdot (X_i \cdot \beta) + \kappa_i \cdot \hat{\theta}_i$, where $\kappa_i = \frac{\tau^2}{\tau^2 + \sigma_i^2}$. Note that the sign of $t(\hat{\theta}_i, \sigma_i, X_i)$ might differ from the sign of $\hat{\theta}_i$, depending on the distribution of $\hat{\theta}_i$ across i, and depending on the value of the covariates X_i.

A closely related decision problem concerns the implementation of a binary decision A_0 for a new problem instance, for which no $\hat{\theta}_0$ or σ_0 has been observed, but covariates X_0 are available. In this setting, the optimal decision is given by

$$A_0 = \mathbf{1}(E[\theta_0|X_0] \geq 0).$$

Under the parametric linear model, $E[\theta_0|X_0] = X_0 \cdot \beta$; we discussed estimation of β using (appropriately weighted) Ridge regression in Section 4.1. The posterior expected welfare for the optimal treatment decision in instance i, when an estimate $\hat{\theta}_i$ is available, equals

$$E[U(A_i, \theta_i)|\hat{\theta}_i, \sigma_i, X_i] = \max(t(\hat{\theta}_i, \sigma_i, X_i), 0).$$

This expression for posterior expected welfare will be important below, when we

consider the allocation of limited research resources, where the implied goal is to maximize the (ex-ante) *value of information* of observing $\hat{\theta}_i$, which can be written as

$$E\left[\max\left(t(\hat{\theta}_i, \sigma_i, X_i), 0\right)\big|X_i\right] - \max\left(E\left[t(\hat{\theta}_i, \sigma_i, X_i)\big|X_i\right], 0\right). \tag{21}$$

D Covariates

Table 9 reports the per-study estimates, standard errors, distances to each test point, citations, and sample sizes for every study that appears in Figure 4 of Section 4.2.2 — the union of the ten highest-$|w|$ studies for each of the two test contexts.

Table 9: Posterior distribution of employment effects of training: Panel B

		Study estimate		UI recipients		Long-term unemployed	
	n	$\hat{\theta}_i$	SE	k(short ℓ)	k(long ℓ)	k(short ℓ)	k(long ℓ)
1. UI/Tr/L	4,664	+0.12	0.009	0.0056	0.0056	0.0017	0.0052
2. UI/Tr/L	7,934	+0.15	0.009	0.0056	0.0056	0.0017	0.0052
3. UI/Tr/L	95,000	+0.04	0.01	0.0056	0.0056	0.0017	0.0052
4. UI/Tr/L	92,500	+0.05	0.01	0.0056	0.0056	0.0017	0.0052
5. UI/Tr/L	85,400	+0.25	0.01	0.0056	0.0056	0.0017	0.0052
6. UI/Tr/L	86,000	+0.27	0.01	0.0056	0.0056	0.0017	0.0052
7. UI/Tr/L	88,400	+0.24	0.01	0.0056	0.0056	0.0017	0.0052
8. UI/Tr/S	28,246	-0.019	0.005	0.0018	0.0052	0.0006	0.0048
9. LTU/Tr/S	23,182	+0.023	0.006	0.0006	0.0048	0.0018	0.0052
10. LTU/Tr/S	15,532	+0.021	0.007	0.0006	0.0048	0.0018	0.0052
11. LTU/Tr/M	23,182	+0.056	0.006	0.0003	0.0046	0.0008	0.0049
12. LTU/Tr/M	15,532	+0.018	0.008	0.0003	0.0046	0.0008	0.0049
13. LTU/Ot/L	2,268	-0.037	0.017	0.0001	0.0043	0.0003	0.0046
14. LTU/Ot/S	3,705	+0.041	0.001	0.0000	0.0040	0.0001	0.0043
15. LTU/Ot/M	3,705	+0.05	0.0014	0.0000	0.0038	0.0000	0.0041
16. Dis/Tr/S	83,145	-0.000295	0.0016	0.0007	0.0049	0.0002	0.0046

Notes: Per-study estimates and per-context kernel covariances for the union of studies displayed in either panel of the GP-weights figure. $\hat{\theta}_i$ and SE are the study estimate and standard error; n is the study sample size. $k(x_0, x_i) = \rho^2 \exp(-d^2/2\ell^2)$ is the squared-exponential kernel covariance between study i and the test point x_0, where d is their Euclidean distance in the standardised covariate space, evaluated at the short and long length scales ℓ ($\rho^2 = 0.0056$, $\ell_{\text{short}} = 1.58$, $\ell_{\text{long}} = 6.32$). A larger covariance means study i is more similar to the test point and receives more posterior weight. $\hat{\theta}_i$ and SE are shown to the number of decimals reported in the original study; n to 0 and k to 4 decimals.

