NET WORK LISER

Discussion Paper Series

IZA DP No. 18701

May 2026

Characterizing the File Drawer: Evidence from a Meta-Analysis of Parent-Interventions Around the World

Peter Bergman
University of Texas, Austin
and IZA@LISER

Nat Chowanajin
University of Texas, Austin

IZA DP No. 18701 May 2026

Characterizing the File Drawer: Evidence from a Meta-Analysis of Parent-Interventions Around the World

Abstract

We conduct a meta-analysis of 82 randomized controlled trials across more than 20 countries to estimate the effects of low-cost, remote parental engagement interventions delivered through text messages, phone calls, and apps. We estimate a joint likelihood function that incorporates both written studies and unwritten studies identified through trial registries, funder records, research labs, evidence clearinghouses, and other sources. By also recording sample sizes for unwritten studies, the model estimates the distribution of standard errors, identifies write-up probabilities conditional on significance, and characterizes the file drawer by estimating effect distributions for written and unwritten studies. Bias-corrected effects are 0.05 SD for test scores, 0.07 SD for grades, 0.05 SD for attendance, and 0.03 SD for enrollment. In the best-identified domain, test scores, statistically insignificant results are still written up at high rates. We also find that larger studies tend to estimate smaller latent effects, which could indicate that true effects are correlated with study precision, violating a common meta-analysis assumption. In smaller-sample domains, our approach helps identify selection probabilities by anchoring the absolute write-up rates. Finally, we estimate the value of additional RCTs to inform adoption decisions. Any single study estimate is unlikely to dissuade adoption because parent interventions have high marginal value of public funds. Instead, future research is most valuable when it can explain heterogeneity across settings.

JEL classification

I24

Keywords

meta-analysis, parent engagement, randomized trials

Corresponding author

Peter Bergman
peterbergman@utexas.edu

1 Introduction

Despite decades of progress in school enrollment, learning levels remain at crisis levels: 70 percent of ten-year-olds in low- and middle-income countries cannot read a simple text (World Bank, 2022). A central challenge is many high-return learning inputs occur *outside* the classroom. Parents make daily decisions that are consequential for skill formation, yet their investment decision faces substantial frictions.[1] Parents have incomplete information about their children's ability, effort, and effective at-home practices (Bergman, 2019; Dizon-Ross, 2019). These frictions are acute for disadvantaged families, where school-to-parent communication is poor and cognitive bandwidth constraints are most severe (Bergman, 2020; Mani et al., 2013; Mullainathan and Shafir, 2013).

Over the last 15 years, a new class of interventions has emerged to target these frictions: remote programs delivered through text messages, phone calls, or mobile apps to engage parents in their child's education. The low cost and remote delivery of these programs means they could be among the highest return, readily scalable education interventions available to policymakers. Public and private entities have expanded these programs to reach millions of families around the world, which has coincided with dozens of randomized controlled trials (RCTs) studying these programs. This body of research creates an opportunity to aggregate the evidence and inform policy makers about its value.

The same features that make these interventions attractive for policy, however, can make it difficult to synthesize this evidence. First, low-cost interventions could imply more RCTs, many of which may never become journal articles, working papers, or policy reports. This creates a file-drawer problem: researchers may never write up statistically insignificant results (Brodeur et al., 2023; Franco et al., 2014). Second, the standard meta-analysis assumption that latent effects are independent of study precision is plausibly violated for RCTs because sample sizes are often *chosen* via power calculations that reflect researchers' expectations about effect sizes. If meta-analyses do not address these issues, their findings could overstate effects, understate uncertainty, and mischaracterize the range of contexts in which programs have been tried.

To overcome these challenges, we take advantage of a unique feature of randomized trials—their pre-registration and documentation in grant reports—to estimate a joint likelihood function that incorporates written and documented, *unwritten* into a publication-bias correction. We culled relevant studies from six RCT registries, 16 funder grant databases, seven evidence clearinghouses, 12 research databases, and Google Scholar. We also conducted forward citation searches, mined acknowledgment sections

[1] See Bergman (2021); Cunha and Heckman (2007); Todd and Wolpin (2007) for example.

for additional relevant funders, and emailed to researchers. Our final sample comprises 82 studies across more than 20 countries, 24 of which are documented but unwritten. The RCTs span four outcome domains: test scores, grades, attendance, and enrollment.

Methodologically, we build on Andrews and Kasy (2019) to operationalize a joint likelihood approach that incorporates documented but unwritten studies. The model combines a random-effect distribution of latent effects within each outcome domain with a selection function that depends on the absolute value of the study's z-statistic. Unwritten studies generally do not report standard errors, but nearly all report sample sizes, which allows us to model the distribution of standard errors as a function of sample size.

This modeling approach is helpful in several ways. First, we can characterize the studies in the file drawer. We estimate the model-implied distribution of effects for written and unwritten studies. We can assess whether the file drawer contains many null or even negative findings. Second, we can identify how write-up rates vary by statistical significance. This distinction is meaningful as studies may not be written due to imprecision or insignificant results (significance-driven selection), but even large-scale or otherwise precise studies may fail to be written because of implementation or funding constraints. Third, we relax the latent effect-precision independence assumption common in meta-analyses and test whether it is violated by modeling the relationship between sample size, standard errors, and latent effects. Lastly, documenting unwritten studies could help identify write-up rates and bias corrections, particularly in domains with small sample sizes.

We find that corrected mean effects are positive for all outcome domains, though enrollment is imprecisely estimated. In our preferred model, the estimated mean effects are 0.05 standard deviations (SD) for test scores and attendance, 0.07 SD for grades, and 0.03 SD for enrollment. Results for test scores, attendance, and grades are statistically significant. However, enrollment outcome is less precise because of the small number of studies and weak identification in that domain.

These effect sizes are meaningful given their low costs. For comparison, a meta-analysis of large-scale implementations of high-dosage tutoring finds an average impact of 0.16 SD, but at costs hundreds of times larger than the typical intervention in our sample (Kraft et al., 2024).[2] Another point of comparison is work by Jackson and Mackevicius (2024), who conduct a meta-analysis of school spending effects. The authors find that a $1,000 per pupil increase in spending sustained for four years improves test scores by 0.03 SD.

We also document that effects *are* negatively correlated with sample size. This suggests the common random-effect meta-analysis assumption that effects are not correlated with precision does not hold. However, relaxing this assumption by modeling the rela-

[2]This 0.16 SD impact is not directly comparable, as it does not adjust for publication bias in the same way, though the authors examine whether publication bias appears likely in their setting.

tionship between sample sizes and latent effects does not substantively change our results.

We formulate three results for policymakers. First, if a new program were implemented, what is the probability that it will have a positive effect? These probabilities are 84% for test scores, 98% for attendance, 99% for grades, and 80% for enrollment.[3] The relatively lower probability for test scores reflects a higher-degree of heterogeneity in effects across studies. We can explain very little of this heterogeneity using a coarse set of study covariates, including geography, student age, and intervention type. This could reflect how program design and implementation matter (Angrist and Meager, 2023; Meager, 2019; Vivalt, 2020).

