Uncertainty in Empirical Economics

Frank Schorfheide*
University of Pennsylvania,
CEPR, PIER, NBER

Zhiheng You
University of Pennsylvania

May 29, 2025

Abstract

Econometricians invest substantial effort in constructing standard errors that yield valid inference under a hypothetical data-generating process. This paper asks a fundamental question: Are the uncertainty statements reported by applied researchers consistent with empirical frequencies? The short answer is no. Drawing on the forecasting literature, we predict estimates from "new" studies using estimates from corresponding baseline studies. By doing this across a large number of study groups and linking parameters through a hierarchical model, we compare stated probabilities to observed empirical frequencies. Alignment occurs only under limited external validity, namely, that the studies estimate different parameters. (JEL C11, C18, C21)

Key words: Bayesian Inference, External Validity, Hierarchical Models, Meta Studies, Standard Errors, Statistical Inference, Treatment Effects, Uncertainty

* Correspondence: F. Schorfheide and Z. You: Department of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297. Email: schorf@ssc.upenn.edu (Schorfheide) and zhyou@sas.upenn.edu (You).
We thank Tim Armstrong, Magne Mogstad (Guest Editor), Alexander Torgovitsky (Guest Editor), seminar participants at the University of Pennsylvania, the University of Oxford, the JPE: Micro Causal Inference Conference, and students from the University of Chicago's Autumn 2024 Econ 21160 "Topics on Causal Inference" class. Schorfheide gratefully acknowledges financial support from the University of Chicago Griffin Economics Incubator Distinguished Visitor Program.

1 Introduction

Empirical results in economics are typically reported as point estimates accompanied by standard errors, allowing readers to construct confidence intervals. This paper asks whether such reporting practices provide meaningful and accurate quantifications of uncertainty. The standard errors offered by the econometrics literature are justified within the internal logic of a probabilistic model. In the frequentist framework, the econometrician posits a data-generating process (DGP), outlines procedures for constructing confidence intervals, and evaluates their properties through asymptotic theory and Monte Carlo simulations. However, empirical analysis occurs outside this controlled environment, raising the question of whether empirical frequencies align with probability statements based on a hypothetical DGP. A central challenge in answering this question is the unobservability of "true" parameters.

Estimation differs fundamentally from forecasting in terms of the ex post verifiability of probability statements. In forecasting, the "true" outcomes, such as future GDP growth, are eventually observed, enabling rigorous evaluation of probabilistic forecasts. For example, one can test whether interval forecasts achieve their stated coverage probabilities. In contrast, the "true" parameter values underlying estimates are unobservable. To apply insights from the forecasting literature to the estimation context, we reframe the estimation problem as a prediction problem. Specifically, we consider the task of predicting a "new" estimate of the same or a closely related parameter using a different sample. This is a common scenario in economics and other social sciences, where researchers frequently re-estimate parameters such as labor supply elasticities, returns to schooling, the slope of the New Keynesian Phillips curve (NKPC), or the effects of monetary and fiscal policy shocks.

To implement the assessment, we form groups of studies that estimate similar parameters or causal effects. We designate one of the studies in each group as validation study and the remaining ones as baseline studies. The baseline studies are used to construct interval or density predictions for the estimate reported the validation study. Averaging across groups, we compare empirical frequencies to nominal probability statements and apply techniques from the interval and density forecast evaluation literature. A practical challenge for this evaluation is that grouped studies may not estimate the same parameter. For instance, treatment effect parameters in microeconometric applications may be specific to the (sub)populations for which they are estimated. Thus, the assessment of probability statements is interwoven with beliefs or evidence about external validity. To allow for parameter variation across grouped studies, we borrow from the literature on meta-analysis and create a hierarchical

Bayes model in which the study-specific parameters are drawn from a distribution that has a common mean and potentially nonzero variance for each group. This variance, denoted by ν, is a hyperparameter of the hierarchical model.

If the reported standard errors are too large and coverage intervals are too wide on average, then we should find that even under the assumption that parameters are equal across studies, average coverage frequency exceeds the nominal coverage probability. In our applications, the opposite is the case. The standard errors reported in the baseline studies are too small to generate predictive intervals that reach the promised nominal coverage probability. Thus, to align the empirical coverage frequency with the coverage probability, one has to acknowledge that external validity is limited and allow for a nonzero within-group differences of parameter values through an appropriate value of ν. Returning to the question posed at the beginning of the introduction, if ν is large relative to the reported standard errors, then the uncertainty quantification provided by the standard errors is of limited usefulness. A reader should mentally scale up the standard errors if (s)he wanted to use the reported estimates for inductive inference.

We focus on microeconometric studies that allow us to assume that the underlying data sets and the resulting estimators are independently distributed.[1] Because the studies considered in our applications report frequentist point estimates and standard errors, we interpret the associated Gaussian limit distributions as likelihood functions that can be embedded in a hierarchical Bayes model. We use a parametric Bayes model because the sample sizes (number of empirical studies) in our applications are relatively small. As mentioned previously, the key parameter of the hierarchical model is the hyperparameter ν that controls the dispersion of population parameters across studies. We either set $\nu = 0$, plot coverage frequencies as a function of ν, or estimate ν from the baseline studies.

We consider three empirical applications. The first application is based on estimates of various psychological effects collected in a study by Klein, Ratcliff, 48 others, and Nosek (2014). We provide a detailed analysis with various robustness exercises that include perturbing the assignment of studies to groups, constructing predictive intervals that rely less heavily on normality assumptions, and allowing ν to be heterogeneous across groups. The second application uses estimates of the impact of microcredit expansions from randomized controlled trials (RCTs) collected by Meager (2019). Here the number of studies is relatively

[1] In empirical macroeconomics, researchers often use the same set of aggregate variables to estimate different model specifications based on overlapping samples. Common observations create correlations among estimates that complicate the analysis.

small, which limits the power of the analysis. The third application examines estimates collected by DellaVigna and Linos (2022) of the effects of nudges on the adoption of certain types of behaviors.

In all three applications the coverage frequency is smaller than the targeted coverage probability if one believes that studies in each group estimate the same population parameter. If one is willing to entertain the possibility that studies within each group estimate somewhat different population parameters, then it is possible to bring empirical coverage frequencies and nominal coverage probabilities into alignment. But for many groups the estimated within-group variation of parameters substantially exceeds the reported standard errors of the estimates, indicating that the standard errors may not capture the most important dimension of our lack of knowledge.

Our paper is connected to several strands of literature. The assessment of the reported standard errors and coverage intervals builds on the time series literature on evaluating interval and density forecasts. Two seminal papers in this literature are Christoffersen (1998) and Diebold, Gunther, and Tay (1998). Related to density forecasting and evaluation, the premise of Dawid (1984)'s prequential approach to statistical inference is that the analysis aims to make probability forecasts for future observations rather than to express information about parameters. This is closely connected to our notion of predicting the estimate in a "new" study based on existing studies. The basic idea of the prequential approach is to assess a prequential forecasting system in light of observed outcomes. This assessment is based on a comparison of empirical frequencies with predicted probabilities. Because the studies that we consider are assumed to be uncorrelated and we are not exploiting any temporal ordering, our evaluation resembles the panel forecasting setting in Liu, Moon, and Schorfheide (2023).

Meta-analysis synthesizes findings from different empirical studies. As in the meta-analysis literature, we link parameters from different studies using a hierarchical Bayes model. Sutton and Abrams (2001) provide a comprehensive review of Bayesian approaches to meta-analysis in the context of medical applications. In economics, meta-studies are less common than in other disciplines. Some examples include the papers by Rusnák, Havranek, and Horváth (2013), Meager (2019), DellaVigna and Linos (2022), Meager (2022), Ehrenbergerova, Bajzik, and Havranek (2023), Gechter and Meager (2022), and Aimone, Ball, Dwibedi, Jackson, and West (2024).

The literature on external validity examines the extent to which parameters obtained from historical studies are relevant for a parameter associated with a "new" population.

This question is related to, but different from our analysis. The baseline studies in our framework correspond to the historical studies in the external validity literature, and our validation studies are similar to the estimation based on a "new" population. One specific task in the external validity literature is to use the historical studies to form a prior for the parameters of the "new" population in settings where the "new" study contains relatively little sample information; see Spiegelhalter (2004), Schmidli, Gsteiger, Roychoudhury, O'Hagan, Spiegelhalter, and Neuenschwander (2014), and the economic application in Iacovone, McKenzie, and Meager (2023). We convert this prior into a predictive distribution for the "new" estimate and examine whether the probability statements align with empirical frequencies.

Spiegelhalter (2004) provides a taxonomy for the relationship between the parameters underlying the historical studies and the "new" study: the "new" population parameter is identical to the parameters in the historical studies; the parameters are equal, but information from the historical studies should be discounted; there is a known functional dependence between the parameters of the previous studies and the "new"; the parameters could be drawn from the same distribution and are exchangeable; there are biases in the historical estimates due to the use of observational studies, or caused by strategic site selection for RCTs; previous studies could be irrelevant because their parameters are unrelated to the parameter associated with the "new" study. Our hierarchical Bayes model assumes that the parameters are different, yet drawn from the same distribution, thereby capturing a subset of the possible relationships between historical and "new" parameters in a reduced form.

Finally, there is a literature on publication bias, e.g., Andrews and Kasy (2019). Journals may be more likely to publish studies that report estimates of parameters that, under conventional hypothesis testing, appear to be significantly different from zero. The presence of publication bias alters the sampling distribution of published estimates because, in the extreme case, an estimate $\hat{\theta}$ and the associated coverage interval would only become observable to us if $|\hat{\theta}/\hat{\sigma}_{\hat{\theta}}| > 1.96$. Hence, the publication process truncates the distribution of $\hat{\theta}|\theta$. In our analysis, such a truncation would be another source of mismatch between stated coverage probabilities and empirical frequencies.

The remainder of the paper is organized as follows: Section 2 provides a motivating example, illustrating that estimates of the same object may vary drastically across studies. Section 3 develops a hierarchical model that allows us to connect parameters across different studies within groups of studies, and shows how to turn estimates from the baseline studies into an interval or density prediction for the estimate of a validation study. The empirical

applications are presented in Sections 4 to 6. Finally, we conclude in Section 7. Derivations and further details on the empirical illustrations are relegated to an Online Appendix that is available from the authors' websites.

2 A Motivating Example: Variation of Estimates

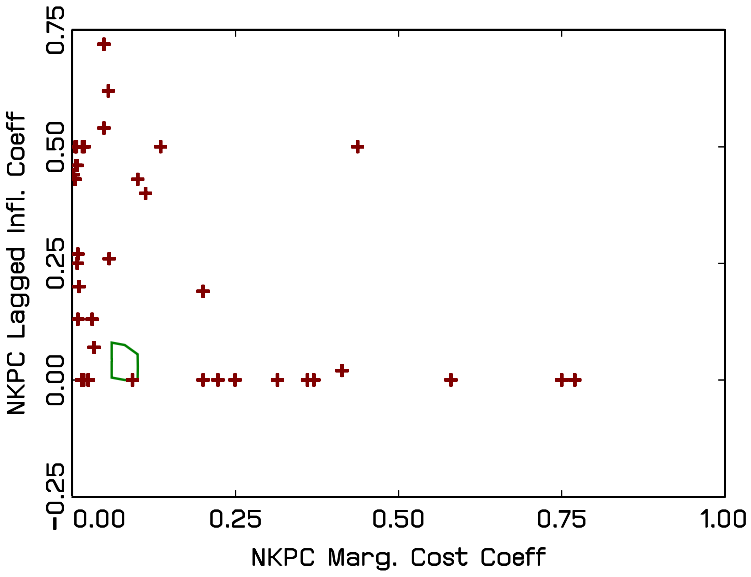
We take the perspective of a reader of empirical studies who trusts that the econometric analysis has been carried out in a competent manner. This reader is typically confronted with tables of point estimates and standard errors that describe the uncertainty associated with these estimates. If the econometric analysis has been carefully executed, then the uncertainty statements should be consistent with the assumed generative model. In applied settings the generative model is an abstraction, and we do not want to rely on it for the evaluation of uncertainty quantifications. Instead, building on the literature on forecast evaluation, we split a collection of empirical papers into paired baseline and validation studies, and examine whether one can use the uncertainty measures reported in the baseline studies to generate accurate probabilistic predictions of the validation study estimates.

Our strategy is partly motivated by Figure 3 of Schorfheide (2013), reproduced here as Figure 1. It shows a scatter plot of point estimates of two NKPC parameters: the coefficient on marginal costs κ and the coefficient on lagged inflation γ_b. The estimates were collected by Schorfheide (2008) from previously published papers reporting on estimated dynamic stochastic general equilibrium (DSGE) models. In addition, the figure also contains a 90% baseline credible set for the two NKPC parameters obtained from a DSGE model estimation by Schorfheide (2013). The troubling aspect of the figure is that the size of the baseline credible set is a magnitude smaller than the dispersion of the point estimates obtained from other studies. In this regard, the baseline coverage set does not seem to convey the information about uncertainty that we should convey in empirical studies.[2]

There could be many reasons for the large dispersion of the empirical estimates, including: (i) the estimates represented by the crosses are very noisy. If one would draw coverage sets around these estimates, these sets might intersect with the baseline credible set. (ii) The studies might estimate different parameters, because parameters could vary over time and

[2]Menkveld, Dreber, 340 others, and Zwinkels (2024) study a related question. They provide identical data sets to teams of researchers, ask them to conduct certain tests or measure particular effects, and examine the variation in the results, which are in part due to data cleaning or auxiliary modeling assumptions.

Figure 1: A Scatter Plot of DSGE Model Based NKPC Estimates



Notes: Source: Schorfheide (2013). "+" indicate DSGE model based point estimates of NKPC parameters obtained from studies surveyed in Schorfheide (2008). The set outlined by green solid line is a 90% credible set based on the model estimated in Schorfheide (2013).

across countries. (iii) Parameters can be model-specific objects, and estimates could differ because model specifications differ across studies. (iv) Related, auxiliary assumptions to identify the parameters of interest could differ across studies.

In the remainder of this paper, we pair baseline and validation studies and examine whether the probabilistic predictions derived from the baseline studies are consistent with the empirical frequencies of the validation study estimates. In case the answer is no, we rephrase the question as follows: on average, how different have the underlying parameters of the baseline and validation studies to be, such that average coverage frequencies match the stated coverage probabilities? In the context of Figure 1, we would use the reported measures of uncertainty for each of the crosses to address point (i). We would then group the studies such that within-group estimands are sufficiently similar to alleviate point (ii). Because addressing points (iii) and (iv) is beyond the scope of this paper, the empirical applications in Sections 4 to 6 use RCT studies that do not rely on sophisticated modeling or identification assumptions.

3 A Framework to Assess Uncertainty Statements

In economics and other social sciences, researchers often re-estimate parameters or causal effects that have been estimated in earlier studies. This will allow us to create pairs of baseline and validation studies that share similar if not equal estimands. We now develop a hierarchical model that links parameters across studies, building on the literatures on meta-analysis and external validity, and allows us to compute predictive intervals for estimates reported in validation studies. The basic framework is introduced in Section 3.1. Several extensions are presented in Section 3.2, and Section 3.3 provides further discussion.

3.1 A Hierarchical Model

We assume that we have access to M studies, which we divide into N groups. Each group comprises J baseline studies and one validation study, such that $M = N(J+1)$. We use the $i = 1, \ldots, N$ subscript to indicate the group, and $j = 1, \ldots, J+1$ to index the studies in each group, with $j = J+1$ being the validation study. The parameters are denoted by θ_{ij}. For instance, if $M = 12$ one could generate six groups of one ($J = 1$) baseline and one validation study, or one could create three groups with three ($J = 3$) baseline studies and one validation study. A discussion on how to choose N and J is deferred to Section 3.3.

Hierarchical Parameter Structure. Building on the Bayesian meta-analysis literature, we assume that the parameters estimated in the surveyed studies are highly correlated, but not identical.[3] This is captured by a hierarchical model in which the parameters θ_{ij}, $j = 1, \ldots, J+1$, of the linked studies have a common mean τ_i. The hierarchical modeling assumption takes the form

$$\theta_{ij} | (\tau_i, \nu) \overset{iid}{\sim} \mathcal{N}\big(\tau_i, \nu\big), \quad j = 1, \ldots, J+1. \tag{1}$$

For now we will assume that ν is identical for all groups $i = 1, \ldots, N$. This assumption will be relaxed in Section 4.4. The variance parameter ν can be interpreted as a measure of (the inverse of) external validity. If $\nu = 0$ then all populations examined in the linked studies share an identical parameter, i.e., $\theta_{ij} = \tau_i$ for all j. The larger ν, the more different the parameters are across studies within group i, which means that estimates from study ij carry less information about the parameters in studies $i\tilde{j}$, where $\tilde{j} \neq j$.