E Smoothness and monotonicity of the p-curve

What properties should the p-curve exhibit in the absence of selectivity? For the following derivation, assume that p-values are based on Z-statistics Z_i for one-sided hypothesis tests of the null that $\theta_i = 0$, $P_i = \Phi(-Z_i)$. Let $z = \Phi^{-1}(1-p)$, so that $\frac{\partial z}{\partial p} = -\frac{1}{\varphi(z)}$. In Equation 19 in the preceding section, we showed that the density of Z is given by a convolution of the distribution ν of ω with the standard normal density φ, $f(z) = \int \varphi(z-\omega)d\nu(\omega)$. From this, we can derive the corresponding distribution of p-values as follows, taking into account the change of variables from Z to P. Here $g(p)$ is the density of p-values, or p-curve:

$$
\begin{aligned}
P(P_i \leq p) &= 1 - F(z) \\
g(p) &= \frac{\partial P(P \leq p)}{\partial p} = \frac{f(z)}{\varphi(z)}, \\
g'(p) &= \frac{-1}{\varphi(z)^3} \cdot (f'(z)\varphi(z) - f(z)\varphi'(z)) \\
&= \frac{-1}{\varphi(z)^3} \cdot \left(\int (\varphi'(z-\omega)\varphi(z) - \varphi(z-\omega)\varphi'(z))\, d\nu(\omega) \right)
\end{aligned}
$$

Smoothness of the density $g(p)$ is immediate from this derivation. In fact, g is analytic (infinitely differentiable) since φ is analytic, f is a convolution of ν with φ, and analyticity is preserved under convolution. The p-curve should thus have no jumps or kinks.

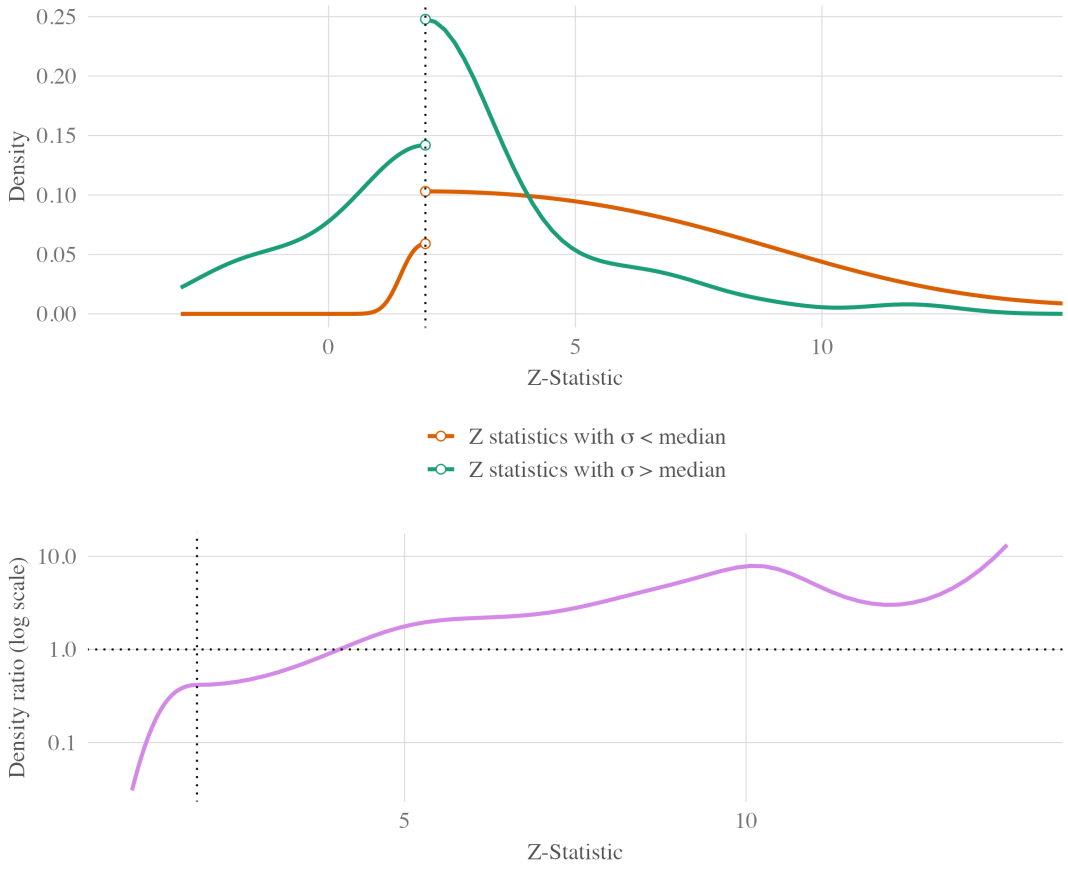
In addition to smoothness, this derivation also implies restrictions on the shape of $g(\cdot)$. Since the standard normal density is given by $\varphi(z) = (2\pi)^{-1/2} \cdot \exp\left(-\frac{z^2}{2}\right)$, it follows that $\frac{\varphi'(z)}{\varphi(z)} = \frac{\partial \log(\varphi(z))}{\partial z} = -z$. Consider the integrand $\varphi'(z-\omega)\varphi(z) - \varphi(z-\omega)\varphi'(z)$ in the expression for $g'(p)$ above. This integrand is positive iff