Second, at a financial break-even threshold, the programs are inexpensive enough that even very small effects can be cost-justified (under strong assumptions on the relationship between earnings and test scores). Programs generally cost only a few dollars per student, so the break-even threshold is close to zero. Our bias-corrected estimates exceed this bar for test scores, attendance, and grades. We show these programs have high marginal values of public funds because their costs are so low (Hendren and Sprung-Keyser, 2020).

Lastly, we write down a program-adoption decision model to measure the value to policymakers of resolving uncertainty around study estimates, where uncertainty reflects both estimation error in the mean and cross-study heterogeneity. The expected value of perfect information is highest when heterogeneity is large and the evidence leaves meaningful uncertainty around the decision threshold. However, when unexplained heterogeneity is large, one additional RCT captures little of this value. This implies that future research may be most valuable when it can explain *why* impacts vary across contexts, rather than simply adding one more randomized trial estimate.

We also find evidence of publication bias and selection-driven write-up rates. In our best-identified domain, test scores, the estimated probability that a statistically insignificant result is written up is 75 percent, compared with 87 percent for a statistically significant result. In smaller-sample domains, these rates are 72 percent versus 99 percent for grades, 74 versus 99 percent for attendance, and 60 percent versus 99 percent for enrollment. Using written studies alone, we can estimate the *ratio* of these probabilities only, but the latter appear weakly identified given the small samples. While the two approaches estimate similar ratios for the larger-sample test score domain, estimates using only written studies imply orders of magnitude more severe selection in the smaller-sample domains. We can also mechanically increase the share of missing studies to examine the robustness our findings about mean effects. Our results are robust to reasonable changes in missingness, and the best-identified domain, test scores, is especially stable.

[3]These probabilities assume, however, that the policymaker is drawing from the full distribution of program designs and implementation contexts represented in the literature.

Our work relates to a large body of syntheses cataloging the effectiveness of education interventions.[4] Angrist et al. (2025) compare over 200 education policies using a unified learning-adjusted years of schooling metric, and Evans and Yuan (2022) benchmark effect sizes across 234 studies, documenting a median learning effect of 0.10 SD. Other papers focus on specific intervention classes: Kraft et al. (2024) and Nickow et al. (2024) analyze tutoring programs, and Escueta et al. (2020) review education technology RCTs. Parental engagement interventions have received comparatively little attention in these syntheses, though Snilstveit et al. (2017) classified information-to-parents as having too few high-quality studies to draw conclusions, which partly motivates our focus on this specific intervention class.

This paper also contributes to a literature that models publication bias as a selection problem in which the probability of observing a result depends on its statistical significance (Copas and Jackson, 2004; Iyengar and Greenhouse, 1988; Hedges, 1992; Vevea and Hedges, 1995). Most important to our work, Andrews and Kasy (2019) show that publication bias can be addressed non-parametrically from written studies alone. The authors also note that when the existence of unpublished studies is known, the problem becomes one of censoring and the absolute level of the selection function is identified. Huang et al. (2021, 2023) incorporate clinical trial registry data into publication-bias corrections using sample size or precision. Other papers document publication bias by assembling comprehensive samples of conducted studies (Brodeur et al., 2020; DellaVigna and Linos, 2022; Driessen et al., 2015; Franco et al., 2014; Leight et al., 2025). Our setting arises when unwritten studies and their sample sizes can be documented, but the studies' estimates are unobserved. The joint likelihood approach developed here is well-suited to such settings and could be applied to a broad topics where RCTs are prevalent.

The rest of the paper proceeds as follows. Section 2 describes the intervention class, our inclusion criteria, and the five-stage audit procedure. Section 3 presents the statistical model and estimation strategy. Section 4 reports bias-corrected estimates, the estimated selection function, prediction intervals, heterogeneity analysis, and benchmarking against alternative interventions. Section 6 presents robustness and validity checks. Section 7 discusses the value of additional research and concluding implications for policy.

2 Background, Data, Search

Remote-parent interventions often target several frictions in the household production of human capital. While parental time and resource investments are key inputs for skill

[4]See Conn (2017); Evans and Popova (2016); Glewwe and Muralidharan (2016); Jackson and Mackevicius (2024); Kremer et al. (2013); McEwan (2015); Snilstveit et al. (2017) for example.

formation (Cunha and Heckman, 2007; Todd and Wolpin, 2007), parents often lack the information needed to allocate these inputs efficiently. Evidence shows that parents hold systematically inaccurate beliefs about their child's academic performance, often overestimating effort and achievement. Correcting these beliefs changes both parental behavior and student outcomes (Barrera-Osorio et al., 2020; Bergman, 2021; Bettinger et al., 2021; Dizon-Ross, 2019). The interventions in our sample address these frictions in two ways.

The first way is *informational*, which aims to reduce information asymmetries between parents, children and schools. The programs in this group provide data to parents on their children's academic progress and effort, which can help parents monitor and motivate their children. This follows a principal-agent logic in which better-informed parents can induce higher effort from their children (Bergman, 2021), but can also operate through salience (Bettinger et al., 2021).

The second way is *instructional*. These programs aim to increase the productivity of parents' time investments by guiding parents on specific activities to do with their children, such as reading exercises and working jointly with remote instructors (Angrist et al., 2022; Mayer et al., 2019; York et al., 2019).

Both forms of interventions are delivered remotely to parents, most commonly through text messages, report cards, or phone calls. This method keeps marginal costs below $10 per student and makes the implementation feasible at scale without requiring significant changes to school operations.

2.1 Inclusion Criteria and Search

To be included, a study had to: (a) use a Randomized Controlled Trial (RCT) to estimate the impact of parent-engagement interventions; (b) provide information or instruction to parents or caregivers rather than to students alone; (c) focus on low-cost, remotely delivered interventions that are easily scalable; and, (d) estimate one or more of the following outcomes: (i) Test scores (standardized test scores or proficiency rates); (ii) Grades (GPA, assignment completion, or failed classes); (iii) Attendance (absenteeism, number of absent days, or attendance rates); and (iv) Enrollment (dropout rates, graduation rates, or enrollment rates).

We identified 58 academic papers with 123 outcomes that met our conditions as of December 2025.[5] To do so we searched six RCT registries, 16 funder grant (both governments and private foundations) databases, seven evidence clearinghouses, and 12 research

[5]See examples for excluded studies in Appendix Section A.

databases.[6] We also conducted forward citation searches, scanned acknowledgement sections of the included papers for additional funders and authors, and searched Google Scholar.

We started from Bergman (2021) and those cited in the paper. Each paper then became a seed in *Connected Papers* and we manually verified if it met the criteria. We iterated the process until we can no longer find another relevant paper. We supplemented this process by identifying the list of funded projects by major research funders, such as Abdul Latif Jameel Poverty Action Lab (JPAL), Arnold Ventures, and IES, and emails to authors inquiring about unwritten studies.

We define an *unwritten* paper as one for which we observe evidence that an RCT was to be conducted but for which no written output, such as a working paper, technical report, or peer-reviewed publication, was available. Drawing on pre-analysis plans, study proposals, and author replies, we collected the information on targeted sample sizes and demographics, research questions, treatment, and primary outcomes. The pool of unwritten studies was narrowed down using the same inclusion criteria mentioned above. In the final sample, there were 24 unwritten projects, 21 of which reported a sample size.