[3] There is an alternative approach to research synthesis that connects parameters across studies through bounds, e.g., Manski (2020) and Ishihara and Kitagawa (2024). However, this approach is not convenient for our goal of assessing coverage statements about validation study parameter estimates.

Aggregating the Baseline Estimates. We assume that study ij reports the estimator $\hat{\theta}_{ij}$ and a consistent standard error σ_{ij}. As most readers would, we interpret the reported estimates $\hat{\theta}_{ij}$ in a frequentist analysis as being (approximately) normally distributed:

$$\hat{\theta}_{ij}|\theta_{ij} \overset{iid}{\sim} \mathcal{N}\big(\theta_{ij}, \sigma_{ij}^2\big). \tag{2}$$

We omitted the $\hat{}$ from the standard error to emphasize that the sampling uncertainty is concentrated in $\hat{\theta}_{ij}$. Going forward, we omit qualifiers such as "approximately" or "asymptotically" because readers of empirical studies typically are left with no other choice than regarding the approximation error as negligible. (2) can be interpreted as a limited-information likelihood function for θ_{ij}; see, for instance, Chapter 18.4 of Pratt, Raiffa, and Schlaifer (1965), Doksum and Lo (1990), Kim (2002), Christensen, Moon, and Schorfheide (2023). We combine (1) and (2) to obtain:

$$\hat{\theta}_{ij}|(\tau_i, \nu) \overset{iid}{\sim} \mathcal{N}\big(\tau_i, \nu + \sigma_{ij}^2\big). \tag{3}$$

From a hierarchical Bayes perspective, the distribution of $\hat{\theta}_{ij}$ is centered at τ_i and its variance is the sum of two components: the variance of θ_{ij} conditional on τ_i, denoted by ν, and the sampling variance of the estimator itself. The joint density of all the baseline estimates in group i is given by

$$p\big(\hat{\theta}_{i,1:J}|\tau_i, \nu\big) \propto \prod_{j=1}^{J}(\nu + \sigma_{ij}^2)^{-1/2} \exp\left\{-\frac{1}{2}\sum_{j=1}^{J}\frac{(\hat{\theta}_{ij} - \tau_i)^2}{\nu + \sigma_{ij}^2}\right\}, \tag{4}$$

where $\hat{\theta}_{i,1:J} = \{\hat{\theta}_{i1}, \ldots, \hat{\theta}_{iJ}\}$.[4] Using this likelihood function, we construct a posterior distribution of τ_i based on the J baseline studies. Starting from a prior distribution for $\tau_i|\xi \sim \mathcal{N}(\underline{\tau}, \underline{V}_\tau)$ with hyperparameters $\xi = [\underline{\tau}, \underline{V}_\tau]$, standard calculations lead to the following posterior distribution for τ_i:

$$\tau_i|\big(\hat{\theta}_{i,1:J}, \nu, \xi\big) \sim \mathcal{N}\big(\bar{\tau}_i, \bar{V}_{\tau_i}\big), \tag{5}$$

where

$$\bar{V}_{\tau_i} = \left(\sum_{j=1}^{J}\frac{1}{\nu + \sigma_{ij}^2} + \underline{V}_\tau^{-1}\right)^{-1}, \quad \bar{\tau}_i = \bar{V}_{\tau_i}\left(\sum_{j=1}^{J}\frac{1}{\nu + \sigma_{ij}^2}\hat{\theta}_{ij} + \underline{V}_\tau^{-1}\underline{\tau}\right). \tag{6}$$

Rather than estimating ξ from the data, as is commonly done for hierarchical models, in the applications we use the improper prior $\underline{V}_\tau \longrightarrow \infty$. This centers the predictive intervals at a precision weighted average of $\hat{\theta}_{ij}$ without any shrinkage.

[4]The model could be extended to allow for a temporal evolution of the parameters θ_{ij}, or dependence in the estimators $\hat{\theta}_{ij}|\theta_{ij}$ caused by, for instance, overlapping estimation samples.

Predicting the Estimate of the Validation Study. We now turn to the predictive distribution for the validation study estimate, denoted by $\hat{\theta}_{i,J+1}$, which allows us to assess the statistical uncertainty quantification conditional on an assumption about the degree of external validity ν. Combining the hierarchical modeling assumption (1) with the posterior distribution of τ_i in (5), we deduce that

$$\hat{\theta}_{i,J+1}|(\hat{\theta}_{i,1:J},\nu,\xi) \sim \mathcal{N}\big(\bar{\tau}_i, \bar{V}_{\hat{\theta}_{i,J+1}}\big), \tag{7}$$

where

$$\bar{V}_{\hat{\theta}_{i,J+1}} = \bar{V}_{\tau_i} + \nu + \sigma^2_{i,J+1}. \tag{8}$$

The $1-\alpha$ highest-posterior-density (HPD) predictive interval takes the form

$$CI^{\hat{\theta}_{i,J+1}}\big(\hat{\theta}_{i,1:J};\nu,\xi\big) = \left[\bar{\tau}_i - z_{\alpha/2}\sqrt{\bar{V}_{\hat{\theta}_{i,J+1}}},\ \bar{\tau}_i + z_{\alpha/2}\sqrt{\bar{V}_{\hat{\theta}_{i,J+1}}}\right], \tag{9}$$

where $z_{\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution.

Example. Suppose that $J=1$. Using (6) we deduce that under the improper prior $\underline{V}_\tau = \infty$ the frequentist confidence interval for θ_{i1} derived from (2) and the predictive interval in (9) are both centered at $\hat{\theta}_{i1}$. The only difference is that the latter is wider, because it has to account for two additional sources of uncertainty. First, the parameter difference between baseline study $i1$ and validation study $i2$ adds ν to the variance of the predictive distribution. Second, the validation is based on the estimate $\hat{\theta}_{i,J+1}$ instead of the "true" value $\theta_{i,J+1}$, which generates the $\sigma^2_{i,J+1}$ term in (8).

Assessment. The preceding formulas make clear that we can only assess the measures of uncertainty provided in the baseline and validation studies jointly with the degree of external validity and the other hierarchical modeling assumptions. Our assessment mostly focuses on the empirical coverage frequency defined as

$$\text{CovFreq}_N = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left\{\hat{\theta}_{i,J+1} \in CI^{\hat{\theta}_{i,J+1}}\big(\hat{\theta}_{i,1:J};\nu,\xi\big)\right\}, \tag{10}$$

where $\mathbb{I}\{x \in A\}$ is the indicator function that is equal to one if $x \in \mathcal{A}$ and otherwise equal to zero. If the uncertainty statements obtained from the studies and the degree of external validity ν are well calibrated, then we should observe the coverage frequency to converge to $1-\alpha$. Use $\mathbb{P}^Y$ to denote the unconditional distribution of a random variable Y, $\mathbb{P}^Y_X$ the conditional distribution of Y given X, and $\mathbb{P}^{Y,X}$ the joint. One can use a suitable Law of Large Numbers to deduce that

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left\{\hat{\theta}_{i,J+1} \in CI^{\hat{\theta}_{i,J+1}}\big(\hat{\theta}_{i,1:J};\nu,\xi\big)\right\} - \mathbb{P}^{\hat{\theta}_{i,1:J+1}}_{\nu,\xi}\Big(\hat{\theta}_{i,J+1} \in CI^{\hat{\theta}_{i,J+1}}\big(\hat{\theta}_{i,1:J};\nu,\xi\big)\Big) \xrightarrow{p} 0 \tag{11}$$

as $N \longrightarrow \infty$. The probabilities can be manipulated as follows:

$$\begin{aligned}
&\mathbb{P}_{\nu,\xi}^{\hat{\theta}_{i,1:J+1}}\left(\hat{\theta}_{i,J+1} \in CI^{\hat{\theta}_{i,J+1}}\left(\hat{\theta}_{i,1:J};\nu,\xi\right)\right) \\
&= \mathbb{P}_{\nu,\xi}^{\hat{\theta}_{i,1:J}}\left(\mathbb{P}_{\nu,\xi,\hat{\theta}_{i,1:J}}^{\hat{\theta}_{i,J+1}}\left(\hat{\theta}_{i,J+1} \in CI^{\hat{\theta}_{i,J+1}}\left(\hat{\theta}_{i,1:J};\nu,\xi\right)\right)\right) \\
&= 1-\alpha.
\end{aligned}$$

Theorem 2.1 in Liu, Moon, and Schorfheide (2023) provides a formal proof for the case in which (ν,ξ) are replaced by estimates $(\hat{\nu},\hat{\xi})$ that are consistent as $N \longrightarrow \infty$.

In addition to empirical coverage frequencies, we also consider statistics derived from probability integral transforms (PITs).[5] For group i the PIT is defined as

$$PIT^{\hat{\theta}_{i,J+1}}\left(\hat{\theta}_{i,1:J};\nu,\xi\right) = \Phi_N^{-1}\left(\frac{\hat{\theta}_{i,J+1}-\bar{\tau}_i}{\sqrt{\bar{V}_{\hat{\theta}_{i,J+1}}}}\right), \tag{12}$$

where $\Phi_N(\cdot)$ is the cumulative distribution function (cdf) of a standard normal random variable. Define the empirical cdf of the PITs across groups as

$$\hat{F}_N^{pit}(u) = \sum_{i=1}^{N} \mathbb{I}\left\{PIT^{\hat{\theta}_{i,J+1}}\left(\hat{\theta}_{i,1:J};\nu,\xi\right) \leq u\right\}. \tag{13}$$

One can show that statistics derived from $\hat{F}_N^{pit}(u)$ converge to the corresponding statistics computed from uniform $\mathcal{U}[0,1]$ distribution as $N \longrightarrow \infty$. The finite sample properties of the PIT statistics are generally more sensitive to the distributional assumptions than the empirical coverage frequencies.

3.2 Implementation Details

Treatment of Hyperparameters. The subsequent empirical analysis is based on the improper prior obtained from $\xi_\infty = [\underline{\tau} = 0, \underline{V}_\tau = \infty]$. In regard to the external validity parameter ν we consider the following options:

Option 1. Impose the belief that all studies in group i estimate the same parameter $\tau_i = \theta_{ij}$ for $j = 1,\ldots,J+1$ by setting $\nu = 0$.

Option 2. Plot the empirical coverage frequency as a function of ν. (8) implies that this function is weakly increasing.

[5] Smith, Tebaldi, Nychka, and Mearns (2009) use PITs to evaluate the exchangeability assumption in a Bayesian hierarchical model.

Option 3. Construct an empirical Bayes estimate $\hat{\nu}$ by maximizing the log marginal data density (MDD) with respect to ν conditional on ξ:

$$\hat{\nu} = \text{argmax}_{\nu \geq 0} \ \ln p(\hat{\theta}_{i,1:J}|\nu,\xi), \quad p(\hat{\theta}_{i,1:J}|\nu,\xi) = \int p(\hat{\theta}_{i,1:J}|\tau_i,\nu)p(\tau_i|\xi)d\tau_i. \tag{14}$$

A formula is provided in the Online Appendix. Strictly speaking, the log MDD is not well defined under the improper prior obtained by setting $\underline{V}_\tau = \infty$. However, the log MDD differential for two values ν and $\tilde{\nu} = 0$ has a well-defined limit as $\underline{V}_\tau \longrightarrow \infty$.

Option 4. Allow the degree of external validity to be group specific and replace the homogeneous ν by group-specific ν_is. This approach requires at least a modest number of baseline studies J and a hyperprior on ν_i which after augmenting the ξ vector could be written as $p(\nu_i|\xi)$; see Liu (2023) and Liu, Moon, and Schorfheide (2023) for possible implementations.

Choice of (J, N). Suppose that the total number of studies is M. From a statistical perspective the trade-off between J and N is as follows: the larger J, the more sample information there is to estimate τ_i and ν_i in case the degree of external validity is treated as heterogeneous. The larger N, the more precise the approximation of population averages through empirical frequencies becomes; see for instance the convergence statement in (11). From a practical perspective, the splits are mostly driven by the plausibility of assuming that subsets of studies have similar population parameters θ_{ij}s so that the degree of external validity is plausibly high and homogeneous across groups.

3.3 Further Discussion

Interpretation of the Assessment. First, at best we can assess average coverage probabilities. Suppose the nominal coverage probability is 80%, but for 50% of the groups we have 70% intervals and the other 50% provide 90% intervals. Then the empirical frequency would converge to the average coverage probability which is the nominal 80% coverage. This is true for any evaluation of interval and density predictions in the forecasting literature. Second, the assessment of coverage statements or, more broadly speaking, measures of uncertainty, is intrinsically linked to beliefs or evidence about external validity, i.e., similarity of population parameters across studies within a group. In fact, one of the recommended diagnostics is to plot the coverage frequency as a function of ν and examine where it crosses the nominal coverage probability. Third, formal tests based on our framework are sensitive to distributional assumptions underlying the hierarchical model. Thus, as an alternative to the Gaussian predictive intervals, we consider the robust intervals recently proposed by Armstrong, Kolesár,

and Plagborg-Møller (2022). For large enough (N, J) one could use more sophisticated nonparametric approaches that have been used in the Bayesian panel data literature, e.g., Liu (2023) and Liu, Moon, and Schorfheide (2023), but we do not pursue this approach in the current paper.

Publication Bias. In the presence of publication bias the distribution of published parameter estimates is different from the pre-publication distribution generated by the research community. In this case our diagnostic tool will find discrepancies between empirical coverage frequencies and stated coverage probabilities, providing a joint test of publication bias and inaccurate uncertainty quantification. Alternatively, if one wants to partial out the effect of publication bias and only test inaccurate uncertainty quantification, then one can replace the normal likelihood $p(\hat{\theta}_{ij}|\theta_{ij})$ in (2) with the one proposed in Andrews and Kasy (2019), which assigns different publishing probabilities when the corresponding t statistics fall in different intervals.

Correlated Parameter Estimates. We assumed that parameter estimates conditional on θ_{ij} are independent across studies. That is a reasonable assumption in microeconomic settings, where studies might rely on data sets from different regions and points in time, as in the empirical applications considered in Sections 4 to 6. In macroeconomics, on the other hand, estimates are often based on overlapping data sets. For instance, one of the estimates plotted in Figure 1 is based on observations from 1980 to 2000, whereas another one uses data from 1965 to 2001. Moreover, in the DSGE model literature and the literature on structural vector autoregressions (VARs), authors often use overlapping sets of variables, e.g., the first study estimates a model based on unemployment, CPI inflation, and interest rates, whereas the second paper uses GDP growth, GDP deflator inflation, and interest rates. In such settings, the conditional independence assumption becomes implausible and the marginal distributions in (2) have to be replaced by a joint distribution that captures the correlation structure.

Identification and Model-Specific Parameters. In the structural VAR literature, it is conceivable that two papers use identical data sets and reduced-form VAR specifications, but estimates of structural parameters, e.g., the response of inflation to a monetary policy shock upon impact, differ because of the identification assumptions that lead from the reduced form to the structural form. In an extreme case, estimates could be based on VARs $j = 1, \ldots, J+1$ of the form $y_t = \Phi y_{t-1} + u_t$, $u_t = \Sigma_{tr}\Omega_j\epsilon_t$. Here Σ_{tr} is the Cholesky factor of the covariance matrix of u_t, (Φ, Σ) are identifiable reduced-form parameters, and the Ω_js are non-identifiable orthonormal matrices that determine the effect of structural shocks ϵ_t on the reduced-form

shocks u_t. Our hierarchical model cannot reconcile differences in impulse response estimates due to differences in Ω_j. Instead, one would have to use a common Ω for all studies in a group.

Some parameters, such as the slope of the NKPC, are intrinsically tied to the specific structure of, say, a DSGE model. A value of 0.1 may have very different effects on observables across two models. Transformations of preference and technology parameters, such as the contemporaneous effect of an unanticipated 25 basis point drop in the central bank's target interest rate, are more closely linked to measurable phenomena that can be defined in the context of a large class of empirical models. Our analysis is more suitable for the latter than the former class of parameters.