$$
\omega - z = \frac{\varphi'(z-\omega)}{\varphi(z-\omega)} \geq \frac{\varphi'(z)}{\varphi(z)} = -z.
$$

The integrand is therefore positive for $\omega \geq 0$. Therefore $g'(z) \leq 0$ if ν is supported on $\mathbb{R}^+$, and monotonicity of g follows. A similar calculation yields monotonicity of g for 2-sided hypothesis tests, where $P_i = 2 \cdot \Phi(-|Z_i|)$.

F Density ratio in Cohen and Ganong (2026)

Figure 12: Cohen & Ganong data: Z-statistic densities by SE group and density ratio



To show how reduced-form evidence of a nonzero latent mean can appear in a real-life empirical setting with selectivity, we apply the plotting framework from the illustrative example to a real application: Cohen and Ganong (2026) in Figure 12. We divide the sample into two groups based on the median standard error. The top panel of the figure shows the density of Z-statistics. The panel shows many of the same features as the simulated environment in Figure 8: the Z-statistics for the more precise estimates in red are much more dispersed than the imprecise estimates in green and there is a jump in the densities at $Z = 1.96$. The key qualitative pattern to focus on is that the precise estimates in red are not simply a mean-preserving spread of the green estimates (as is the case in the left panel of Figure 8). Instead, the density of the red estimates is shifted to the right relative

to the density of the green estimates (as is the case in the right panel of Figure 8). This is the "signature" of a positive value for the latent mean of the distribution.

The bottom panel of Figure 12 makes this "signature" legible. The density ratio (of the imprecise estimates to the precise estimates) is less than one everywhere below 1.96 and is rising monotonically in the Z-statistic through $Z = 10$. This pattern is quite similar to the right panel of Figure 9 and is noticeably different from the U-shape in the density ratio in the left panel of Figure 9. We emphasize that one should not use this procedure of splitting the sample in half to actually estimate a latent mean. Maximum likelihood, which uses all available information in σ_i, is optimal. Nevertheless, the patterns here are useful as a visual diagnostic.

G Density Estimation and Bandwidth Selection

In the main text, Section 5.2, we test for selectivity by examining whether the p-curve is discontinuous at 0.05 (Elliott et al., 2022, 2025). Our method of local polynomial density estimation requires choosing a bandwidth (Cattaneo et al., 2020). With a small set of studies, bandwidth choice becomes challenging and could meaningfully affect whether one rejects the null of no discontinuity in the p-curve at 0.05. To assess whether our results are robust to different bandwidth choices, Table 10 shows the p-values for three bandwidth choices: MSE optimal, 1/2 MSE optional, and 2x MSE optimal.