2.2 Data Collected from Each Study

After manually verifying that each project met our inclusion criteria (see Section 2.1), we collected available information on whether the intervention was instructional (i.e. aiming to teach) or informational (e.g. providing information about their child's academic performance), whether the intervention targeted younger or older children (elementary school children or younger versus not), the geographic context (e.g., OECD versus non-OECD country), the outcomes of interest and whether it was primary or not, the cost of the intervention, and the intended sample size. Outcomes were standardized to be in a standard deviation unit. When applicable and feasible, we contacted authors directly for needed details if information was missing. For unwritten studies, we obtained this information from grant reports, the RCT analysis plans, or other available documentation.

[6]List of funders: International Initiative for Impact Evaluation (3ie), RAND Corporation, Arnold Ventures, Mathematica, Abdul Latif Jameel Poverty Action Lab (JPAL), Institute of Education Sciences (IES), International Growth Centre (IGC), Education Endowment Foundation (EEF), and American Institutes for Research (AIR), and Innovations for Poverty Action (IPA). List of project registries: AEA RCT Registry, Registry of Efficacy and Effectiveness Studies (REES), and Center for Open Science (OSF). Many researchers submitted a pre-analysis plan to funders or RCT registries, which allowed us to extract such information.

2.3 Descriptive Overview

Table 1 summarizes basic statistics about the included studies by each outcome domain.[7] Columns 2 and 3 indicate the number of studies in each domain. The raw write-up rates range between 67% and 84%. One study can estimate effects several different measures of outcomes within the same domain, so we report the total number of estimates in Columns 5 and 6. The last two columns indicate the median sample size of each domain.[8]

The written papers vary widely in terms location, as shown in Figure 1a. Red indicates studies conducted in OECD member countries. For the United States, blue indicates the states in which studies were located. Green indicates studies in non-OECD countries, which are lower-income country in this context. The geographic coverage of unwritten papers largely follows that of written ones, with a few additional countries, as shown in Figure 1b.

Figure 2 presents the number of studies by age group the study targeted and the write-up status. Most papers focused on younger students, including those in pre-school, kindergarten, and elementary school. The patterns are consistent between studies in lower and higher-income countries.

In line with Banerjee et al. (2016b) and the general rise of the "credibility revolution" (cf. Angrist and Pischke (2010)), all the studies were conducted in the 2000s or later as depicted in Figure 3. Although we identify the first intervention in 2004, most studies were conducted in the 2010s. Approximately half of the studies were conducted in higher income countries, predominantly in the United States.

Figure 4 plots the number of written and unwritten projects by their registration year. The green line shows the cumulative percentage of unwritten papers over year. The gap between registrations and written papers widens after 2015, consistent with a file-drawer problem as well as a lag between project registration and write-up completion. Because this time lag can be endogenous, we do not exclude studies any time before December of 2025, which we view as conservative, though we can allow for a time trend in our analysis to account for this lag.

[7] Appendix tables list the studies and estimates included in each outcome domain.

[8] These numbers refer to the recruited samples (e.g., number of students) to keep it comparable to the analysis plan for the unwritten studies.

3 Identification and Estimation

In this section, we construct a joint likelihood of publication and selection model by incorporating both written and unwritten studies. By extracting information from studies that were conducted but never written up we can identify the absolute probabilities that a study is written conditional on statistical significance or insignificance.

3.1 Set up

Within each domain d we assume the average latent effect follows a normal distribution with mean μ_d and heterogeneity across studies is captured by the standard deviation τ_d. Each paper i in the domain then has the true effect Θ_{id} drawn from $\mathcal{N}(\mu_d, \tau_d^2)$. Conditional on the true effect, Θ_{id}, and σ_{id}, we assume the observed result X_{id} is drawn from the normal distribution $\mathcal{N}(\Theta_{id}, \sigma_{id}^2)$.[9]

A standard assumption in meta-analyses that latent effects are independent of precision: $\Theta_{id} \perp \sigma_{id}$ is likely violated. We can test this assumption, and, by modeling the relationship between sample size and effect size, relax it if our model is correct.

The setup can be summarized as follows.[10]

$$\Theta \sim \mathcal{N}(\mu_d, \tau_d^2), \tag{1}$$

$$x \mid \Theta, \sigma \sim \mathcal{N}(\Theta, \sigma^2), \tag{2}$$

$$x \mid \sigma \sim \mathcal{N}\big(\mu_d, \tau_d^2 + \sigma^2\big) \tag{3}$$

For unwritten studies we observe neither standard errors nor study estimates. We model standard errors σ as a function of sample size n and other study characteristics under the assumption that conditional on sample size, standard errors are orthogonal to the write-up status D. We model this relationship as a log-normal distribution, where a and b are a domain-specific intercept and elasticity, respectively, and c captures how other study characteristics relate to σ.

$$\log \sigma \mid n \sim \mathcal{N}\big(a + b \log n + cW_{id},\ s^2\big) \tag{4}$$

A paper is written up if $D = 1$, which occurs with probability $p(z)$ where $z = x/\sigma$.

[9] This is arguably reasonable assumption given that estimates are often assumed to be distributed asymptotically normal.

[10] For simplicity, at times, we suppress the subscripts i and d.

We parametrize this probability with a symmetric step function with one cutoff at 1.96.[11]

$$p(z) = \begin{cases} q_\ell, & |z| < 1.96, \\ q_h, & |z| \geq 1.96 \end{cases} \tag{5}$$

For unwritten studies, the test statistic z is not observed. From Equation (3) we have

$$z \mid \sigma \sim \mathcal{N}\left(\frac{\mu_d}{\sigma},\ 1 + \frac{\tau_d^2}{\sigma^2}\right),$$

The probability that a study is statistically insignificant at a critical value c, conditional on its standard error, is

$$P_\ell(\sigma) \equiv \Pr(|z| < c \mid \sigma). \tag{6}$$

Given the selection function in Equation (5), the write-up probability conditional on σ is

$$\pi_d(\sigma) \equiv \Pr(D = 1 \mid \sigma) = q_\ell\, P_\ell(\sigma) + q_h \left(1 - P_\ell(\sigma)\right). \tag{7}$$

3.2 Likelihood Function

Let $f_N(\cdot; \mu, v)$ denote the normal density function with mean μ and variance v. Let $f_{LN}(\cdot; m, s^2)$ denote the lognormal density such that $\log \sigma \sim \mathcal{N}(m, s^2)$.