Studies Reporting Posteriors. We have assumed that the underlying studies report frequentist estimates of θ_{ij}, which are then linked through a Bayesian model. However, it is conceivable that a subset or all of the studies report posterior distributions for θ_{ij}. One way of proceeding would be to appeal to a Bernstein-von-Mises result of large sample equivalence between the posterior distribution $\theta_{ij}|\hat{\theta}_{ij}$ and the sampling distribution $\hat{\theta}_{ij}|\theta_{ij}$, and then proceed as described in Section 3.1. Alternatively, one could use a version of the Bayesian predictive synthesis approach proposed by McAlinn and West (2019) and adapted to the synthesis of causal estimates by Sugasawa, Takanashi, and McAlinn (2023). Because the estimates used in the empirical applications in Sections 4 to 6 are frequentist, we do not pursue the assessment of posterior uncertainty statements.

4 Estimates From the "Many Labs" Replication

The first application is based on studies that were conducted as part of a large-scale replication project, published by Klein, Ratcliff, 48 others, and Nosek (2014). This paper collects estimates of 13 classic and contemporary effects in psychology, such as gain versus loss framing, anchoring, and gambler's fallacy, among others. For instance, in the gain versus loss framing experiment, individuals are divided into two groups. For Group 1 outcomes are described as gains, e.g., Program A will save 200 people, whereas for Group 2 outcomes will be described in terms of losses, e.g., 400 people will die under Program A. The experimenter then records what fraction of subjects choose Program A (or one of the other programs) within Group 1 and Group 2. Based on this information a causal effect estimate can be computed.

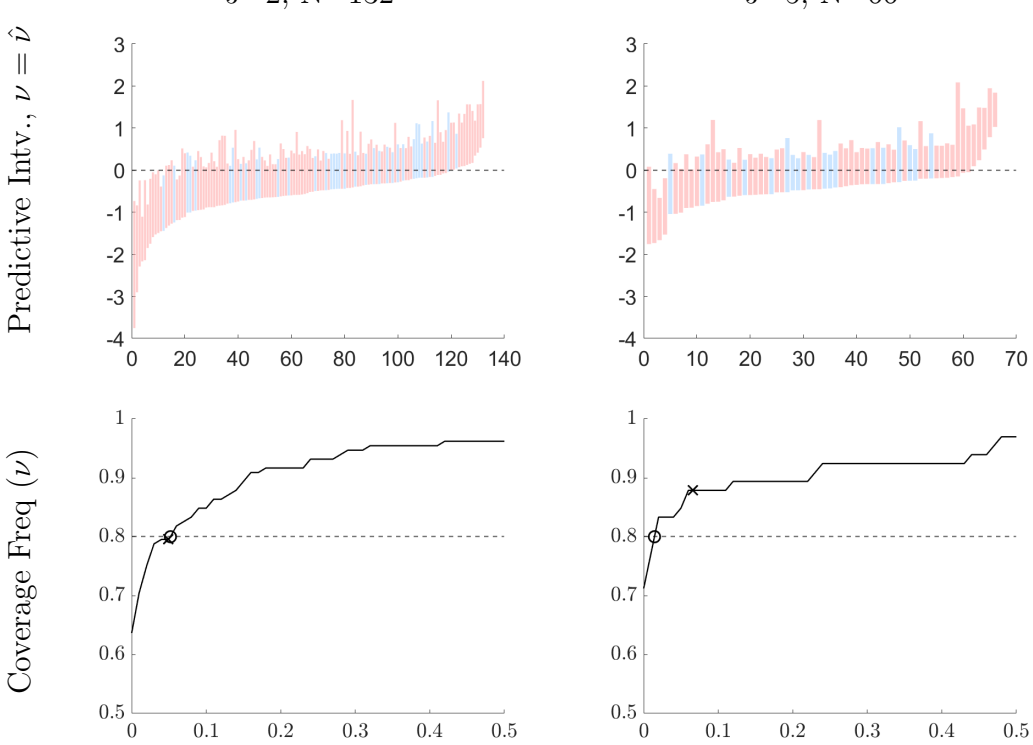
Klein, Ratcliff, 48 others, and Nosek (2014) examined whether the 13 aforementioned effects, that had been documented in the original studies, could be replicated in other laboratories. Here replicability refers to the similarity of the estimated effect between the original study and the new experiments conducted by the authors. Overall, 36 study sites participated in this project and collected data from a total of 6,344 participants. Of these sites, 27 have studies conducted in a laboratory setting and nine online; 25 are located in the U.S., and 11 in other countries. The authors found that ten of the 13 effects replicated consistently. From our perspective, the replication study provides a set of independent estimates and standard errors $(\hat{\theta}_{ij}, \sigma_{ij})$ of comparable population parameters θ_{ij}. Section 4.1 presents the benchmark analysis using the hierarchical model of Section 3. We then consider three extensions: interval predictions that are robust to deviations from the normality assumption in (1) (Section 4.2), an analysis under an alternative grouping of studies (Section 4.3), and heterogeneous ν_is (Section 4.4).

4.1 Benchmark Analysis

Grouping of Studies and Estimates. The hierarchical model postulates a common τ_i among the $J+1$ studies belonging to group i. Thus, our goal is to group the studies such that the estimands θ_{ij} are similar within each group i. We use all 36 site locations and consider estimates of eight effects for which we have point estimates and standard errors. Out of the eight effects, seven allow us to construct one estimate from each study site. For the anchoring effect, which examines participants' estimates of four specific outcomes, we can generate four estimates from each site. This leads to 11 experimental designs and gives us a total of $M_f = 36 * 11 = 396$ estimates. We refer to this set of estimates as the *full sample*. We also construct a *reduced* sample by dropping three effects that show weak or no replicability. This sample comprises $M_r = 36 * 8 = 288$ estimates.

In each of the 11 experimental designs there is a control group of N_c participants and a treatment group of N_t participants. Each site reports the mean and standard error of the outcome variable for participants in each group, denoted as μ_c, μ_t, s_c, and s_t, respectively. We then compute the estimate $\hat{\theta}$ as $\hat{\theta} = \mu_t - \mu_c$, and the variance σ^2 as $\sigma^2 = s_c^2/N_c + s_t^2/N_t$ (assuming that the treatment and control samples are independent). The plausibility of the hierarchical modeling assumptions in Section 3.1, in particular the assumption of a homogeneous ν, depends on the scaling of θ_{ij} and $\hat{\theta}_{ij}$. Only the estimates from the four anchoring-effect experiments need to be rescaled to ensure that all $\hat{\theta}_{ij}$ are commensurable.

Figure 2: Predictive Intervals and Coverage for "Many Labs," Full Sample



Notes: Top row: predictive intervals conditional on $\nu = \hat{\nu}$. The bars are arranged in ascending order of the predictive intervals' lower bounds. Blue bars represent studies with designs identified as weakly- or non-replicable by Klein, Ratcliff, 48 others, and Nosek (2014). Bottom row: empirical coverage frequency as a function of ν. Circles indicate the values of ν that achieve an empirical coverage frequency of 0.8, while crosses represent the $\hat{\nu}$ estimates.

We consider $J = 2$ and $J = 5$ baseline studies per group, which allows us to either generate $N = 132$ or $N = 66$ groups for the full sample. To create a benchmark grouping, we group sites with similar characteristics (U.S. versus international, online versus lab) together, conjecturing that they have similar θ_{ij}s. For the full sample and $J = 5$, each i corresponds to a group of six sites and one of the 11 experimental designs. For $J = 2$, we divide each of the six site-groups into two separate groups and combine them with one of the eleven designs. For the reduced sample we have the same site structure, but only eight designs, which leads to $N = 96$ for $J = 2$ and $N = 48$ for $J = 5$. A table with the group definitions and additional details about the construction of $\hat{\theta}_{ij}$ and σ^2_{ij} can be found in the Online Appendix. Within each group, we randomly choose one study to be the validation study $J + 1$ and designate the remaining ones as baseline studies.

Results. Based on the full sample, the panels in the top row of Figure 2 plot the predictive

Table 1: Empirical Coverage Frequency for "Many Labs"

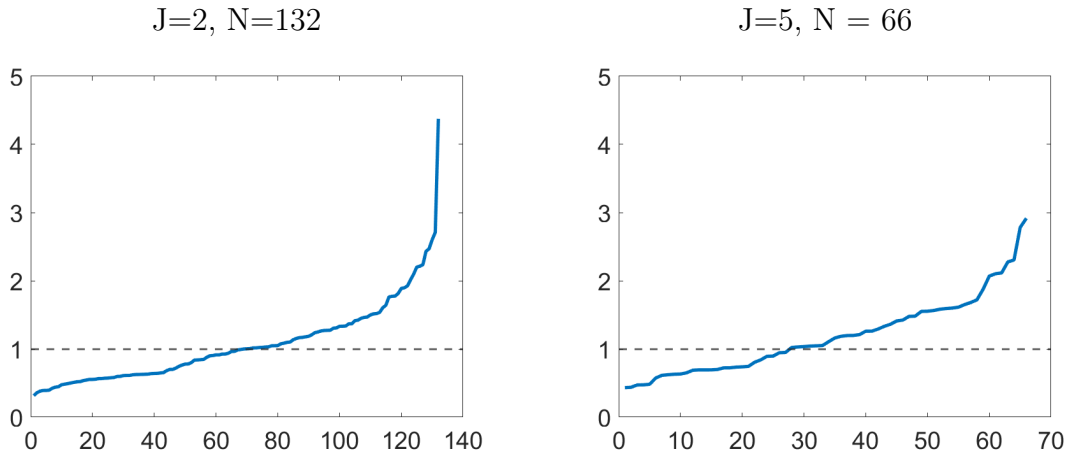
	J=2		J=5	
	N = 132	N = 96	N = 66	N = 48
Identical Parameters: $\nu = 0$				
Emp. Coverage Freq.	0.64	0.54	0.71	0.63
Coverage p-value	0.00	0.00	0.09	0.00
Empirical Bayes Estimate: $\nu = \hat{\nu}$				
Emp. Coverage Freq.	0.80	0.76	0.88	0.81
Coverage p-value	0.91	0.37	0.13	0.87
$\hat{\nu}$	.049	.072	.066	.099

interval for $\hat{\theta}_{i,J+1}$ for each group i. Here we consider the empirical Bayes approach that conditions on an estimate $\hat{\nu}$. The left panel corresponds to $J = 2$ and the right panel to $J = 5$. We normalize $\hat{\theta}_{i,J+1}$ to zero and adjust the corresponding predictive intervals accordingly, allowing us to interpret the horizontal line $y = 0$ as actual values. The groups are sorted in ascending order based on the lower bound of the predictive interval. Because the nominal coverage probability is set to $\alpha = 0.8$ we would expect 20% of the re-centered intervals to exclude zero. Predictive intervals for studies that are considered to be weakly- or non-replicable are indicated by blue bars.

The panels in the bottom row of Figure 2 show the empirical coverage frequency across groups as function of ν. As previously discussed under *Option 2* in Section 3.2, this function is monotonically increasing because the width of the predictive interval is increasing in ν. The horizontal line indicates the nominal coverage probability of 80%. The circle indicates the value of ν at which the coverage frequency function intersects with the 80% line and the cross indicates the coverage frequency at the empirical Bayes estimate $\hat{\nu}$. In the application $\hat{\nu}$ yields coverage frequencies that slightly exceed the nominal level of 80%. For $\nu = 0$ the coverage frequencies are 0.64 ($J = 2$) and 0.71 ($J = 5$), respectively, indicating that the predictive intervals are too small if one believes that the studies in each group estimate identical population parameters.

Table 1 summarizes the coverage frequencies for $\nu = 0$ and $\nu = \hat{\nu}$, using the full and reduced sample. Conditional on the belief that the studies within a group i estimate the same population parameter τ_i, the researchers understate the uncertainty associated with their estimates. The empirical coverage frequency is at most 0.71 across the different samples

Figure 3: Ratio of $\sqrt{\hat{\nu}}$ to Average Standard Error, Full Sample



Notes: The figure depicts $r_i = \sqrt{\hat{\nu}}/\frac{1}{J+1}\sum_{j=1}^{J+1}\sigma_{ij}$ and groups i are sorted in ascending order of r_i.

and groupings. Moreover, three out of four p-values for the hypothesis that the coverage probability is 80% are zero.[6] Once we allow for differences of population parameters within groups, the coverage frequencies are much closer to the target level of 80% and all p-values exceed 0.10. The estimates of ν range from 0.05 to 0.1.

Without further information, the numerical values of $\hat{\nu}$ are difficult to interpret. In view of our discussion of Figure 1 in Section 2 it seems sensible to relate the estimated degree of external validity $\hat{\nu}$ to the variance of the estimators σ_{ij}^2. For each group i we compute the ratio

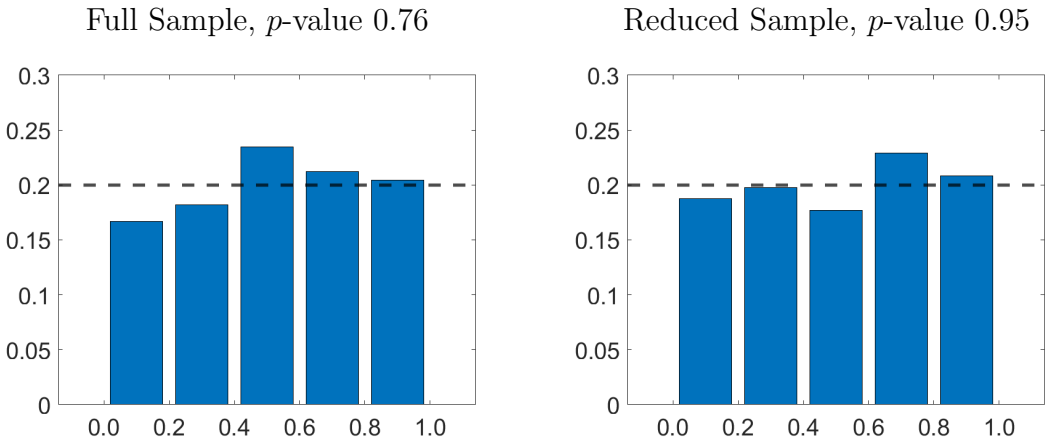
$$r_i = \frac{\sqrt{\hat{\nu}}}{\frac{1}{J+1}\sum_{j=1}^{J+1}\sigma_{ij}}$$

and plot it in Figure 3. If r_i is close to one, the dispersion of the population coefficients is approximately equal to the average standard errors of the estimators. The groups are now sorted in ascending order of r_i which leads to a function that by construction is monotonically increasing. For groups with a high r_i the average dispersion of population parameters is substantially larger than the standard errors, and concerns about external validity are more important than sampling uncertainty.

Finally we compute PITs and sort their values into five equally spaced bins. The resulting histograms for $J = 2$ are plotted in Figure 4. To assess how far these histograms deviate

[6]Under the null hypothesis that the coverage probability is 80%, the indicator functions from which the coverage frequency is constructed, see (10), are Bernoulli random variables, which makes it straightforward to simulate the distribution of CovFreq_N.

Figure 4: PIT Histograms for "Many Labs," $J = 2$ and $\nu = \hat{\nu}$



Notes: The p-values are computed via simulation for S_{pit} statistic under the null hypothesis that the PITs are uniformly distributed.

from uniformity we compute the S_{pit} statistic

$$S_{pit} = \sum_{j=1}^{5} \frac{(n_j - N/5)^2}{N/5},$$

where n_j is the number of PITs in bin j. The statistic is zero if all bins contain the same number of PITs, and is greater than zero otherwise. Because under the null hypothesis that the predictive distribution is correctly specified, the PITs have a uniform distribution, it is straightforward to simulate the sampling distribution of the S_{pit} statistic and compute p-values which we also report in the figure. For $\hat{\nu}$ the PIT histograms look fairly uniform and the p-values are 0.76 and 0.95, respectively.