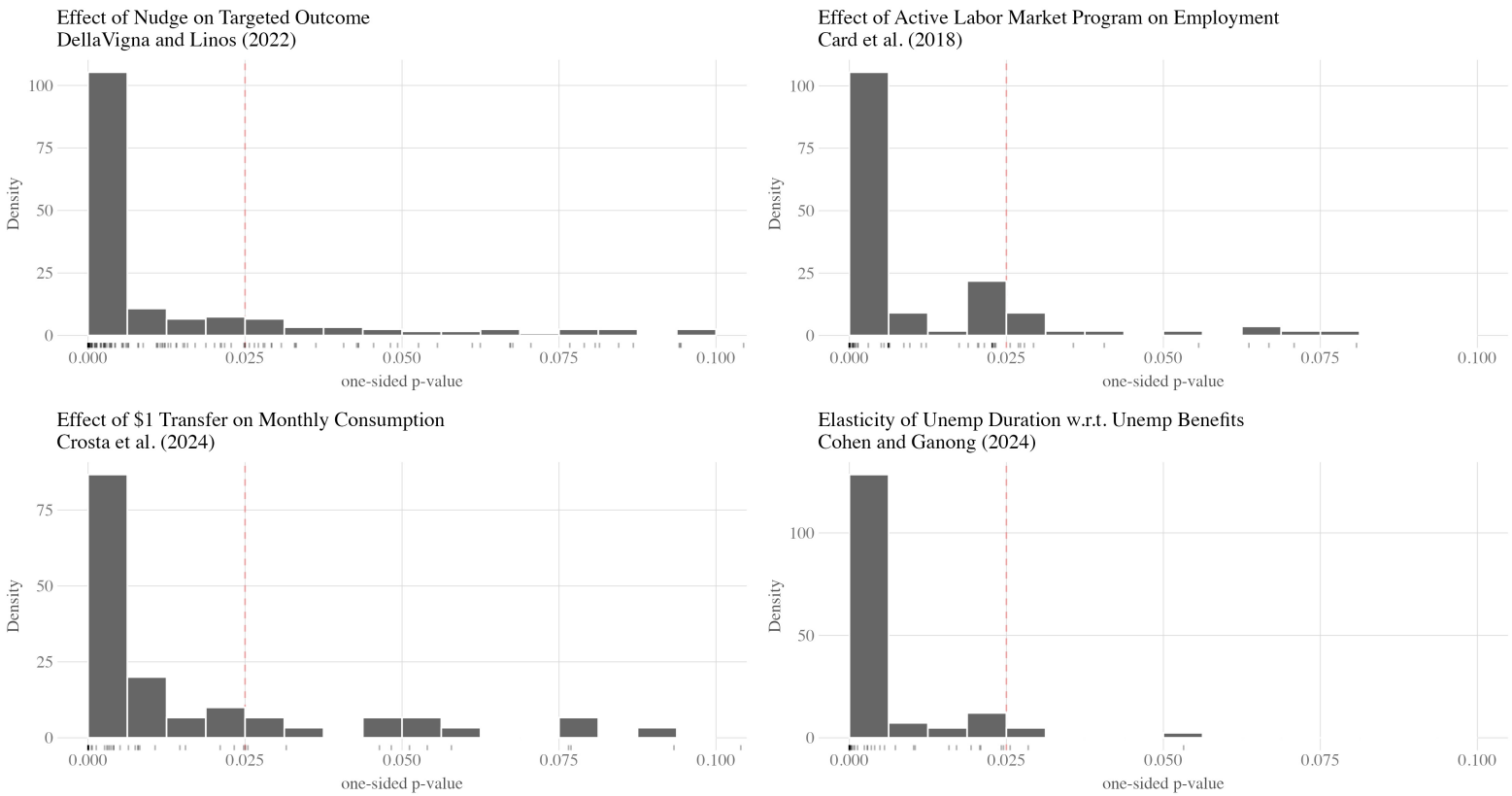
Table 10: Bandwidth Sensitivity in Discontinuity Tests.

		Bandwidths			
Study		$h_{p<0.025}$	$h_{p\geq 0.025}$	Density Ratio	p-value
Effect of Nudge on Targeted Outcome DellaVigna and Linos (2022)	0.5x MSE-optimal	0.012	0.098	0.191	0.287
	MSE-optimal	0.025	0.196	0.634	0.221
	2.0x MSE-optimal	0.025	0.392	0.642	0.208
Effect of Active Labor Market Program on Employment Card et al. (2018)	0.5x MSE-optimal	0.011	0.126	0.039	0.075
	MSE-optimal	0.022	0.252	0.068	0.000
	2.0x MSE-optimal	0.025	0.505	0.062	0.001
Effect of $1 Transfer on Monthly Consumption Crosta et al. (2024)	0.5x MSE-optimal	0.012	0.151	0.269	0.198
	MSE-optimal	0.025	0.301	-0.814	0.095
	2.0x MSE-optimal	0.025	0.603	-0.539	0.091
Elasticity of Unemp Duration w.r.t. Unemp Benefits Cohen and Ganong (2026)	0.5x MSE-optimal	0.012	0.263	0.068	0.374
	MSE-optimal	0.025	0.526	0.077	0.021
	2.0x MSE-optimal	0.025	0.975	0.070	0.015

Notes: This table presents p-values from tests of selectivity with different choices of bandwidths and with de-rounded data. It reports the p-value using MSE optimal bandwidths, 1/2 MSE optimal bandwidth, and 2x MSE optimal bandwidth. If the bandwidth below the cutoff is greater than 0.025, we use 0.025. If the bandwidth above the cutoff is greater than 0.975, we use 0.975.