For written studies, denoted by $D = 1$, the likelihood contribution is characterized by three components: (1) precision σ given its sample size n and study characteristics W; (2) outcome x given the precision σ; and (3) the selection given $z \equiv x/\sigma$. Thus, for each written paper, the likelihood function is:

$$L^{(1)} = \underbrace{f_{LN}\left(\sigma; a + b\log n + cW, s^2\right)}_{\text{precision model}} \times \underbrace{f_N\left(x; \mu, \tau^2 + \sigma^2\right)}_{\text{effect density}} \times \underbrace{p_d(z)}_{\text{selection function}} \tag{8}$$

For unwritten studies ($D = 0$) we integrate over the distribution of σ. The probability of *not written up* is

$$L^{(0)} = \int_0^\infty \underbrace{\left[1 - \pi_d(\sigma)\right]}_{\text{not written at } \sigma} \times \underbrace{f_{LN}\left(\sigma; a + b\log n, s^2\right)}_{\text{precision model}}\, d\sigma \tag{9}$$

Combining Equations 8 and 9, for a given domain d, the likelihood function is :

$$\mathcal{L}(\theta) = \prod_{D=1} L^{(1)} \prod_{D=0} L^{(0)} \tag{10}$$

[11] We can vary this threshold or also include a threshold at the 10% level.

We can then recover parameter estimates $\hat{\psi}_d$ from

$$\hat{\psi} = \arg\max_{\psi} \mathcal{L}(\psi) \tag{11}$$

Where,

$$\psi_d = \Bigg(\underbrace{(\mu_d,\ \tau_d)}_{\substack{\text{effects,} \\ \text{heterogeneity}}} \ , \ \underbrace{(a_d,\ b_d,\ c_d\ ;\ s_d)}_{\text{standard errors}} \ , \ \underbrace{(q_{d,\ell},\ q_{d,h})}_{\text{selection}} \Bigg)$$

We estimate the model separately for each domain by maximum likelihood. See the appendix for estimation details.

3.3 Inference

For inference, the model is estimated by constrained maximum likelihood separately by outcome domain. Inference is challenging because the likelihood can be nonlinear and asymmetric in small samples, and because several parameters are subject to inequality constraints. In particular, heterogeneity must satisfy

$$\tau_d \geq 0,$$

and the write-up probabilities satisfy

$$0.05 \leq q_{d,\ell} \leq q_{d,h} \leq 0.995.$$

When estimates are close to these boundaries, standard Wald intervals can be misleading. A Wald interval imposes symmetry and can assign probability to infeasible values, such as negative heterogeneity. We therefore use likelihood-ratio inversion as our approach to inference.

Let

$$\omega_d = (\mu_d, \tau_d, a_d, b_d, s_d, q_{d,\ell}, q_{d,h})$$

denote the full parameter vector in outcome domain d. Let $\widehat{\omega}_d$ be the unrestricted maximum likelihood estimate:

$$\widehat{\omega}_d = \arg\max_{\omega_d \in \Omega} \ell_d(\omega_d),$$

where Ω is the constrained parameter space. For a scalar parameter of interest ψ_d, such as μ_d or τ_d, and a null value ψ_0, define the restricted estimator

$$\widehat{\omega}_d(\psi_0) = \arg\max_{\omega_d \in \Omega:\, \psi_d = \psi_0} \ell_d(\omega_d).$$

That is, we impose the candidate value $\psi_d = \psi_0$ and re-optimize all remaining nuisance parameters. This re-optimization produces the profile likelihood for ψ_d: for each candidate value ψ_0, the likelihood is evaluated at the best-fitting values of all other parameters.

The likelihood-ratio statistic is

$$LR_d(\psi_0) = 2\left[\ell_d(\widehat{\omega}_d) - \ell_d(\widehat{\omega}_d(\psi_0))\right]. \tag{12}$$

This statistic asks how much worse the model fits when ψ_d is fixed at ψ_0, allowing all other parameters to adjust. Candidate values that fit almost as well as the unrestricted estimate remain in the confidence set while candidate values that substantially worsen the fit are rejected.

Our confidence intervals invert this likelihood-ratio statistic using the standard χ^2_1 distribution. When the tables report standard errors rather than confidence intervals, we report confidence-interval-equivalent standard errors implied by these likelihood-ratio intervals.

Likelihood-ratio inversion has several advantages in this setting. It respects the constrained parameter space, permits asymmetric intervals, and accounts for nuisance-parameter adjustment. When testing a candidate value of μ_d, for example, the heterogeneity parameter, write-up probabilities, and precision-model parameters are all re-estimated under that restriction.

The χ^2_1 reference distribution is still an asymptotic approximation. It is most reliable when the parameter of interest is locally interior and the nuisance parameters are well identified. It may be less reliable when estimates lie on or near inequality boundaries. This concern is most relevant for the smaller domains, especially enrollment, where the estimated heterogeneity is noisy and some write-up probabilities can be near their bounds. We therefore interpret enrollment inference, and inference on boundary-prone parameters more generally, with additional caution.

4 Results

4.1 Main Effects and Heterogeneity

Table 2 reports bias-corrected mean effects from the joint likelihood model with standard errors in parentheses. $\hat{\mu}$ is the corrected mean effect for each domain, and $\hat{\tau}$ is the estimated heterogeneity in a given domain. The unit is standard deviations.

The corrected average effect for test scores is 0.05 SD. For a student at the 50th

percentile of a standardized test, this corresponds to moving to approximately the 52nd percentile, which is a modest shift. The effect on grade outcomes is larger, at 0.07 SD. The standard deviation of Grade Point Average (GPA) is often around one, so if this effect were focused on average grades, it's roughly equivalent to a student moving up half a letter grade in two of their seven courses.[12] Attendance effect is 0.05 SD, equivalent to roughly 1 fewer day absences, which is based on Rogers and Feller (2018). All three outcomes are statistically significant at the 5% level. However, enrollment effects, which are 0.03 SD, is imprecise and not statistically significant. Its wide confidence interval reflects the small sample size in this domain (13 total observations, of which 9 are written).

When factoring in their costs, these effects compare favorably to several other interventions. For instance, one recent meta-analysis of tutoring programs by Kraft et al. (2024) finds pooled effects of 0.4 SD across all studies. However, these effects decline steeply with scale: studies of programs serving more than 1,000 students show an average effect of 0.16 SD (unadjusted for publication bias), and quasi-experimental evaluations at larger scale often find even smaller effects. Our bias-corrected test score effect is roughly one-quarter of the large-scale tutoring benchmark at a fraction of the cost. High-dosage tutoring at scale costs \$1,500–\$4,000 per student per year, while remote parental engagement programs typically cost closer to \$3 per student. This implies a cost-per-SD-gain for remote parental engagement that is an order of magnitude lower than large-scale tutoring.

Another reference point is a meta-analysis of school spending effects by Jackson and Mackevicius (2024). The authors find that \$1,000 per pupil increase sustained for four years improves test scores by 0.03 SD. Our corrected test score effect is comparable in magnitude, but the per-student cost of remote programs is orders of magnitude smaller than a \$1,000/pupil spending increase. School spending effects may have impacts beyond those captured by test scores alone, however.

4.2 Heterogeneity by Study Covariates

We extend the model to examine whether latent mean effects vary systematically with study-level covariates. We do not estimate separate selection models for each subgroup because of small sample sizes. Instead, we assume a common covariate coefficient for each subgroup across domains to gain additional power. Even then, precision is an issue.

For binary covariates, $W = 1$ indicates membership in the subgroup of interest: LMIC settings, younger students (elementary school or younger), and remote instruction

[12]See NCES statistics for average courses taken per year here, and Denning et al. (2026) for GPA standard deviations.