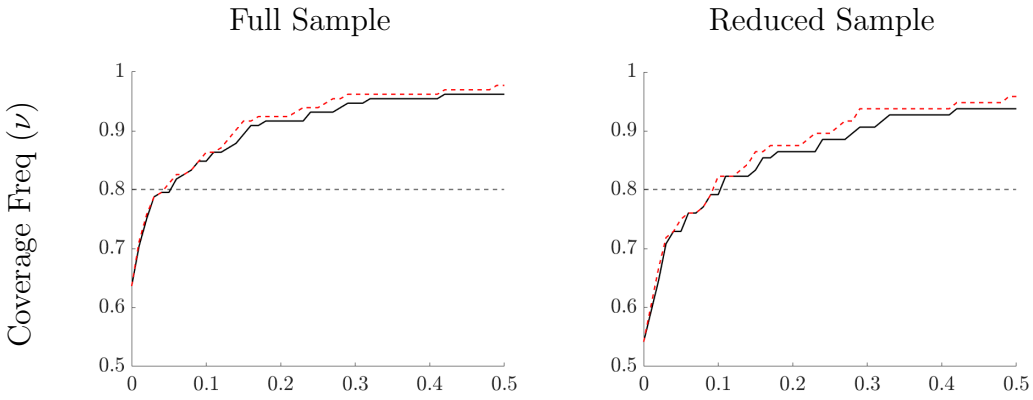
4.2 Robust Predictive Intervals

The calculations in the previous section rely on the assumption that $\theta_{ij}|(\tau_i, \nu)$ is normally distributed. Armstrong, Kolesár, and Plagborg-Møller (2022) developed a method to robustify the predictive intervals against deviations from normality.[7] Point of departure from the previous analysis is the assumption that

$$\theta_{ij}|p_i \overset{iid}{\sim} p_i(\cdot), \quad \mathbb{E}[\theta_{ij}|p_i] = \tau_i, \quad \mathbb{V}[\theta_{ij}|p_i] = \nu < \infty, \tag{15}$$

[7]We are grateful to our discussant Tim Armstrong for proposing this robustness analysis.

Figure 5: Robust Intervals: Coverage Freq. for $J = 2$



Notes: The solid black lines depict coverage as a function of ν under benchmark normal assumption, dashed red lines are based on robust predictive intervals.

where $p_i(\cdot)$ is a distribution with finite second moments. Notice from (6) that for $\underline{V}_\tau = \infty$ we can write the posterior mean as $\bar{\tau}_i = \sum_{i=1}^J w_{ij}\hat{\theta}_{ij}$. The weights w_{ij} are fixed conditional on ν. The prediction error $\hat{\theta}_{i,J+1} - \bar{\tau}_i$ can be decomposed as follows:

$$\hat{\theta}_{i,J+1} - \bar{\tau}_i = \underbrace{\left(\theta_{i,J+1} - \sum_{j=1}^J w_{ij}\theta_{ij}\right)}_{A_i} + \underbrace{\left(\sigma_{i,J+1}u_{i,J+1} - \sum_{j=1}^J w_{ij}\sigma_{ij}u_{ij}\right)}_{B_i}, \tag{16}$$

where $u_{ij} = (\hat{\theta}_{ij} - \theta_{ij})/\sigma_{ij}$ is iid $\mathcal{N}(0,1)$ conditional on the θ_{ij}s, $j = 1, \ldots, J+1$ according to (2). We deduce that $B_i|(A_i, p_i) \sim \mathcal{N}(0, V_{B_i})$ with $V_{B_i} = \sigma^2_{i,J+1} + \Sigma_{j=1}^J w_{ij}^2\sigma_{ij}^2$. The distribution of A_i depends on p_i, but we can characterize its first two moments:

$$\mathbb{E}[A_i|p_i] = 0, \quad \mathbb{E}[A_i^2|p_i] = \left(1 + \sum_{j=1}^J w_{ij}^2\right)\nu. \tag{17}$$

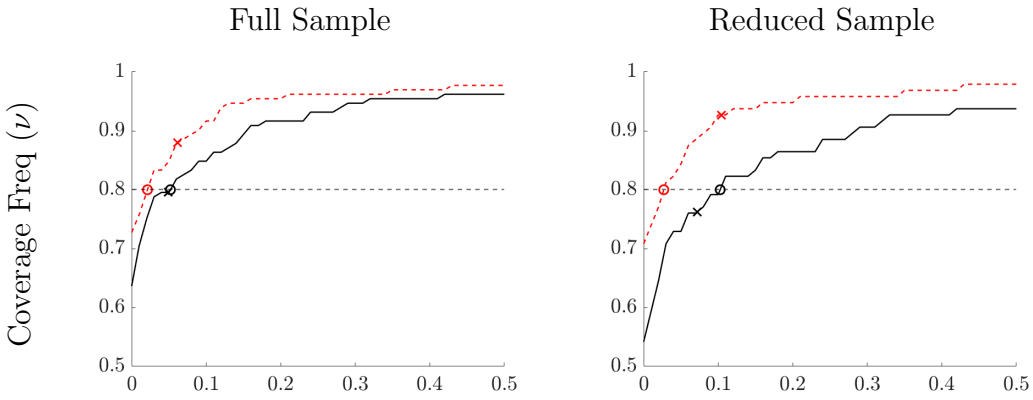
The robust predictive interval is of the form

$$CI^{\hat{\theta}_{i,J+1}}\left(\hat{\theta}_{i,1:J}; \nu\right) = \left[\bar{\tau}_i - \chi_i,\ \bar{\tau}_i + \chi_i\right]. \tag{18}$$

The bound χ_i is the largest possible $1-\alpha$ quantile of $|A_i(p_i) + B_i|$. The maximization is over the distributions p_i subject to the moment constraints (17). We use the software provided by Armstrong, Kolesár, and Plagborg-Møller (2022) to compute χ_i.

Figure 5 overlays the empirical coverage frequencies of predictive intervals constructed with and without the normality assumption for the "Many Labs" application. The latter

Figure 6: Alternative Grouping: Coverage Freq. for $J = 2$



Notes: Solid black and dashed red lines correspond to the benchmark and alternative groupings, respectively. Circles indicate the values of ν that achieve an empirical coverage frequency of 0.8, while crosses represent the $\hat{\nu}$ estimates. Symbol colors match their respective groupings.

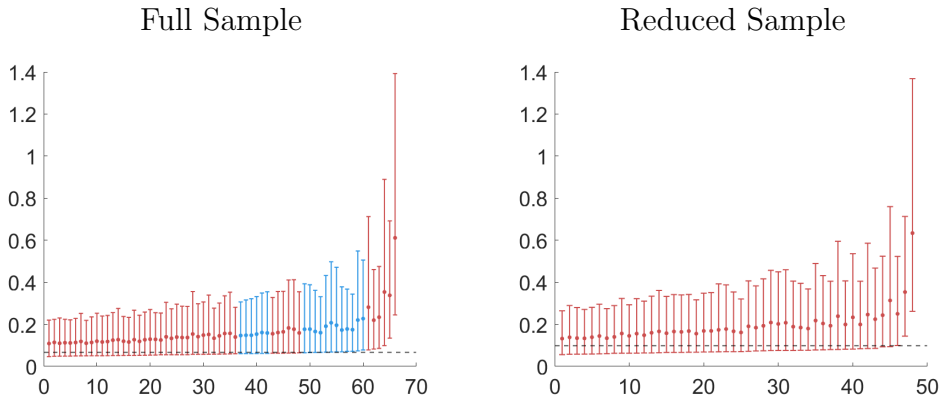
always yield higher coverage frequencies, because the robust predictive intervals are wider by design. However, the increase in coverage is relatively small and the conclusions from the benchmark analysis do not change.

4.3 Alternative Groupings

We now consider an alternative grouping of the estimates. As in the benchmark analysis, we group sites with similar characteristics (U.S. versus international, online versus lab) together. However, we reshuffle the studies that share the same characteristics. We manually construct the alternative grouping to maximize its deviation from the original assignment. For instance, study sites WL, UVA, and VCU originally belonged to Groups 4, 5, and 6, respectively, but are deliberately grouped together in Group 4 under the alternative assignment. The specific grouping can be found in the Online Appendix.

Empirical coverage frequencies for the benchmark grouping and the alternative grouping are plotted in Figure 6. For a given ν the coverage frequency under the alternative grouping is larger than under the benchmark grouping, which implies that for small values of ν it is closer to the nominal coverage probability of 80%. For instance, the coverage frequency at $\nu = 0$ rises from 0.64 to 0.73 for the full sample. The p-value also increases, from 0 to 0.03, but it stays below 5%.

Figure 7: Posterior for ν_i for $J = 5$



Notes: 90% credible intervals for ν_i with posterior mean estimates. The intervals are arranged in ascending order of their lower bounds. Blue intervals in the left panel represent studies with designs identified as weakly- or non-replicable by Klein, Ratcliff, 48 Others, and Nosek (2014). The dashed black line horizontal lines represent the $\hat{\nu}$ estimates for the two samples.

4.4 Heterogeneous ν_i

In the benchmark analysis we assumed that the external validity parameter ν is homogeneous across groups i. We now relax this assumption and replace (1) by

$$\theta_{ij}|(\tau_i, \nu) \overset{iid}{\sim} \mathcal{N}\big(\tau_i, \nu_i\big), \quad j = 1, \ldots, J+1. \tag{19}$$

Because J is small, we will use $J = 5$ below. We add a prior distribution for ν_i, following Liu, Moon, and Schorfheide (2023):

$$\nu_i \sim IG\left(3, 2\left(\frac{\kappa}{N}\sum_{i=1}^{N}\hat{\mathbb{V}}(\hat{\theta}_{i\cdot})\right)\right), \tag{20}$$

where $\hat{\mathbb{V}}(\hat{\theta}_{i\cdot})$ is the sample variance of $\hat{\theta}_{i1}, \ldots, \hat{\theta}_{iJ}$ and κ is a tuning constant to be chosen by the researcher. We use a simple Random-Walk Metropolis-Hastings algorithm (see, for instance, Herbst and Schorfheide (2015) for a description) to sample from the posterior of ν_i and generate draws from the posterior predictive distribution of $\hat{\theta}_{i,J+1}$:

$$p(\nu_i|\hat{\theta}_{i,1:J}, \xi) \propto p(\hat{\theta}_{i,1:J}|\nu_i, \xi)p(\nu_i|\xi). \tag{21}$$

Here $\propto$ denotes proportionality, the MDD $p(\hat{\theta}_{i,1:J}|\nu_i, \xi)$ is the same as in (14) with ν replaced by ν_i and $p(\nu_i|\xi)$ is the probability density function (pdf) associated with (20).

90% posterior credible intervals and posterior mean estimates for the ν_is are plotted in Figure 7. The posterior mean estimates are slightly larger than the $\hat{\nu}$ estimate generated

Table 2: Empirical Coverage Frequency for "Many Labs," Homo. and Hetero. ν_i

	N=66, J=5		N=48, J=5	
	Homosk.	Heterosk.	Homosk.	Heterosk.
Emp. Coverage Freq.	0.88	0.91	0.81	0.85
Coverage p-value	0.13	0.03	0.87	0.38

Notes: Homosk(edasticity) is the benchmark specification with homogeneous ν, and Heterosk(edasticity) is the alternative specification with heterogeneous ν_is.

previously. Overall, there is no evidence for heterogeneity, except possibly in a small number of groups. Empirical coverage frequencies and associated p-values are reported in Table 2. Because the heterogeneous ν_i estimates are slightly larger than $\hat{\nu}$, so are the predictive intervals, which moves the coverage frequency further away from the target level of 80% and reduces the p-value. The main problem with the heterogeneous specification is that the group size is too small to generate precise estimates of ν_i.

5 Estimates of the Impacts of Microcredit Expansions

The second application uses estimates of the impact of microcredit expansions collected by Meager (2019). The estimates are based on RCTs that adopt various sampling strategies, experimental designs, and econometric strategies to identify causal effects of expanded access to credit from microfinance institutions on household business and consumption outcomes.

Grouping of Studies and Estimates. We use the term "articles" to refer to the research papers synthesized in Meager (2019), and "studies" to denote results associated with a particular outcome variable. Each article contains multiple studies. We create random groups i with $J_i = 1$ or $J_i = 2$ from studies that examine the same outcome variable.[8] Out of the six outcomes considered in the set of articles, we can construct three groups each for "Profit," "Revenues," and "Expenditures," and two groups each for "Consumption," "Durables," and "Temptation." In total, this leads to $N = 15$ groups. Table 3 reports the estimates and standard errors of causal effects from the seven papers summarized by Meager (2019). We observe significant variation in the estimates across studies.

Results for Benchmark Grouping. The predictive intervals for the $N = 15$ groups are plotted in the left and center panels of Figure 8 for $\nu = 0$ and $\nu = \hat{\nu} = 8.81$. The ordering

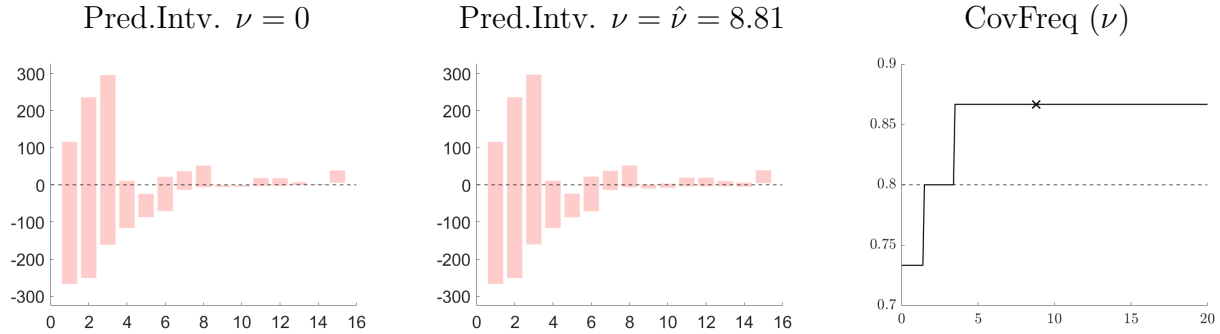
[8]The $J_i = 1$ groups do not contribute to the estimation of ν.

Table 3: Estimates from Papers Included in Meager (2019)

Paper	Profit		Revenues		Expenditures	
	OLS	SE	OLS	SE	OLS	SE
Angelucci, Karlan, and Zinman (2015)	-4.55	5.88	9.74	4.29	12.94	6.10
Attanasio et al. (2015)	-0.34	0.22	-0.07	0.18	0.27	0.32
Augsburg et al. (2015)	37.53	19.78	78.41	36.86	32.40	21.19
Banerjee et al. (2015)	16.72	11.83	25.56	33.17	7.85	29.21
Crépon et al. (2015)	17.54	11.40	65.46	24.03	31.84	17.24
Karlan and Zinman (2011)	66.56	78.13	-67.03	178.25	-62.59	149.08
Tarozzi, Desai, and Johnson (2015)	7.29	7.89	10.85	8.09	3.56	1.73
Paper	Consumption		Durables		Temptation	
	OLS	SE	OLS	SE	OLS	SE
Angelucci, Karlan, and Zinman (2015)	5.51	4.10			-0.08	0.09
Attanasio et al. (2015)	50.45	32.18	1.07	0.50	1.52	2.10
Augsburg et al. (2015)	-5.17	19.24	-6.50	189.80	-5.80	2.82
Banerjee et al. (2015)	4.64	5.48	4.44	2.25	-1.64	0.58
Crépon et al. (2015)	-2.94	6.02	1.38	2.26	-0.42	0.72

Notes: As in Meager (2019), we re-estimate the causal effects using OLS on the data provided by the papers. All units are standardized to USD PPP over a two week period (indexed to 2010 dollars). The blank spaces in the table indicate that the original papers did not provide the relevant data.

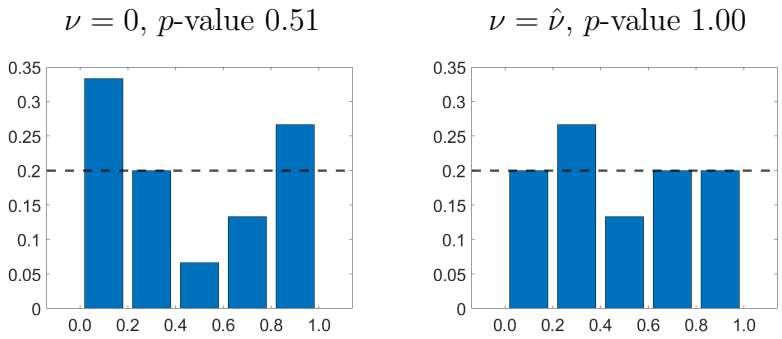
Figure 8: Predictive Intervals and Coverage Frequencies for Meager (2019)



Notes: Left and center panels: the groups in the two bar charts are arranged in ascending order of the predictive intervals' lower bounds in the left panel. Right panel: empirical coverage frequency as a function of ν. The cross represents $\hat{\nu}$. The coverage frequency is 0.8 for values of ν ranging from 1.5 to 3.4.

of the groups is identical in both panels. Unlike in the top row of Figure 2, which showed the predictive intervals for the "Many Labs" application, we see a considerable variation in interval lengths across groups in the microcredit application. This variation is driven by

Figure 9: PIT Histograms for Meager (2019)



Notes: The p-values are computed via simulation for S_{pit} statistic under the null hypothesis that the PITs are uniformly distributed.

the large variation in standard errors across studies; see Table 3. The effect of increasing ν on the predictive intervals is barely visible, but shifts two intervals from not covering zero to covering zero. For 12 of the 15 groups the ratio of $\sqrt{\nu}$ to the average standard error of estimates in each group is below one. The largest value is 5.8.