In general, the results are not robust to different bandwidth choices. However, we consistently estimate a density ratio less than 1 but find insufficient evidence to reject the null hypothesis of no selection.

Figure 13: Density Discontinuity in the P-curve at 0.05.



Notes: This figure shows the histogram of p-values for estimates of studies in four meta-analyses. The y-axis is the number of estimates. Plot annotations are MSE-optimal bandwidth h on each side of the cutoff, the estimated density (number of studies), the difference of the densities, and the p-value of the difference of these densities. For DellaVigna and Linos (2022), we include only estimates from the academic journal subsample since we assume nudge unit estimates are always reported.

Appendix References

Bailey, M., Byker, T., Patel, E., and Ramnath, S. (2025). The long-run effects of california's paid family leave act on women's careers and childbearing: New evidence from a regression discontinuity design and US tax data. *American Economic Journal: Economic Policy*, 17(1):401–431.

Baker, M., Halberstam, Y., Kroft, K., Mas, A., and Messacar, D. (2023). Pay transparency and the gender gap. *American Economic Journal: Applied Economics*, 15(2):157–183.

Baum, Charles L., I. and Ruhm, C. J. (2016). The effects of paid family leave in california on labor market outcomes. *Journal of Policy Analysis and Management*, 35(2):333–356.

Bennedsen, M., Simintzi, E., Tsoutsoura, M., and Wolfenzon, D. (2022). Do firms respond to gender pay gap transparency? *The Journal of Finance*, 77(4):2051–2091.

Bento, A., Freedman, M., and Lang, C. (2015). Who benefits from environmental regulation? evidence from the clean air act amendments. *The Review of Economics and Statistics*, 97(3):610–622.

Blundell, J., Duchini, E., Simion, Ş., and Turrell, A. (2025). Pay transparency and gender equality. *American Economic Journal: Economic Policy*, 17(2):418–445.

Campbell, Z., Chin, I., Chyn, E., and Hastings, J. (2017). The impact of paid leave: Evidence from temporary disability insurance in Rhode Island. Technical report. Working paper, preliminary draft dated July 27, 2017.

Cattaneo, M. D., Jansson, M., and Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531):1449–1455.

Chay, K. Y. and Greenstone, M. (2005). Does air quality matter? evidence from the housing market. *Journal of Political Economy*, 113(2):376–424.

Cohen, J. P. and Ganong, P. (2026). Disemployment effects of unemployment insurance: A meta-analysis. *American Economic Review: Insights*, 8(1):1–18.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614.

Elliott, G., Kudrin, N., and Wüthrich, K. (2022). Detecting p-hacking. *Econometrica*, 90(2):887–906.

Elliott, G., Kudrin, N., and Wüthrich, K. (2025). The power of tests for detecting p-hacking. *The Review of Economics and Statistics*, pages 1–44.

Grainger, C. A. (2012). The distributional effects of pollution regulations: Do renters fully pay for cleaner air? *Journal of Public Economics*, 96(9-10):840–852.

Gulyas, A., Seitz, S., and Sinha, S. (2023). Does pay transparency affect the gender wage gap? evidence from austria. *American Economic Journal: Economic Policy*, 15(2):236–255.

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109(506):674–685.

Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. Springer.

Rossin-Slater, M., Ruhm, C. J., and Waldfogel, J. (2013). The effects of california's paid family leave program on mothers' leave-taking and subsequent labor market outcomes. *Journal of Policy Analysis and Management*, 32(2):224–245.

Sager, L. and Singer, G. (2025). Clean identification? the effects of the clean air act on air pollution, exposure disparities, and house prices. *American Economic Journal: Economic Policy*, 17(1):1–36.

Timpe, B. (2024). The labor market impacts of america's first paid maternity leave policy. *Journal of Public Economics*, 231:105067.