(or information). Within outcome domain d, the binary-covariate model is

$$\Theta \mid W, d \sim \mathcal{N}(\mu_d + \beta W, \tau_d^2),$$

$$x \mid \Theta, \sigma, W, d \sim \mathcal{N}(\Theta, \sigma^2),$$

and therefore

$$x \mid \sigma, W, d \sim \mathcal{N}(\mu_d + \beta W, \tau_d^2 + \sigma^2).$$

μ_d is the domain-specific mean effect for $W = 0$. β is the pooled covariate effect for W, and τ_d is residual cross-study heterogeneity within domain d after accounting for W. The precision model and selection probabilities remain domain-specific.

For sample-size effect heterogeneity, we use a separate specification because sample size is continuous. Let

$$\widetilde{n} = \exp(\overline{\log n})$$

denote the global geometric mean sample size in the estimation sample. We estimate

$$\Theta \mid n, d \sim \mathcal{N}\left(\mu_d(\widetilde{n}) + \beta \left(\log n - \log \widetilde{n}\right), \tau_d^2\right).$$

Study precision typically improves with sample size at a diminishing rate, so a log specification is a reduced-form way to capture this relationship. $\mu_d(\widetilde{n})$ is the fitted mean effect in domain d at the global mean sample size, and β is the change in the latent mean effect associated with a one-log-point increase in sample size. Because the model is linear in $\log n$, this coefficient is also the average marginal effect of log sample size.

Table 3 reports the binary-covariate results. The differential effects associated with LMIC settings, younger students, and remote instruction are economically small and statistically indistinguishable from zero. The estimated effect for LMIC settings is $\hat{\beta} = -0.01$, for younger students, $\hat{\beta} = 0.03$, and for remote instruction, $\hat{\beta} = -0.01$. The residual heterogeneity estimates $\hat{\tau}_d$ remain close to their baseline values across these specifications, especially for test scores. This suggests that the sizeable variance in effects, particularly for test scores, is not well explained by the observed binary covariates for geography, student age, or delivery mode.

Table 4 Panel A reports the sample-size specification. The pooled coefficient on log sample size is negative, $\hat{\beta} = -0.016$ (95% CI $[-0.019, -0.014]$ and $p = 0.04$). A one-log-point increase in sample size is therefore associated with a 0.016 SD lower latent mean effect. Equivalently, a doubling of sample size changes the covariate by

$$\Delta \log n = \log(2n) - \log(n) = \log(2) + \log(n) - \log(n) = \log(2).$$

Thus, the predicted change in the latent mean effect is $\hat{\beta}\log(2) \approx -0.011$ SD.

This result is important because a common identifying assumption in meta-study models is that latent effects are independent of study precision. Our estimates suggest that this assumption is questionable in this RCT setting. Randomized trials are often designed using ex ante power calculations, so interventions expected to generate smaller effects are *chosen* to have larger samples. We therefore interpret the negative sample-size correlation as evidence that effect sizes and precision are jointly related to study design, which may be important to allow for in meta-analyses of RCTs.[13]

Panel B of Table 4 illustrates how the sample size changes fitted domain means. Column $\hat{\mu}_d(\widetilde{n})$ reports the fitted effect at the global geometric mean sample size, $\widetilde{n} = 2216$, while column $\hat{\mu}_d(\widetilde{n}_d)$ reports the fitted effect at the domain-specific geometric mean sample size. Thus, despite the correlation between effects and study precision, the predicted means are close to the main estimates. Allowing the latent mean to vary with sample size does not substantively change our initial findings.

4.3 Empirical Bayes Posterior Summary

Lastly, Table 5 summarizes the empirical Bayes (EB) posterior estimates in each domain and Table 11 reports the estimate at the individual estimate-level. We estimate these posteriors by combining each written estimate and standard error, (x, σ), with the domain-level distribution from the model, $(\hat{\mu}_d, \hat{\tau}_d)$.[14] The EB posterior means θ_{EB} partially pool noisy studies toward the domain mean, with the amount of shrinkage determined by the precision of each study. Precise studies move little, while imprecise studies, especially those with very large positive or negative raw estimates, are pulled substantially toward $\hat{\mu}_d$. The posterior probability $\Pr(\theta_{EB} > 0)$ provides a study-level measure of how likely it is that the underlying effect is actually positive once both sampling error and cross-study heterogeneity are taken into account.

The test scores domain has the largest heterogeneity, $\hat{\tau}_d = 0.06$, so shrinkage can be meaningful for those effects. For grades, the posterior means are tightly concentrated around the domain mean of about 0.08, and essentially every study has an overwhelmingly high posterior probability of a positive effect. Attendance shows the same pattern, though with somewhat more uncertainty. Overall, the EB estimates suggest that the favorable grades and attendance findings are not being driven by a few outlier studies.

Enrollment should still be interpreted with caution given the small sample size. The

[13] We model this relationship to address this point, but this relies on our model being a good approximation to the truth.

[14] This is only possible for the written studies.

EB summaries therefore provide a useful descriptive shrinkage exercise, not strong evidence that enrollment effects are precisely characterized.

5 Selection and Characterizing the File Drawer

5.1 Selection

Tables 6 and 7 show that unwritten studies are not a random subset of all conducted studies, though the resulting file-drawer problem is, arguably, not extreme. Written studies tend to have larger sample sizes than unwritten studies in three of the four domains, with the exception of enrollment. This pattern is important because smaller studies are mechanically less precise, and therefore less likely to generate large absolute z-statistics. Descriptively, Table 6 also provides motivation for using sample-size information from unwritten studies rather than treating the missing part of the literature as a random subset of all conducted trials.

Table 7 shows that overall write-up rates are high, ranging from 0.69 in enrollment to 0.90 in grades.[15] Overall write-up rates, however, do not reveal how selected these write-up rates are.

The joint-likelihood approach with unwritten studies allows us to answer this questions. The model can separate the unconditional probability that a study is written up from the conditional probability of write-up given whether the result is statistically significant. This distinguishes overall write-up rates from significance-driven, selective write-up rates.

Table 7 shows these conditional write-up rates in columns $\hat{q}_\ell$ and $\hat{q}_h$. The probability a study outcome is written conditional on finding a statistically-significant estimate is $\hat{q}_h$, while the probability a study outcome is written-up conditional on *not* finding a statistically significant estimate is $\hat{q}_\ell$. Attendance and grades both have high observed write-up rates, 0.89 and 0.90, but statistically significant results are far more likely to be written up in both domains. For attendance, $\hat{q}_\ell = 0.74$ and for grades $\hat{q}_\ell = 0.72$. Put differently, high average write-up rates do not rule out significance-based selection within the written evidence base.

Taken together, the two tables suggest that the file drawer problem operates through two mechanisms: unwritten studies tend to be less precise due to smaller sample sizes,

[15] One caveat is that just under one quarter of studies report multiple estimates within a domain, while each unwritten study signals the *existence* of a potential set of outcomes in each domain. We can reweight studies to account for this issue, which does not significantly alter the results.

and statistically significant results are more likely to be written up. The extent and form of selection varies across outcomes, with the largest gaps between write-up probabilities for statistically significant and insignificant results appearing in attendance, grades, and enrollment. Test scores show a more modest gap.