The right panel of Figure 8 plots the empirical coverage frequency as a function of ν. Because the number of groups is considerably smaller than in the "Many Labs" application, the steps in the coverage frequency function are more pronounced. At $\nu = 0$ the coverage frequency is 0.73, whereas it is 0.87 for $\nu = \hat{\nu} = 8.81$. The coverage p-values are 0.53 and 0.75, respectively. Due to the small sample size, the assessment does not have a lot of power to detect violations.[9] The nominal coverage probability of 0.8 is achieved for values of ν ranging from 1.5 to 3.4, which are smaller than the estimate $\hat{\nu} = 8.81$. The discrepancy is larger than in the "Many Labs" application. $\hat{\nu}$ is obtained without information about $(\hat{\theta}_{i,J+1}, \sigma_{i,J+1})$ which makes the empirical coverage frequency a pseudo-out-of-sample statistic, whereas the value of ν for which the coverage frequency equals 80% is an in-sample statistic.

We plot PIT histograms in Figure 9. For $\nu = 0$ the histogram has a U-shape, meaning that the Gaussian predictive distribution is too concentrated and disproportionately many validation sample estimates fall into the tails. Due to the small sample of $N = 15$, the visual deviation of the histogram from uniformity is, however, not significant: the p-value is 0.51. For $\nu = \hat{\nu} = 8.81$ the U-shape vanishes,the histogram looks almost uniform, and the p-value rises to one.

[9]The variance of a Bernoulli$(1-\alpha)$ random variable is $\alpha(1-\alpha)$. If $\alpha = 0.2$, then the standard deviation of a sample average of 15 observations would be close to 0.1.

Table 4: Results for Alternative Groupings

	$\nu = 0$		$\nu = \hat{\nu}$		
Grouping	CovFreq	p-Value	$\hat{\nu}$	CovFreq	p-Value
Benchmark	0.73	0.53	8.81	0.87	0.75
Alt 1	0.53	0.02	0.00	0.53	0.02
Alt 2	0.60	0.06	0.00	0.60	0.06
Alt 3	0.60	0.06	1.06	0.60	0.06
Alt 4	0.67	0.20	1.04	0.67	0.20
Alt 5	0.60	0.06	4.14	0.67	0.20
Alt 6	0.73	0.52	0.34	0.73	0.52
Alt 7	0.73	0.52	9.04	0.80	1.00
Alt 8	0.67	0.20	162.5	0.87	0.74
Alt 9	0.73	0.52	336.9	1.00	0.09
Alt 10	0.73	0.52	490.6	1.00	0.09

Alternative Groupings. We now perturb the group assignments. As before, we group studies that examine the same outcome variable. We then randomly reorder the list of study indices, e.g., $1, 2 \ldots, 7$ for "Profits," to create new groupings. Table 4 contains $\nu = 0$ and $\nu = \hat{\nu}$ results for the benchmark grouping and ten alternative groupings. The alternative groupings are sorted based on the empirical coverage frequencies.

For alternative groupings 1, 2, 3, 4, and 6 $\hat{\nu}$ is close to zero in the sense that coverage frequencies and p-values are exactly identical to the $\nu = 0$ case because N is small. Although the coverage frequency is below 0.8, due to the small sample size two out of the five p-values exceed 0.05. Under these groupings the baseline studies indicate a large degree of external validity, but the resulting predictive distribution is too concentrated, either because standard errors in the baseline samples are too small, or the parameter estimate for the validation sample is very different. For the last three alternative groupings $\hat{\nu}$ is very large, which indicates that there is a large dispersion among the estimates in the baseline studies. In turn, the predictive intervals for $\theta_{i,J+1}$ are large. For groupings 9 and 10 this leads to coverage frequencies of one. Setting $\nu = 0$ for these groupings leads to coverage frequencies below 0.8, but the p-values are both 0.52.

6 Estimates of the Effects of Nudges

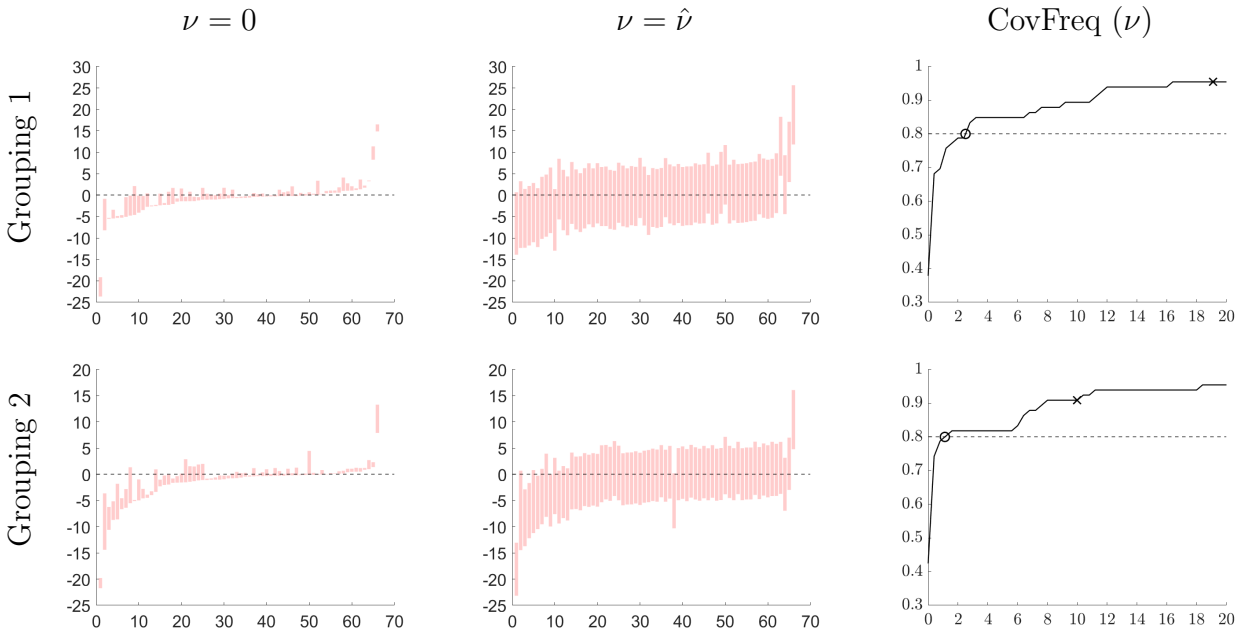
DellaVigna and Linos (2022) examine the scalability of behavioral interventions by RCTs run by two major U.S. government Nudge Units. We use their database of estimates from unpublished RCTs conducted by the Office of Evaluation Sciences (OES) and the Behavioral Insights Team's North America office (BIT-NA) to conduct our assessment. In each trial, there may be one control group and multiple treatment groups, each receiving a different form of nudge intervention (e.g., phone call, email message, letter). This generates a treatment effect estimate and standard error for each type of nudge. The outcome variable is typically the take-up rate, e.g., fraction of individuals donating blood, ensuring that the effects are measured on the same scale and are thus directly comparable across studies. From now on, we refer to each nudge treatment as a separate study, resulting in a total of 207 studies.

Despite the comparable scaling, the grouping of studies is more difficult in the previous two applications because the interpretation of the treatment effect depends on the outcome variable and the type of nudge. We consider two groupings. Under Grouping 1, we aim to assign studies with similar characteristics (policy area, communication medium, behavioral mechanism) to the same group. However, the targeted outcomes might differ. The grouping is generated by applying spectral clustering to binary vectors of characteristics. To obtain Grouping 2 we extract the trial titles associated with each study and generate semantic embeddings using the OpenAI API. This leads to numerical vectors to which we apply constrained k-means clustering. Grouping 2 tends to combine studies with similar outcomes, often drawn from the same trial, but the treatments within each group may vary considerably. We consider $J = 2$, which leads to $N = 66$ under Grouping 1 and $N = 69$ under Grouping 2. Further details are provided in the Online Appendix.

The results are summarized in Figure 10. In the left and center panels we plot the predictive intervals for $\hat{\theta}_{i,J+1}$ for $\nu = 0$ and $\nu = \hat{\nu}$. As before, we normalize $\hat{\theta}_{i,J+1}$ to zero. The most striking feature of the plots is that the predictive intervals at $\nu = \hat{\nu}$ are substantially wider than for $\nu = 0$, indicating that parameter variation across studies is one to two order of magnitude larger than sampling uncertainty. The median values of the ratio of $\sqrt{\hat{\nu}}$ to the group average standard error are around 15 and the maximum values close to 100.

The right panels in Figure 10 show the coverage frequency as a function of ν. For $\nu = 0$ the coverage frequencies are 0.38 and 0.42, respectively. As ν increases, the coverage frequency rise and intersect the 0.8 line at $\nu \approx 2$ (Grouping 1) and $\nu \approx 1$ (Grouping 2). At

Figure 10: Predictive Intervals and Coverage Frequencies for "RCTs to Scale"



Notes: The top row corresponds to Grouping 1, and the bottom row to Grouping 2. In the left and center columns, the groups in the bar charts are arranged in ascending order of the predictive intervals' lower bounds in the left column panels. The rightmost column displays the empirical coverage frequency as a function of ν. Circles indicate the values of ν that achieve an empirical coverage frequency of 0.8, while crosses represent the $\hat{\nu}$ estimates. $\hat{\nu} = 19.1$ under Grouping 1, and $\hat{\nu} = 10.0$ under Grouping 2.

$\nu = \hat{\nu}$ the predictive intervals are too wide, leading to coverage frequencies of 0.95 and 0.91, respectively. Three out of four p-values for the hypothesis that the coverage probability is 0.8 are 0, whereas the fourth one is 0.03. The general pattern is consistent with the previous applications: the reported standard error estimates are too small (or the underlying "true" parameters are too different) to achieve the desired nominal probability at $\nu = 0$. A value of ν that implies substantial variation of population parameters is needed to bring empirical and nominal coverage probabilities in alignment.

7 Conclusion

This paper began with a simple but fundamental question: Do economists provide meaningful and accurate quantifications of uncertainty? While econometricians dedicate significant effort to constructing standard errors and coverage intervals that are valid under hypotheti-

cal DGPs, we examine whether these uncertainty statements align with empirical frequencies observed across studies. Leveraging insights from the forecasting literature, we predict estimates from "new" studies based on those from baseline studies, linking them through a hierarchical model that captures variation in population parameters. This framework leads us to jointly assess the credibility of reported uncertainty in the baseline studies and the assumptions embedded in the hierarchical structure—specifically, the level of heterogeneity among true parameters required to reconcile empirical coverage frequencies with nominal ones.

In our three applications, we find that empirical coverage falls short of stated probabilities if one assumes the grouped studies estimate the same parameter. Achieving alignment often requires the variance across population parameters to substantially exceed the reported variance of the estimates themselves. This finding raises concerns about the practical value and scope of standard uncertainty quantification in applied economics. We are not arguing against reporting standard errors. Rather, we suggest that if the goal is to generalize findings beyond the studied sample, these measures should be adjusted — potentially inflated — by an appropriate "factor of safety." We hope our work motivates further scrutiny of reported uncertainty measures in economics and the social sciences, possibly through more sophisticated hierarchical modeling.

References

Aimone, J. A., S. Ball, E. Dwibedi, J. J. Jackson, and J. E. West (2024): "Macro-Level Institutions and Micro-Level Economic Behavior: A Meta-Meta Analysis of 1,126 Studies," *NBER Working Paper No. 33129*.

Andrews, I., and M. Kasy (2019): "Identification of and Correction for Publication Bias," *American Economic Review*, 109(8), 2766–2794.

Angelucci, M., D. Karlan, and J. Zinman (2015): "Microcredit Impacts: Evidence From a Randomized Microcredit Program Placement Experiment by Compartamos Banco," *American Economic Journal: Applied Economics*, 7(1), 151–182.

Armstrong, T., M. Kolesár, and Plagborg-Møller (2022): "Robust Empirical Bayes Confidence Intervals," *Econometrica*, 90(6), 2567–2602.

Attanasio, O., B. Augsburg, R. De Haas, E. Fitzsimons, and H. Harmgart (2015): "The Impacts of Microfinance: Evidence From Joint-liability Lending in Mongolia," *American Economic Journal: Applied Economics*, 7(1), 90–122.

Augsburg, B., R. De Haas, H. Harmgart, and C. Meghir (2015): "The Impacts of Microcredit: Evidence From Bosnia and Herzegovina," *American Economic Journal: Applied Economics*, 7(1), 183–203.

Banerjee, A., E. Duflo, R. Glennerster, and C. Kinnan (2015): "The Miracle of Microfinance? Evidence From a Randomized Evaluation," *American economic journal: Applied economics*, 7(1), 22–53.

Christensen, T., H. R. Moon, and F. Schorfheide (2023): "Optimal Decision Rules When Payoffs are Partially Identified," *Manuscript, University of Pennsylvania*.

Christoffersen, P. F. (1998): "Evaluating Interval Forecasts," *International Economic Review*, pp. 841–862.

Crépon, B., F. Devoto, E. Duflo, and W. Parienté (2015): "Estimating the Impact of Microcredit on Those Who Take It Up: Evidence From a Randomized Experiment in Morocco," *American Economic Journal: Applied Economics*, 7(1), 123–150.

Dawid, A. P. (1984): "Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach," *Journal of the Royal Statistical Society, Series A*, 147(2), 278–292.

DellaVigna, S., and E. Linos (2022): "RCTs to Scale: Comprehensive Evidence From Two Nudge Units," *Econometrica*, 90(1), 81–116.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998): "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, pp. 863–883.

Doksum, K. A., and A. Y. Lo (1990): "Consistent and Robust Bayes Procedures for Location Based on Partial Information," *Annals of Statistics*, 18(1), 443–453.

Ehrenbergerova, D., J. Bajzik, and T. Havranek (2023): "When Does Monetary Policy Sway House Prices? A Meta-Analysis," *IMF Economic Review*, 71, 538–573.

Gechter, M., and R. Meager (2022): "Combining Experimental and Observational Studies in Meta-Analysis: A Debiasing Approach," *Manuscript, Penn State University*.

Herbst, E., and F. Schorfheide (2015): *Bayesian Estimation of DSGE Models*. Princeton University Press.

Iacovone, L., D. J. McKenzie, and R. Meager (2023): "Bayesian Impact Evaluation with Informative Priors: An Application to a Colombian Management and Export Improvement Program," *World Bank Policy Research Working Paper No. 10274*.

Ishihara, T., and T. Kitagawa (2024): "Evidence Aggregation for Treatment Choice," *Manuscript arXiv:2108.06473v2*.

Karlan, D., and J. Zinman (2011): "Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation," *Science*, 332(6035), 1278–1284.

Kim, J.-Y. (2002): "Limited Information Likelihood and Bayesian Analysis," *Journal of Econometrics*, 107(1-2), 175–193.

Klein, R. A., K. A. Ratcliff, 48 others, and B. A. Nosek (2014): "Investigating Variation in Replicability: A "Many Labs" Replication Project," *Social Psychology*, 45(3), 142–152.

Liu, L. (2023): "Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective," *Journal of Business & Economics Statistics*, 41(2), 349–363.

Liu, L., H. R. Moon, and F. Schorfheide (2023): "Forecasting with a Panel Tobit Model," *Quantitative Economics*, 14(1), 117–159.

Manski, C. F. (2020): "Toward Credible Patient-centered Meta-analysis," *Epidemiology*, 31(3), 345–352.

McAlinn, K., and M. West (2019): "Dynamic Bayesian Predictive Synthesis in Time Series Forecasting," *Journal of Econometrics*, 210(1), 155–169.

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.

——— (2022): "Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature," *American Economic Review*, 112(6), 1818–1847.

Menkveld, A. J., A. Dreber, 340 others, and R. Zwinkels (2024): "Nonstandard Errors," *The Journal of Finance*, 79(3), 2339–2390.

Pratt, J. W., H. Raiffa, and R. Schlaifer (1965): *Introduction to Statistical Decision Theory*. Wiley, New York.

Rusnák, M., T. Havranek, and R. Horváth (2013): "How to Solve the Price Puzzle? A Meta-Analysis," *Journal of Money, Credit and Banking*, 45(1), 37–70.

Schmidli, H., S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, and B. Neuenschwander (2014): "Robust Meta-Analytic-Predictive Priors in Clinical Trials with Historical Control Information," *Biometrics*, 70, 1023–1032.

Schorfheide, F. (2008): "DSGE Model-Based Estimation of the New Keynesian Phillips Curve," *FRB Richmond Economic Quarterly*, 94(4), 397–433.

——— (2013): "Estimation and Evaluation of DSGE Models: Progress and Challenges," in *Advances in Economics and Econometrics*, ed. by D. Acemoglyu, M. Arellano, and E. Deckel, vol. 3, pp. 184–230. Cambridge University Press.

Smith, R. L., C. Tebaldi, D. Nychka, and L. O. Mearns (2009): "Bayesian Modeling of Uncertainty in Ensembles of Climate Models," *Journal of the American Statistical Association*, 104(485), 97–116.

Spiegelhalter, D. J. (2004): "Incorporating Bayesian Ideas into Health-Care Evaluation," *Statistical Science*, 19(1), 156–174.

Sugasawa, S., K. Takanashi, and K. McAlinn (2023): "Bayesian Causal Synthesis for Supra-Inference on Heterogeneous Treatment Effects," *arXiv:2304.07726v1*.

Sutton, A. J., and K. R. Abrams (2001): "Bayesian Methods in Meta-Analysis and Evidence Synthesis," *Statistical Methods in Medical Research*, 10, 277–303.

Tarozzi, A., J. Desai, and K. Johnson (2015): "The Impacts of Microcredit: Evidence From Ethiopia," *American Economic Journal: Applied Economics*, 7(1), 54–89.

Online Appendix: Uncertainty in Empirical Economics

Frank Schorfheide and Zhiheng You

This Appendix consists of the following sections:

A. Derivations for Section 3

B. Details and Additional Results for the Empirical Application in Section 4

C. Details and Additional Results for the Empirical Application in Section **??**

D. Details and Additional Results for the Empirical Application in Section **??**

A Derivations for Section 3

Marginal Data Density. The following calculation is conducted for a particular group i. Because of the independence assumptions, the MDD that covers all groups is given by the product of the i-specific MDDs. Inverting Bayes Theorem, yields

$$\begin{aligned} & p(\hat{\theta}_{i,1:J}|\nu,\xi) \qquad (A.1)\\ &= \frac{p\big(\hat{\theta}_{i,1:J}|\tau_i,\nu\big)p(\tau_i|\xi)}{p\big(\tau_i|\hat{\theta}_{i,1:J},\nu,\xi\big)} \\ &= (2\pi)^{-J/2}\left(\prod_{j=1}^{J}(\nu+\sigma_{ij}^2)^{-1/2}\right)\exp\left\{-\frac{1}{2}\sum_{j=1}^{J}\frac{\hat{\theta}_{ij}^2}{\nu+\sigma_{ij}^2}\right\}|\underline{V}_\tau|^{-1/2}\exp\left\{-\frac{1}{2}\underline{V}_\tau^{-1}\underline{\tau}^2\right\} \\ &\quad \times|\bar{V}_{\tau_i}|^{1/2}\exp\left\{\frac{1}{2}\bar{V}_{\tau_i}^{-1}\bar{\tau}_i^2\right\}. \end{aligned}$$

Now consider the $\underline{V}_\tau \longrightarrow \infty$ limit. Strictly speaking, the MDD is not well defined under this limit; but the limit of MDD ratios, say ν versus $\tilde{\nu}=0$ would be. Thus, we will ignore the terms stemming from $p(\tau_i|\xi)$ by setting the density to one. Under the improper prior posterior mean and variance are given by

$$\bar{V}_{\tau_i,\infty}=\left(\sum_{j=1}^{J}\frac{1}{\nu+\sigma_{ij}^2}\right)^{-1}, \quad \bar{\tau}_{i,\infty}=\bar{V}_{\tau_i,\infty}\left(\sum_{j=1}^{J}\frac{1}{\nu+\sigma_{ij}^2}\hat{\theta}_{ij}\right).$$

This leads to

$$\begin{aligned} & p_\infty(\hat{\theta}_{i,1:J}|\nu,\xi) \qquad (A.2)\\ &= (2\pi)^{-J/2}\left(\prod_{j=1}^{J}(\nu+\sigma_{ij}^2)^{-1/2}\right)\exp\left\{-\frac{1}{2}\sum_{j=1}^{J}\frac{\hat{\theta}_{ij}^2}{\nu+\sigma_{ij}^2}\right\} \\ &\quad \times\left(\sum_{j=1}^{J}\frac{1}{\nu+\sigma_{ij}^2}\right)^{-1/2}\exp\left\{\frac{1}{2}\left(\sum_{j=1}^{J}\frac{1}{\nu+\sigma_{ij}^2}\right)^{-1}\left(\sum_{j=1}^{J}\frac{1}{\nu+\sigma_{ij}^2}\hat{\theta}_{ij}\right)^2\right\}. \end{aligned}$$

B Details and Additional Results for the Empirical Analysis in Section 4

B.1 Benchmark Analysis

Construction of Study Estimates. We download the ManyLabs summary statistics spreadsheets from https://osf.io/dmf62. The 11 designs we consider are Sunk costs, Anchoring 1, Anchoring 2, Anchoring 3, Anchoring 4, Gambler's fallacy, Flag priming, Quote attribution, Money priming, Imagined contact, Math-art-gender, IAT correlation, Gain loss, Allowed forbidden, Reciprocity, and Scales. Each spreadsheet contains information on one single design. Figure A-1 illustrates an example of the "Flag priming" experiment. In each sheet, we use the columns labeled "N(T)," "N(C)," "Mean(T)," "Mean(C)," "SD(T)," and "SD(C)," where T represents the treatment group and C represents the control group. The specific labels for T and C may vary based on the design. For this "Flag priming" example, these correspond to columns B "N (Flag Prime)," C "N (Control)," E "Mean (Flag Prime)," F "Mean (Control)," G "SD (Flag Prime)," and H "SD (Control)". To ensure consistency in magnitude across estimates, we rescale columns "Mean(High)," "Mean(Low)," "SD(High)," and "SD(Low)" in "Anchoring 1," "Anchoring 2," "Anchoring 3," and "Anchoring 4" by a factor of 1/1000. We apply the formulas described in the main paper to compute the estimates and standard errors.

Figure A-1: "Many Labs" Data Example

	A	B	C	D	E	F	G	H
1	Site	N (Flag	N (Control)	N (Excluded	Mean (Flag	Mean (Control)	SD (Flag	SD (Control)
2	**Overall:**	3106	3145	93	3.10	3.07	1.02	1.00
3	**Overall for US**	2,424	2,472		3.17	3.15	1.07	1.04
4	**Mean across**	86.28	87.36		3.11	3.10	0.83	0.82
5	abington	39	44	1	3.10	3.05	0.61	0.90
6	brasilia	62	58	0	2.82	2.82	0.80	0.95
7	charles	51	33	0	3.07	3.09	0.65	0.68
8	conncoll	42	52	1	2.56	2.88	0.69	0.72
9	csun	47	47	2	3.26	3.32	0.77	0.85
10	help	45	57	0	3.17	3.25	0.85	0.54
11	ithaca	43	46	1	2.86	2.99	0.77	0.75
12	jmu	82	90	2	3.41	3.37	0.90	0.89
13	ku	54	56	3	2.69	2.59	0.76	0.60
14	laurier	58	53	1	2.90	2.87	0.63	0.71
15	lse	132	143	2	2.73	2.70	0.78	0.70
16	luc	78	67	1	3.08	3.03	0.97	0.75
17	mcdaniel	48	50	0	3.20	3.12	0.76	0.88
18	msvu	45	40	0	2.60	2.54	0.56	0.73
19	mturk	487	495	18	3.21	3.10	1.19	1.15
20	osu	35	62	10	3.59	3.35	0.86	0.90
21	oxy	58	63	2	2.30	2.30	0.71	0.72
22	pi	654	666	9	2.92	2.89	1.12	1.06
23	psu	46	45	4	3.39	3.37	0.73	0.78
24	qccuny	48	51	4	3.13	3.30	0.79	0.77
25	qccuny2	48	37	1	3.10	3.09	0.74	0.77
26	sdsu	86	74	2	3.06	3.15	0.77	0.78
27	swps	34	44	1	3.17	3.03	0.72	0.75
28	swpson	90	79	0	2.83	3.03	0.89	0.96
29	tamu	107	75	5	3.92	4.19	0.95	1.04

Flag Priming | Quote Attribution | Money Priming | Imagined Contact | Math_Art Gender | IAT correlation | Gain_Loss

Additional Tables and Figures.

- Table A-1 provides RCT site information for the experiments.
- The grouping for the benchmark analysis is summarized in Table A-2. The alternative grouping is described in Table A-3.
- Figure A-2 shows predictive intervals and coverage frequencies as function of ν for the reduced sample. It is similar to Figure 2 in the main text.
- Figure A-3 shows the ratio of $\sqrt{\hat{\nu}}$ to the average standard error for the reduced sample. It is similar to Figure 3 in the main text.
- Figure A-4 shows predictive intervals for the case of $\nu = 0$.
- Figure A-5 shows additional PIT histograms.

Table A-1: Site Information

Site Identifier	Location	Participants	Online (O) or Lab (L)	U.S. or International (I)
Abington	Penn State Abington, Abington, PA	84	L	U.S.
Brasilia	University of Brasilia, Brasilia, Brazil	120	L	I
Charles	Charles University, Prague, Czech Republic	84	L	I
Conncoll	Connecticut College, New London, CT	95	L	U.S.
CSUN	California State University, Northridge, LA, CA	96	O	U.S.
Help	HELP University, Malaysia	102	L	I
Ithaca	Ithaca College, Ithaca, NY	90	L	U.S.
JMU	James Madison University, Harrisonburg, VA	174	O	U.S.
KU	KoÅ University, Istanbul, Turkey	113	O	I
Laurier	Wilfrid Laurier University, Waterloo, Ontario, Canada	112	L	I
LSE	London School of Economics and Political Science, London, UK	277	L	I
Luc	Loyola University Chicago, Chicago, IL	146	L	U.S.
McDaniel	McDaniel College, Westminster, MD	98	O	U.S.
MSVU	Mount Saint Vincent University, Halifax, Nova Scotia, Canada	85	L	I
MTURK	Amazon Mechanical Turk (U.S. workers only)	1000	O	U.S.
OSU	Ohio State University, Columbus, OH	107	L	U.S.
Oxy	Occidental College, LA, CA	123	L	U.S.
PI	Project Implicit Volunteers (U.S. citizens/residents only)	1329	O	U.S.
PSU	Penn State University, University Park, PA	95	L	U.S.
QCCUNY	Queens College, City University of New York, NY	103	L	U.S.
QCCUNY2	Queens College, City University of New York, NY	86	L	U.S.
SDSU	SDSU, San Diego, CA	162	L	U.S.
SWPS	University of Social Sciences and Humanities Campus Sopot, Sopot, Poland	79	L	I
SWPSON	Volunteers visiting www.badania.net	169	O	I
TAMU	Texas A&M University, College Station, TX	187	L	U.S.
TAMUC	Texas A&M University-Commerce, Commerce, TX	87	L	U.S.
TAMUON	Texas A&M University, College Station, TX (Online participants)	225	O	U.S.
Tilburg	Tilburg University, Tilburg, Netherlands	80	L	I
UFL	University of Florida, Gainesville, FL	127	L	U.S.
UNIPD	University of Padua, Padua, Italy	144	O	I
UVA	University of Virginia, Charlottesville, VA	81	L	U.S.
VCU	VCU, Richmond, VA	108	L	U.S.
Wisc	University of Wisconsin-Madison, Madison, WI	96	L	U.S.
WKU	Western Kentucky University, Bowling Green, KY	103	L	U.S.
WL	Washington & Lee University, Lexington, VA	90	L	U.S.
WPI	Worcester Polytechnic Institute, Worcester, MA	87	L	U.S.

Table A-2: "Many Labs" Benchmark Grouping

Panel A: $J = 5$

Group	Study Sites					
1	Brasilia	Charles	Help	Laurier	MSVU	SWPS
2	KU	SWPSON	UNIPD	LSE	Tilburg	WPI
3	CSUN	JMU	McDaniel	MTURK	PI	TAMUON
4	WL	TAMU	Abington	QCCUNY	OSU	Luc
5	UVA	TAMUC	PSU	QCCUNY2	Wisc	SDSU
6	VCU	UFL	WKU	Ithaca	Conncoll	Oxy

Panel B: $J = 2$

Group	Study Sites			Group	Study Sites		
1	Brasilia	Charles	Help	7	WL	TAMU	Abington
2	Laurier	MSVU	SWPS	8	QCCUNY	OSU	Luc
3	KU	SWPSON	UNIPD	9	UVA	TAMUC	PSU
4	LSE	Tilburg	WPI	10	QCCUNY2	Wisc	SDSU
5	CSUN	JMU	McDaniel	11	VCU	UFL	WKU
6	MTURK	PI	TAMUON	12	Ithaca	Conncoll	Oxy

Notes: See Table A-1 for explanation of acronyms.

Table A-3: "Many Labs" Alternative Grouping

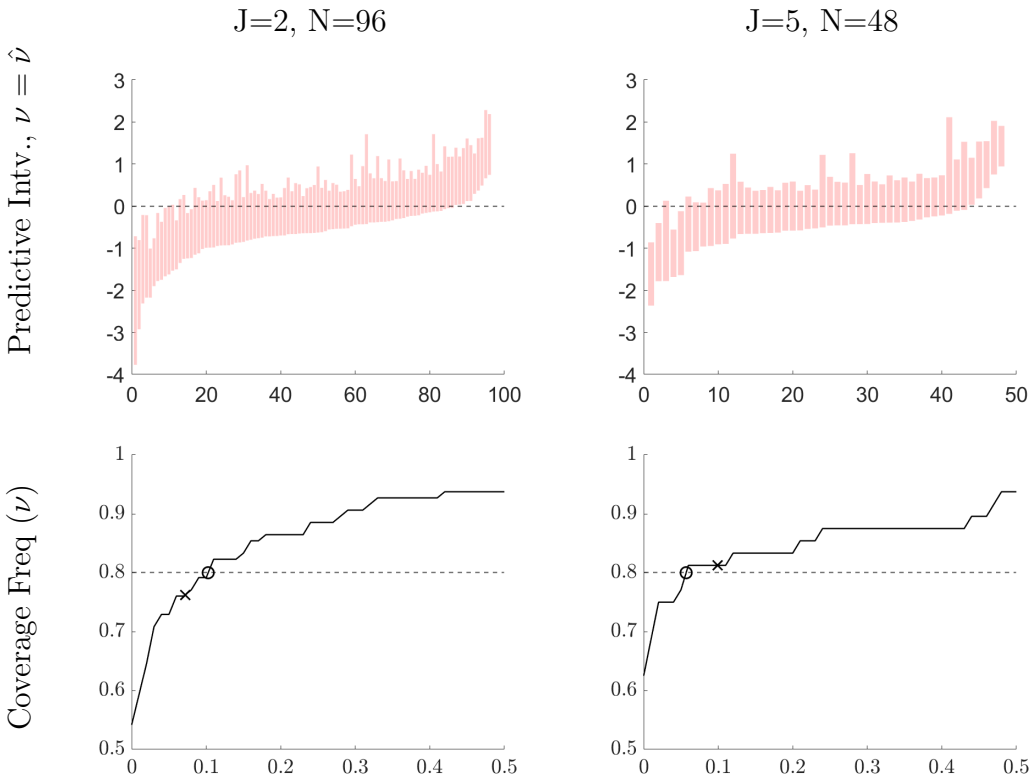
Panel A: $J = 5$

Group	Study Sites					
1	Brasilia	LSE	MSVU	Tilburg	Help	SWPS
2	KU	Charles	SWPSON	Laurier	UNIPD	Conncoll
3	CSUN	MTURK	JMU	PI	McDaniel	TAMUON
4	TAMU	TAMUC	UFL	UVA	VCU	WL
5	QCCUNY	QCCUNY2	Ithaca	Abington	PSU	WPI
6	Oxy	SDSU	Wisc	OSU	Luc	WKU