5.2 Characterizing the File Drawer

To characterize the file drawer, we plot the distribution of effects for written and unwritten studies. The purpose is to show what the selection model implies about the file drawer: conditional on the observed estimation sample and the estimated parameters, how would the distribution of observed estimates differ between studies that are written up and studies that remain unwritten?

For each outcome domain d, we use the fitted parameters to estimate the model-implied density of the observed estimate X conditional on write-up status:

$$f_d(x \mid D = 1) \qquad \text{and} \qquad f_d(x \mid D = 0).$$

The first is the density of observed estimates among written studies. The second is the counterfactual density of observed estimates that unwritten studies would have produced had their estimates been observed.

It is helpful to begin with the joint density of an estimate, standard error, and write-up status, conditional on sample size. For a given sample size n, the written joint density is

$$f_d(x, \sigma, D = 1 \mid n) = p_d(x/\sigma)\, \phi\big(x; \widehat{\mu}_d, \widehat{\tau}_d^2 + \sigma^2\big)\, f_{LN}\Big(\sigma; \widehat{a}_d + \widehat{b}_d \log n + cW, \widehat{s}_d^2\Big),$$

and the analogous unwritten joint density is

$$f_d(x, \sigma, D = 0 \mid n) = [1 - p_d(x/\sigma)]\, \phi\big(x; \widehat{\mu}_d, \widehat{\tau}_d^2 + \sigma^2\big)\, f_{LN}\Big(\sigma; \widehat{a}_d + \widehat{b}_d \log n + cW, \widehat{s}_d^2\Big).$$

The likelihood for written studies evaluates the joint density at the observed (x, σ). The likelihood for unwritten studies integrates over the unobserved estimate and standard error because neither x nor σ is observed. For each n, we can also integrate out σ to have a density over x:

$$f_d(x, D = 1 \mid n) = \int_0^\infty f_d(x, \sigma, D = 1 \mid n)\, d\sigma,$$

and

$$f_d(x, D = 0 \mid n) = \int_0^\infty f_d(x, \sigma, D = 0 \mid n)\, d\sigma.$$

Figure 5 plots the model-implied distribution of effects estimates separately for written and unwritten studies.[16] Unwritten distributions are generally shifted to the left of written distributions: ranging from 0.02 to 0.03 SD smaller across most domains, but 0.04 SD smaller in grades. This suggests that the file drawer contains studies with smaller effects, but not a large set of strongly negative findings. The pattern is consistent with our main findings that the bias correction attenuates the written literature without overturning the positive effects.

Overall, selection-driven write up choices appear to shift the evidence base toward larger positive effects rather than by suppressing an underlying body of strongly negative results. Grades show this pattern most clearly: written studies over-represent the upper part of the positive effect distribution, but the unwritten distribution remains centered above zero. Similar, but milder, patterns are also observed in test scores and attendance. Enrollment is more weakly identified and more sensitive to individual observations, so its distribution should be interpreted with additional caution.

Taken together, the figure complements the selection-probability estimates by showing how the file drawer in this setting mainly trims magnitudes rather than reversing substantive conclusions. Nonetheless, this analysis is model-dependent and sample sizes in meta-analyses are often small.

5.3 Selection-reweighted Distribution of Study z-statistics

In addition to the approach above in Section 5.2, we also estimate the selection-adjusted distribution of z-statistics using a Horvitz–Thompson (HT) approach. While Section 5.2 uses the full parametric model to construct the counterfactual distribution of effect estimates among unwritten studies whose x and z are unobserved, the HT approach reweights observed z-statistics to represent the complete conducted evidence base under the fitted selection rule.

If all conducted studies are written up, then the z-statistics distribution should be directly observable. Since z_i is observed only for written studies, we recover this distribution by weighting each written estimate by the inverse of its fitted write-up probability. For each estimate $i \in \mathcal{I}_d$ in domain d, it has an inverse probability weight of

$$\widehat{\omega}_i = \frac{1}{\widehat{p}_d(z_i)}.$$

The intuition is that written estimates in regions of the z-distribution with lower fitted

[16] Each reports domain-level densities rather than densities at a particular sample size n. See calculation details in the Appendix Section E.

write-up probabilities represent more conducted but unwritten evidence and therefore receive larger weights.

For a histogram bin B_b with bin width h, the unweighted written density is

$$\widehat{f}_d^{\,\text{obs}}(B_b) = \frac{\sum_{i \in \mathcal{I}_d : D_i = 1} \mathbb{1}\{z_i \in B_b\}}{N_{wd} \cdot h},$$

where N_{wd} is the number of written estimates in domain d. The selection-reweighted density is

$$\widehat{f}_d^{\,\text{HT}}(B_b) = \frac{1}{h} \frac{\sum_{i \in \mathcal{I}_d : D_i = 1} \widehat{\omega}_i \, \mathbb{1}\{z_i \in B_b\}}{\sum_{i \in \mathcal{I}_d : D_i = 1} \widehat{\omega}_i}.$$

Both histograms are normalized to integrate to one within domain, so the figure compares the shape of the written and selection-reweighted z-distributions.

Figure 6 is a selection-reweighted view of the observed written evidence (red bars) in contrast to the actual observed distribution (blue bars). The dashed vertical lines mark $|z| = 1.96$, the threshold at which the write-up probability changes (by construction). The comparison between the unweighted and selection-reweighted bars shows where the fitted model adds mass to the observed evidence distribution. If insignificant results are estimated to have lower write-up probabilities, for example, this histogram mechanically places more weight below $|z| = 1.96$.

Across domains, however, these corrections are generally modest. The red and blue histograms remain fairly close, and the corrected distributions continue to be concentrated on positive z-statistics. This suggests that the written literature somewhat overstates the magnitude and concentration of favorable findings. As before, however, it is not dominated by a large hidden group of null or negative results. Put differently, if statistically insignificant findings were rarely written up, the red bars would rise much more dramatically below the $|z| = 1.96$ threshold than they do here. The overall picture is one of attenuation.

6 Robustness

6.1 Estimates across Different Meta-analytic Models

First, we address three distinct, potential sources of adjustment in our meta-analysis: precision weighting, random-effect heterogeneity, and correction for write-up selection. To do so, we compare and contrast our estimates against different specifications used in meta analysis as well as the additional adjustments made in ours. All specifications are

estimated separately by outcome domain.

1. Unweighted mean among written estimates. The average effect is given by

$$\widehat{\mu}_d^{UW} = \frac{1}{N_{wd}} \sum_{i \in \mathcal{I}_d : D_i = 1} x_i.$$

   This estimate is descriptive. It gives each written estimate equal weight and does not adjust for sampling precision, cross-study heterogeneity, or publication and write-up selection.