Panel B: $J = 2$

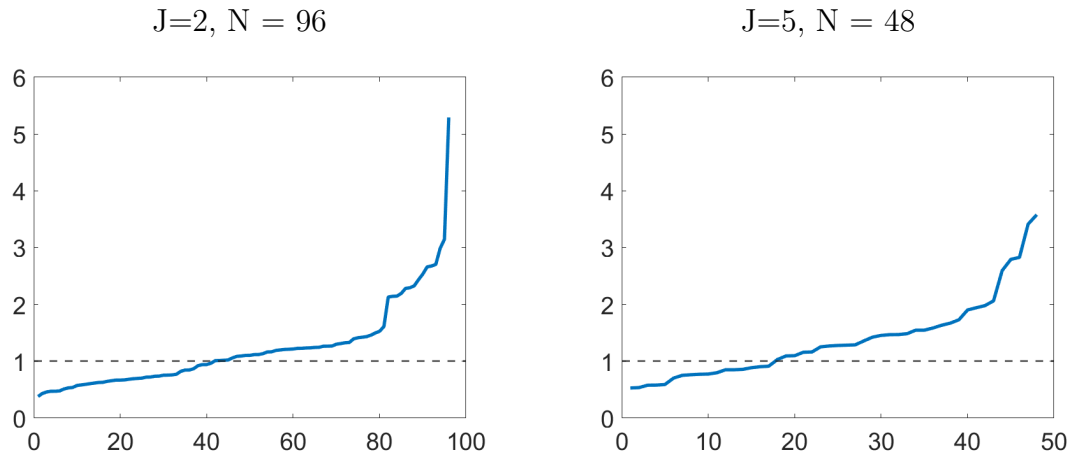
Group	Study Sites			Group	Study Sites		
1	Brasilia	LSE	MSVU	7	TAMU	TAMUC	UFL
2	Tilburg	Help	SWPS	8	UVA	VCU	WL
3	KU	Charles	SWPSON	9	QCCUNY	QCCUNY2	Ithaca
4	Laurier	UNIPD	Conncoll	10	Abington	PSU	WPI
5	CSUN	MTURK	JMU	11	Oxy	SDSU	Wisc
6	PI	McDaniel	TAMUON	12	OSU	Luc	WKU

Figure A-2: Predictive Intervals and Coverage for "Many Labs," Reduced Sample



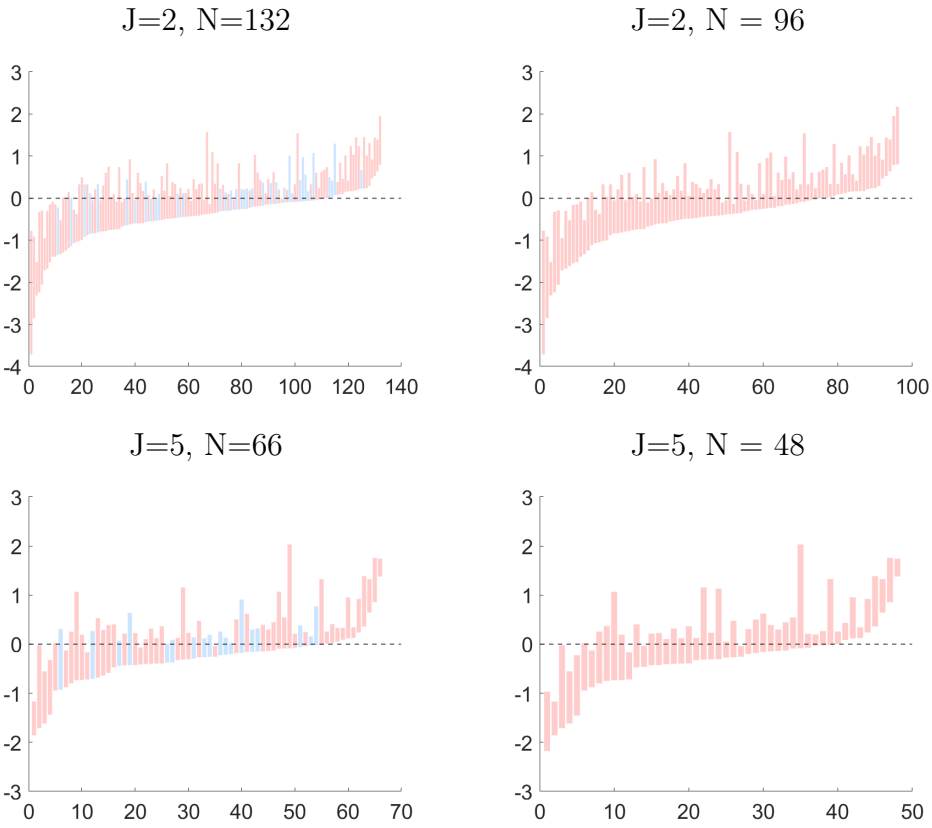
Notes: Top row: predictive intervals conditional on $\hat{\nu}$ (Empirical Bayes approach). The bars are arranged in ascending order of the predictive intervals' lower bounds. Bottom row: empirical coverage frequency as a function of ν. Circles indicate the values of ν that achieve an empirical coverage frequency of 0.8, while crosses represent the $\hat{\nu}$ estimates.

Figure A-3: Ratio of $\sqrt{\hat{\nu}}$ to Average Standard Error, Reduced Sample



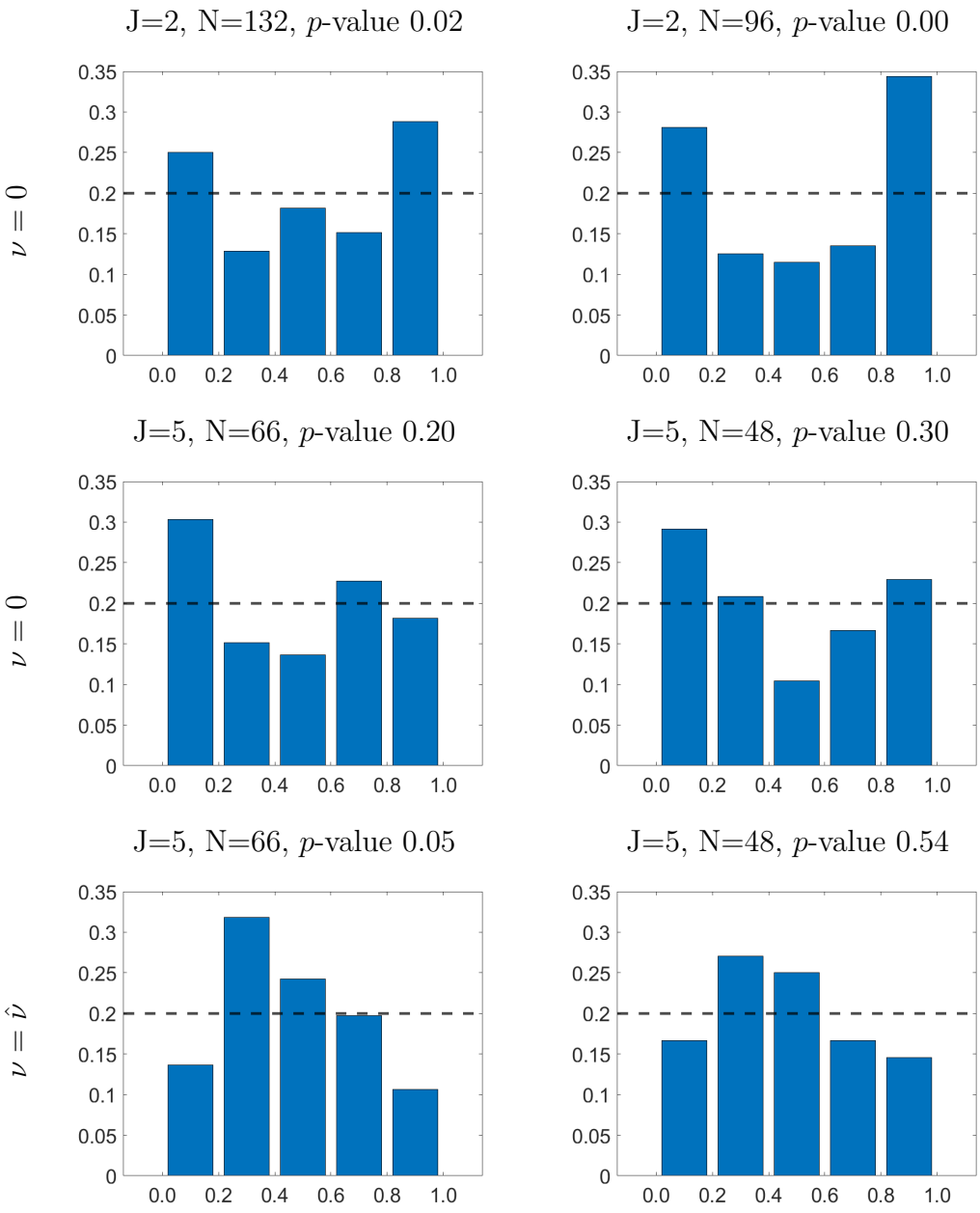
Notes: The figure depicts $r_i = \sqrt{\hat{\nu} / \frac{1}{J+1} \sum_{j=1}^{J+1} \sigma_{ij}^2}$ and groups i are sorted in ascending order of r_i.

Figure A-4: Predictive Intervals for "Many Labs," $\nu = 0$



Notes: Predictive intervals conditional on $\nu = 0$. The bars are arranged in ascending order of the predictive intervals' lower bounds. Blue bars represent studies with designs identified as weakly- or non-replicable by Klein, Ratcliff, 48 others, and Nosek (2014).

Figure A-5: Additional PIT Histograms for "Many Labs"



B.2 Robust Predictive Analysis

Recall that under the improper prior the posterior mean $\bar{\tau}_i$ can be expressed as

$$\bar{\tau}_i = \sum_{j=1}^{J} w_{ij}\hat{\theta}_{ij}, \quad w_{ij} = \left(\sum_{j=1}^{J} \frac{1}{\nu + \sigma_{ij}^2}\right)^{-1} \frac{1}{\nu + \sigma_{ij}^2}. \tag{A.3}$$

The distribution of B_i defined in the main text takes the from

$$B_i|(A_i, p_i) \sim N\big(0, V_{B_i}\big), \text{ where } V_{B_i} = \sigma_{i,J+1}^2 + \sum_{j=1}^{J} w_{ij}^2\sigma_{ij}^2. \tag{A.4}$$

V_{B_i} is known because it only depends on the weights w_{ij} and the variances of the estimators σ_{ij}^2. To construct a predictive interval robust to the normality assumption of $A_i|p_i$, we can let χ_i be largest possible $1-\alpha$ quantile of $|A_i+B_i|$, where $\mathbb{E}[A_i] = 0$, $\mathbb{E}[A_i^2] = \big(1+\Sigma_{j=1}^{J} w_{ij}^2\big)\nu := V_{A_i}$, and $B_i|A_i \sim N(0, V_{B_i})$.

Formally, the non-coverage probability of interval $[\bar{\tau}_i-\chi_i, \bar{\tau}_i+\chi_i]$, conditional on $\{\theta_{ij}\}_{j=1}^{J+1}$, i.e., holding A_i fixed, is

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{\theta}_{i,J+1} - \bar{\tau}_i\right| \geq \chi_i \;\middle|\; \{\theta_{ij}\}_{j=1}^{J+1}\right) \\
&= \mathbb{P}\left(|A_i + B_i| \geq \chi_i \;\middle|\; \{\theta_{ij}\}_{j=1}^{J+1}\right) \\
&= \mathbb{P}\left(\left|\frac{A_i}{\sqrt{V_{B_i}}} + Z\right| \geq \frac{\chi_i}{\sqrt{V_{B_i}}} \;\middle|\; \{\theta_{ij}\}_{j=1}^{J+1}\right) \\
&= \Phi_N\left(-\frac{\chi_i}{\sqrt{V_{B_i}}} - \frac{A_i}{\sqrt{V_{B_i}}}\right) + \Phi_N\left(-\frac{\chi_i}{\sqrt{V_{B_i}}} + \frac{A_i}{\sqrt{V_{B_i}}}\right) \\
&:= r\left(\frac{A_i}{\sqrt{V_{B_i}}}, \frac{\chi_i}{\sqrt{V_{B_i}}}\right),
\end{aligned} \tag{A.5}$$

where Z is a standard normal random variable, and Φ_N is its cdf. Thus, by iterated expectations, the non-coverage is bounded by

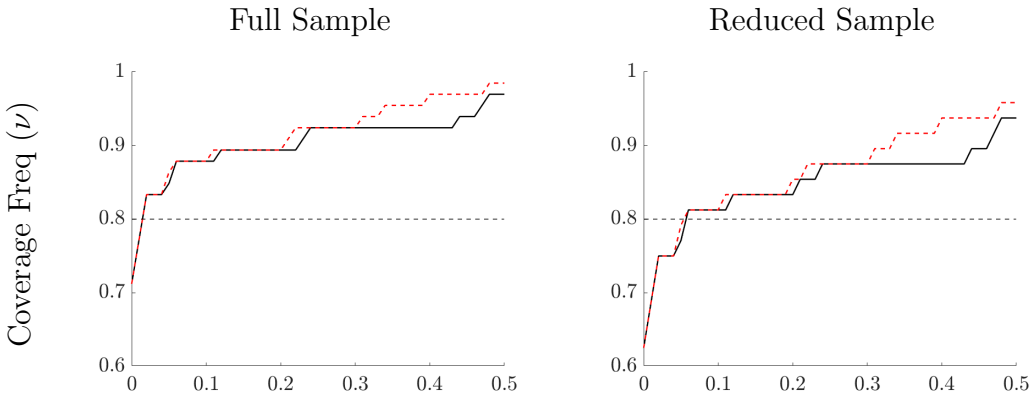
$$\rho\left(\frac{V_{A_i}}{V_{B_i}}, \frac{\chi_i}{\sqrt{V_{B_i}}}\right) = \sup_{p_i} \mathbb{E}\left[r\left(a, \frac{\chi_i}{\sqrt{V_{B_i}}}\right) \;\middle|\; p_i\right], \tag{A.6}$$

subject to

$$\mathbb{E}[a|p_i] = 0, \quad \mathbb{E}[a^2|p_i] = \frac{V_{A_i}}{V_{B_i}}.$$

The critical value χ_i can be expressed as $\chi_i = \sqrt{V_{B_i}} \cdot \text{cva}_\alpha\left(V_{A_i}/V_{B_i}\right)$, where $\text{cva}_\alpha(t) = \rho^{-1}(t, \alpha)$, and the inverse is with respect to the second argument. Armstrong, Kolesár, and Plagborg-Møller (2022) provide code to compute $\rho(\cdot)$ and $\rho^{-1}(\cdot)$.

Figure A-6: Robust Intervals: Coverage Freq. for $J = 5$



Notes: The solid black lines depict coverage as a function of ν under benchmark Normal assumption, dashed red lines are based on robust predictive intervals.