2. DerSimonian–Laird (DL) random-effects meta-analysis estimated on written study estimates (DerSimonian and Laird, 1986). It assumes

$$x_i \mid \Theta_i, \sigma_i \sim \mathcal{N}(\Theta_i, \sigma_i^2), \qquad \Theta_i \sim \mathcal{N}(\mu_d, \tau_d^2),$$

   and uses inverse-variance weights

$$w_i(\widehat{\tau}_d^2) = \frac{1}{\sigma_i^2 + \widehat{\tau}_d^2}.$$

   The resulting estimate is

$$\widehat{\mu}_d^{DL} = \frac{\sum_{i \in \mathcal{I}_d : D_i = 1} w_i(\widehat{\tau}_d^2) x_i}{\sum_{i \in \mathcal{I}_d : D_i = 1} w_i(\widehat{\tau}_d^2)}.$$

   This specification accounts for precision and random-effects heterogeneity, but it treats the written evidence base as if it were selected independently of statistical significance.

3. Andrews–Kasy's (AK) publication bias correction (Andrews and Kasy, 2019). This is estimated only using written studies and it corrects publication bias. We implement it using the same normal random-effects distribution and two-bin selection function in the model we estimate. As this specification uses only written studies, the absolute levels of conditional write-up probabilities are not identified.

4. We also incorporate the precision model into the AK estimator to see whether the precision model drives differences in estimates:

$$\log \sigma_i \mid n_i \sim \mathcal{N}(a_d + b_d \log n_i + c_d W_i, s_d^2).$$

   For written studies, its likelihood contribution is

$$L_{id}^{AK+\sigma} = \frac{p_d(z_i)\, \phi(x_i; \mu_d, \tau_d^2 + \sigma_i^2)\, f_{LN}(\sigma_i; a_d + b_d \log n_i + c_d W_i, s_d^2)}{\int p_d(x/\sigma_i)\, \phi(x; \mu_d, \tau_d^2 + \sigma_i^2)\, dx}.$$

This specification is included as a minor check on the model. Since σ_i is observed for written studies, adding the precision model should not materially change the written-only estimates of μ_d, τ_d, or relative selection. This model just helps assess whether the precision model on its own is altering selection-correction estimates.

5. Lastly, the paper's joint likelihood approach presented in Section 3. The key difference between the written-only AK specifications and the joint likelihood is identification of absolute selection probabilities. Written-only selection models identify only the relative write-up rate $q_{d,\ell}/q_{d,h}$, as multiplying all write-up probabilities by a common constant leaves the distribution of written estimates unchanged. Observing both written and unwritten studies identifies the absolute probability that a conducted study-domain observation is written up.

Table 8 reports estimates from the five specifications by each outcome domain. $\hat{\mu}$ is the grand mean and $\hat{\tau}$ is the cross-study heterogeneity. $q_{d,\ell}/q_{d,h}$ is the relative publication probability between the significant and non-significant results, while $\hat{q}_\ell$ and $\hat{q}_h$ are the absolute publication probability. Dash indicates that that specification does not produce that estimate.

The first comparison in Table 8, the unweighted written mean compared to the DerSimonian–Laird (DL) random-effects estimate, shows that the largest adjustment often comes before any selection correction at all. The mean falls only slightly for test scores, from 0.06 to 0.05, but much more for grades, attendance, and enrollment. This pattern implies that equal-weight averages are heavily influenced by noisy estimates, especially in the domains with fewer studies. In our setting, then, the biggest differences often arise from weighting and heterogeneity rather than from the file-drawer adjustment alone.

Comparing specifications (3) and (4) shows that adding the precision model to the Andrews–Kasy estimator changes almost nothing. Across all four domains, the estimated mean effect, heterogeneity, and relative write-up probability $\hat{q}_\ell/\hat{q}_h$ are nearly unchanged once the precision model is layered onto written studies only. This indicates that the precision model itself is not mechanically generating the selection results. The important additional identifying content comes only when unwritten studies enter the likelihood.

For test scores, which have by far the largest number of studies, all model-based specifications deliver essentially the same mean effect, and the estimated relative selection ratio remains close to one. This stability suggests that the test-score results are comparatively well identified and that, in this domain, the main contribution of the joint likelihood is to separately pin down the conditional write up probabilities.

Much larger differences show when comparing specification (4) to (5) in domains with fewer studies. In grades, attendance, and enrollment, the AK specifications imply much

lower relative selection ratios. Under the AK-plus-precision specification, $\hat{q}_\ell/\hat{q}_h$ is 0.06 for grades, 0.20 for attendance, and 0.11 for enrollment. These ratios are much more severe than the corresponding estimate for test scores, and they pose a tension with the relatively high observed write-up shares in the full sample of written and unwritten studies. Once the unwritten studies are incorporated, the joint likelihood still estimates significance-based selection, but the implied absolute write-up probabilities are less extreme: $(\hat{q}_\ell, \hat{q}_h) = (0.72, 0.99)$ for grades, $(0.74, 0.99)$ for attendance, and $(0.60, 0.99)$ for enrollment.

One interpretation is that we have drastically underestimated the number of unwritten studies. That seems inconsistent with what we find in the test-score domain, and it is unclear why relative write-up rates for insignificant versus significant studies would be so different across domains, especially when many of these remote interventions directly target grades or attendance. We believe the evidence suggests the low written-only AK ratios in the smaller-N domains are better interpreted as reflecting weak-identification than as literal evidence of severe suppression of statistically insignificant results.

The last point suggests the unwritten studies data help identify the selection model when the sample size is small and estimates are less stable. This is most evident for grades, where the joint likelihood moves the estimated mean back toward the conventional random-effects estimate after the aggressive downward correction implied by the AK model. For attendance, the mean changes less. The joint model adds value by preventing the low-N domains from telling what might be a less plausible severe-selection story.

6.2 Robustness to Unwritten Study Incompleteness

We can more directly address the concern that we might have missed many unwritten studies. In this section we examine how the joint estimates would change if the true number of unwritten studies were larger than the number observed in our audit of registries, reports and author inquiries.

The exercise expands the unwritten component of the joint likelihood in Equation 10. For domain d, let $L_d^{(1)}(\theta_d)$ and $L_d^{(0)}(\theta_d)$ denote the written and unwritten likelihood contributions. The overall likelihood is:

$$\ell_d(\theta_d; \kappa) = \sum_{i:D_i=1} \log L_{id}^{(1)}(\theta_d) + \kappa \sum_{i:D_i=0} \log L_{id}^{(0)}(\theta_d).$$

where $\kappa \geq 1$ is an expansion factor for the unwritten studies. While the case $\kappa = 1$ is the baseline audited model, values $\kappa > 1$ are equivalent to fractionally adding additional unwritten studies with the same empirical sample-size distribution as the audited unwritten studies in that domain.

In other words, the robustness check asks what happens if the audit found the right kind of unwritten studies but too few of them. Additional missingness is allowed to change the fitted selection probabilities, the heterogeneity parameter, and the mapping from sample size to precision. The key assumption is that the audit recovered the representative *type* of unwritten studies but too few of them.

Figure 7 plots the re-estimated mean at each value of κ. The dashed line indicates the implied missing share from the joint likelihood model, and the dotted line indicates that from the AK specification.[17] We extend the assumed unwritten share to 80%.

The corrected means for scores, grades, and attendance remain positive over a wide range of assumed unwritten shares and the main conclusions are not very sensitive to moderate incompleteness. For scores, grades, and attendance, the results are also fairly stable, even when the model doubles the assumed missing share. The enrollment mean drops much more sharply, however the results were neither significant nor well-identified to begin with.