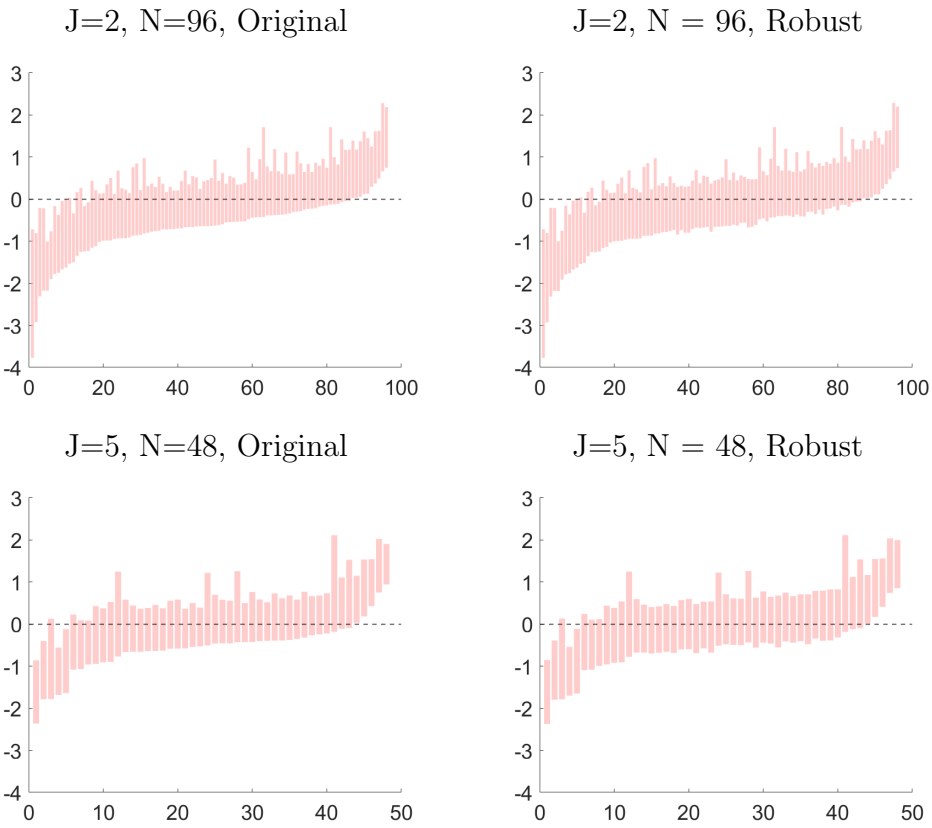
Figure A-6 shows the coverage frequency of the robust intervals as a function of ν for $J = 5$.

Figure A-7 shows that the robust predictive intervals are not substantially different from the original ones. To understand why the increase in coverage is relatively modest, note that the ratio of the robust to original predictive interval lengths can be expressed as a function of variance ratio $t_i := V_{A_i}/V_{B_i}$:

$$r(t_i) = \frac{\sqrt{V_{B_i}} \cdot \text{cva}_\alpha\left(V_{A_i}/V_{B_i}\right)}{z_{1-\frac{\alpha}{2}} \cdot \sqrt{V_{A_i} + V_{B_i}}} = \frac{\text{cva}_\alpha\left(t_i\right)}{z_{1-\frac{\alpha}{2}}} \cdot \frac{1}{\sqrt{t_i + 1}}. \tag{A.7}$$

Figure A-8 plots this function. Although $r(t_i)$ is increasing in t_i, the increase is modest—even when the variance ratio reaches 100, the interval length ratio rises only to 1.56. Figure A-9 shows histograms for the interval length ratio and variance ratio t_i across groups and ν values used to form the left panel of Figure A-6 ($N = 66, J = 5$).

Figure A-7: Predictive intervals for "Many Labs"



Notes: Predictive intervals conditional on $\nu = \hat{\nu}$ (Empirical Bayes approach). Top and bottom left panels: orignial predictive intervals under the normality assumption. The bars are arranged in ascending order of the predictive intervals' lower bounds. Top and bottom right panels: robust predictive intervals. The bars are arranged in the same order as in the left panels.

Figure A-8: Interval Length Ratio as a Function of t_i

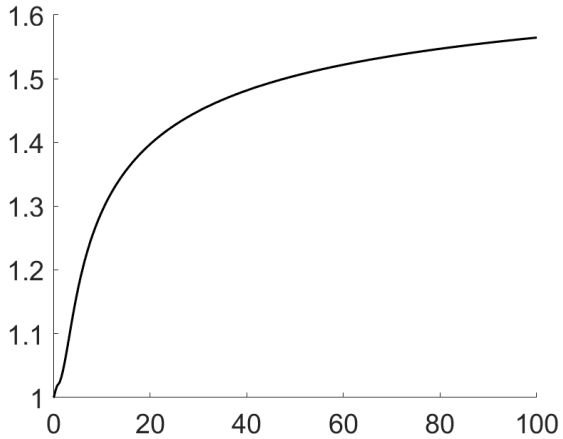
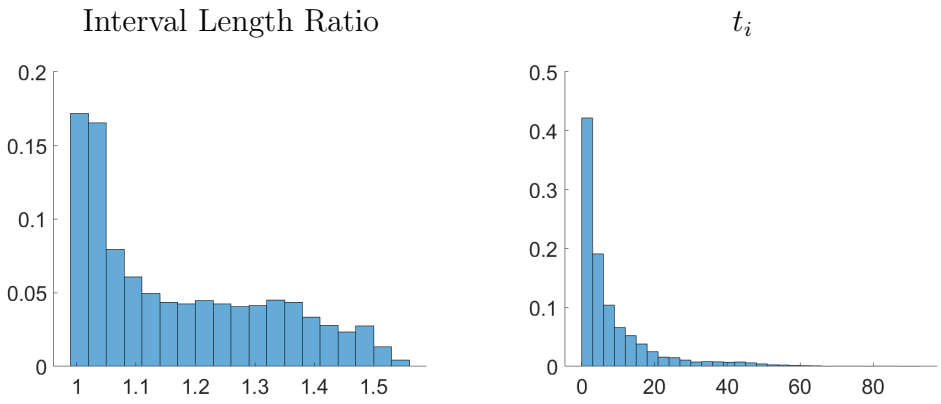


Figure A-9: Histograms for "Many Labs"



Notes: The y-axis shows the probability (relative frequency) of observations in each bin. Probabilities sum to 1. Left panel: the ratio of the robust interval length to the original interval length. Right panel: the ratio of V_{A_i} to V_{B_i}.

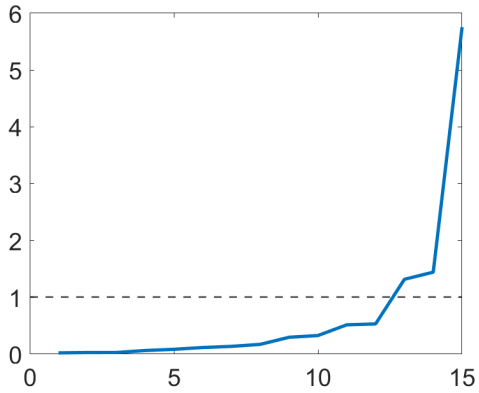
C Details and Additional Results for the Empirical Analysis in Section 5

Background. Meager (2019)'s analysis is based on seven randomized experiments: Angelucci, Karlan, and Zinman (2015), Attanasio, Augsburg, De Haas, Fitzsimons, and Harmgart (2015), Augsburg, De Haas, Harmgart, and Meghir (2015), Banerjee, Duflo, Glennerster, and Kinnan (2015), Crépon, Devoto, Duflo, and Parienté (2015), Karlan and Zinman (2011), and Tarozzi, Desai, and Johnson (2015).

Additional Tables and Figures.

- Figure A-10 shows the ratio of $\sqrt{\hat{\nu}}$ to the average standard error.

Figure A-10: Ratio of $\sqrt{\hat{\nu}}$ to Average Standard Error



Notes: The figure depicts $r_i = \sqrt{\hat{\nu}/\frac{1}{J+1}\sum_{j=1}^{J+1}\sigma_{ij}^2}$ and groups i are sorted in ascending order of r_i.

D Details and Additional Results for the Empirical Analysis in Section 6

The analysis draws on a comprehensive dataset of 126 RCTs conducted by two large Nudge Units — the Office of Evaluation Sciences (OES) and the Behavioral Insights Team's North America office (BIT-NA) — involving 23.5 million participants. The Nudge Unit data includes all eligible trials conducted between 2015–2019, most of which were unpublished and documented regardless of outcome. In general, these RCTs vary in policy area (e.g., revenue and debt, workforce and education, health), medium of communication (e.g., email, physical letter, in person), and behavioral mechanism (e.g., simplification and information, personal motivation, social cues).

In each trial, there may be one control group and multiple treatment groups, each receiving a different form of nudge intervention. For instance, in the trial titled "Increasing Use of Patient Generated Health Data through Patient Reminders," the control group receives no communication. The first treatment group receives a reminder signed by the patient's physician, emphasizing that the patient-entered data will be discussed at their next office visit ("Physician Accountability"), while the second treatment group receives a generic reminder signed by Inova Health System ("Basic"). Each nudge treatment corresponds to a treatment effect estimate and its associated standard error. The outcome variable is typically the take-up rate, ensuring that the effects are measured on the same scale and are thus directly comparable across studies. From now on, we refer to each nudge treatment as a separate study, resulting in a total of 207 studies.

Under Grouping 1, we aim to assign studies with similar characteristics to the same group. The grouping criteria is as follows: first, studies are grouped together only if they fall under the same policy area; second, within each policy area, studies that share similar communication mediums and behavioral mechanisms are more likely to be assigned to the same group. The procedure is implemented as follows. In the dataset, each study is described by a vector of binary characteristics (e.g., whether the mechanism involves framing, or whether the communication medium is email). Based on these vectors, we compute the Euclidean distances between studies within the same policy area to construct an adjacency matrix reflecting their similarity. Spectral clustering is then applied to this matrix to partition the studies into groups of $J+1$. [10]

[10]The standard spectral clustering algorithm only guarantees the number of resulting groups to be $\lfloor N_p/J+$

One drawback of Grouping 1 is that it may group together studies with entirely unrelated topics, even though they fall under the same broad policy area. For example, in one such group, one trial aims to increase enrollment in veteran health care benefits, while another focuses on encouraging the use of large item collection services.

Grouping 2 is designed to cluster studies with similar topics into the same group. To implement this, we first extract the trial title associated with each study. We then use the *text-embedding-ada-002* model provided by the OpenAI API to generate semantic embeddings of the titles, transforming the textual information into numerical vectors that capture their underlying meanings. Based on these vector representations, we apply constrained k-means clustering to partition the studies into groups of $J+1$, ensuring that grouping is guided by topical similarity.

Under Grouping 2, studies from the same trial are grouped together, and even when studies come from different trials, their topics are similar. For example, in one group, the titles of the three studies are: "Using Social Norms to Increase Payment of Delinquent Parking Citations in [REDACTED]", "Increasing Closure Rate of Parking Violation Cases in [REDACTED]", and "Using Social Norms to Improve Payment of Parking Fines in [REDACTED]". These studies are clearly centered around a common topic—encouraging payment of parking-related fines—highlighting the topical coherence achieved under Grouping 2.

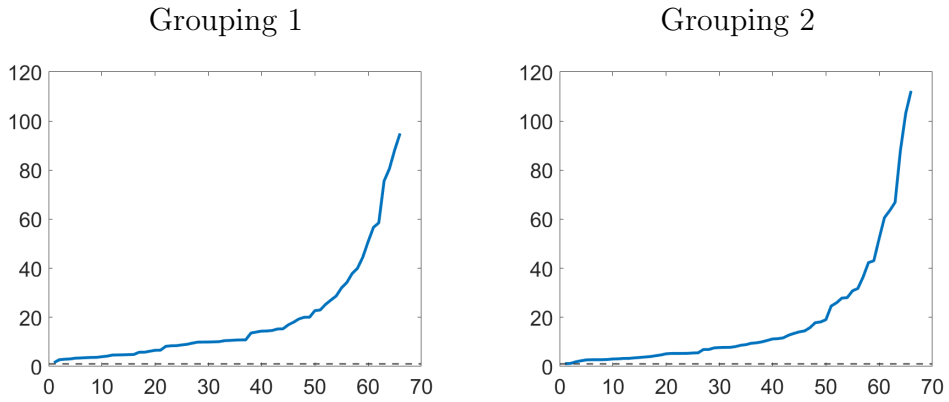
Additional Tables and Figures.

- Figure A-11 compares the estimated degree of external validity $\hat{\nu}$ to the group average of study standard errors. Specifically, for each group i, we calculate the ratio $r_i = \frac{\sqrt{\hat{\nu}}}{\frac{1}{J+1}\sum_{j=1}^{J+1}\sigma_{ij}}$.
- Table A-4 summarizes the coverage frequencies and reports the associated p-values. Under both groupings, the coverage frequencies are well below the nominal coverage probability of 0.8 when $\nu = 0$. Once ν is estimated, both coverage frequencies are well above 0.8.
- Figure A-12 shows PIT histograms. For each panel we also report the p-value associated with the S statistic. Under both groupings, the intervals seem to be too narrow when

$1\rfloor$, where N_p is the number of studies within each policy area. As a result, after applying spectral clustering, studies from any group with more than $J+1$ members are randomly reassigned to groups with fewer than $J+1$ members.

$\nu = 0$, but the intervals become too large when ν is estimated by empirical Bayes. Three out of four p-values are 0.00.

Figure A-11: Ratio of $\sqrt{\hat{\nu}}$ to Average Standard Error for "RCTs to Scale"

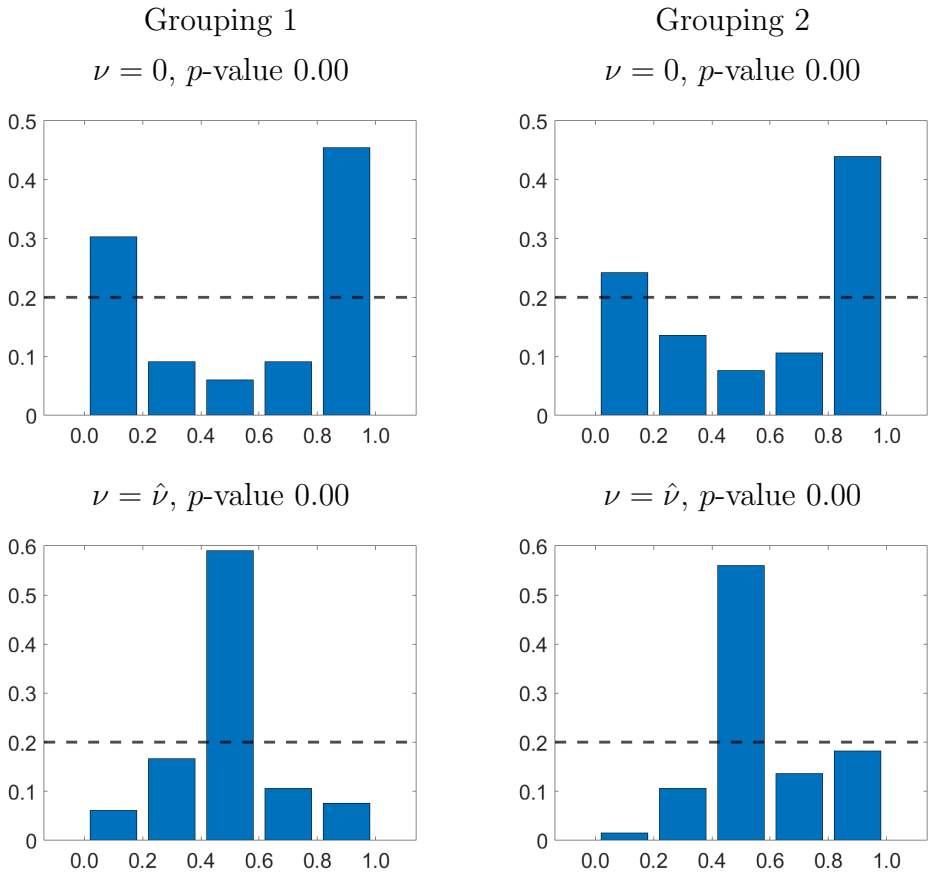


Notes: The figure depicts $r_i = \sqrt{\hat{\nu} / \frac{1}{J+1} \sum_{j=1}^{J+1} \sigma_{ij}^2}$ and groups i are sorted in ascending order of r_i.

Table A-4: Emp. Coverage Freq. for "RCTs to Scale"

	Grouping 1		Grouping 2	
	$\nu = \hat{\nu}$	$\nu = 0$	$\nu = \hat{\nu}$	$\nu = 0$
Emp. Coverage Freq.	0.95	0.38	0.91	0.42
Coverage p-value	0.00	0.00	0.03	0.00

Figure A-12: PIT Histograms for "RCTs to Scale"



Notes: The left column corresponds to Grouping 1, and the right column to Grouping 2. The p-value is computed for S statistics $S = \sum_{j=1}^{5} \frac{(n_j - N/5)^2}{N/5}$, where n_j is the number of PITs in the bin $[(j-1)/5, j/5]$. It is then derived based on the finite sample distribution of S, obtained via simulation.