Lastly, the AK estimates imply much more severe selection than the audited joint model in the smaller sample domains. Matching the AK benchmark would require a very large amount of additional missingness beyond what we uncovered.

6.3 Distributional Diagnostics

The model assumes that latent true effects within each domain are drawn from a normal random-effects distribution,

$$\Theta_i \mid d \sim \mathcal{N}(\mu_d, \tau_d^2).$$

This is distinct from the assumption of asymptotic normality of treatment-effect estimators, which motivates the sampling model

$$x_i \mid \Theta_i, \sigma_i, d \sim \mathcal{N}(\Theta_i, \sigma_i^2).$$

Because the latent effects Θ_i are not observed directly, the normality assumption for Θ_i cannot be tested directly from the observed estimates. The observed estimates combine latent heterogeneity, sampling error, and write-up selection. Therefore, these diagnostics based on the observed estimates should be interpreted cautiously and not as formal tests of latent-effect normality.

Nonetheless, we conduct three diagnostics typical in the meta-analysis literature. First, we compare deconvolved density estimates to fitted normal densities. Deconvolu-

[17] See calculation details in Appendix F.

tion is useful because it attempts to separate the dispersion of true effects from sampling noise. However, standard deconvolution methods do not fully reproduce the joint likelihood used in the main analysis, particularly the write-up selection component and the precision model for unwritten studies. Second, we report QQ plots for the observed estimates. These plots show whether the empirical distribution of estimates departs strongly from a normal benchmark. However, because the estimates have different standard errors and are selected into the written sample, departures from the QQ line can reflect sampling-error heterogeneity or write-up selection, not necessarily non-normality of the latent effects. Third, we report Shapiro–Wilk tests of normality, but again, these are not direct tests of the latent random-effects distribution.

With these caveats in mind, Figure 8 compares the deconvolved density estimates with normal densities fitted to the same mean and variance. We use a deconvolution kernel density approach following Delaigle et al. (2008) and Wang and Wang (2011). The tuning parameter is set using the optimal bandwidth following Fan (1991) (see Appendix Table 10). Figure 8 reports the comparison between the deconvolution results and a normal distribution. The deconvolved density function closely follows the normal distribution for all outcomes.

We also use the Shapiro-Wilk test with the estimated τ from the main result in Table 2. The test fails to reject the null hypothesis that the data are normally distributed for all outcomes, except for attendance with p-value of 0.03. When allowing for multiple estimates per study, we also reject the null for test scores due to an outlier in Angrist et al. (2023). The QQ-plot in Figure 9 also descriptively suggests that the observed data are roughly normal.

7 Information Policy Decisions

The preceding sections summarize the average effects of remote parental engagement interventions within each outcome domain. A policymaker, however, faces a different question: if a new program were implemented in a new context, how likely is it to clear a policy-relevant benchmark, what is the marginal value of public funds, and how valuable would additional evidence be before making that decision? This section translates the selection-corrected estimates into that decision problem.

7.1 Set up

Let $\mathcal{D}$ denote the current meta-analysis evidence base of written and unwritten study-domain observations used to estimate the model. Suppose a policy maker implemented a new program targeting outcomes d with true latent effect of $\Theta_{\text{new},d}$.[18]

The policy calculations distinguish three related objects. First, μ_d is the latent domain-level mean effect. Second, $\Theta_{\text{new},d}$ may differ from μ_d as effects vary across settings, programs, and populations. Third, a future program would not observe $\Theta_{new,d}$ perfectly. Instead, it would observe $x_{\text{new},d}$, an estimate with sampling error.

We summarize estimation uncertainty in the domain mean using the standard error of the estimated grand mean $\widehat{\mu}_d$:

$$\mu_d \mid \mathcal{D} \approx \mathcal{N}\left(\widehat{\mu}_d, V_{\mu d}\right), \qquad V_{\mu d} = SE(\widehat{\mu}_d)^2. \tag{13}$$

This is a normal approximation to uncertainty about the true mean μ_d, centered around the estimate $\widehat{\mu}_d$. Conditional on μ_d, a new implementation has true effect

$$\Theta_{\text{new},d} = \mu_d + \eta_d, \qquad \eta_d \sim \mathcal{N}(0, \tau_d^2).$$

We can plug in the fitted heterogeneity parameter $\widehat{\tau}_d$ so that:

$$\Theta_{\text{new},d} \mid \mu_d \sim \mathcal{N}\left(\mu_d, \widehat{\tau}_d^2\right). \tag{14}$$

This means we account for uncertainty in μ_d, but we do not integrate over uncertainty in τ_d. The calculations should therefore be interpreted as conditional on the estimated heterogeneity parameter $\widehat{\tau}_d$.

A future study would observe the true effect with sampling error. If $x_{\text{new},d}$ is the estimate from one additional study, then

$$x_{\text{new},d} = \Theta_{\text{new},d} + \varepsilon_d = \mu_d + \eta_d + \varepsilon_d, \qquad \varepsilon_d \sim \mathcal{N}(0, \sigma^2_{\text{new},d}),$$

Therefore,

$$x_{\text{new},d} \mid \mu_d \sim \mathcal{N}\left(\mu_d, \widehat{\tau}_d^2 + \sigma^2_{\text{new},d}\right). \tag{15}$$

[18] The notations follow the main set up in Section 3.

7.2 Decision Thresholds

Under the set up above, we can model the adoption decision as a threshold problem. Let $a \in \{0, 1\}$ denote the action, where $a = 1$ means adopting the program and $a = 0$ otherwise. For a threshold T measured in standard-deviation units, the payoff is

$$U(a, \Theta_{\text{new},d}) = a\,(\Theta_{\text{new},d} - T)\,. \tag{16}$$

The policymaker adopts a program when the expected benefit of adoption exceeds the threshold. Here, we consider two values of T.

First, we consider a financial break-even threshold. A program is worth implementing if its benefit exceed the cost. Let K_{SD} be the present value of a one-standard-deviation gain in achievement and c_{PE} be the implementation cost per student.

We use $c_{PE} = \$3$ which corresponds to the median cost for the engagement programs in our sample. To calculate K_{SD}, we assume that

$$K_{SD} = \eta Y_0 \frac{1 - (1 + r)^{-H}}{r}$$

where η is the relationship between one SD gain in test scores and future earnings and Y_0 is baseline earnings. H is the working horizon with a real discount rate of r. We assume $\eta = 0.05$, meaning that one SD gain in test scores increases annual earnings by 5%.[19] In a lower-income setting with baseline earnings Y_0 of \$3,000 and 40 years of working horizon with $r = 0.05$, K_{SD} is approximately \$2,600.

Thus the financial break-even threshold is

$$T_{\text{financial}} = \frac{c_{PE}}{K_{\text{SD}}} = \frac{3}{2600} \approx 0.001.$$

This threshold is close to zero in effect-size units. Under this calibration, the main question is essentially whether effects are positive.

This threshold formulation is equivalent to the Marginal Value of Public Funds (MVPF) comparison. In dollar units, the net benefit from adoption is

$$NB(a, \Theta_{\text{new},d}) = a\,(K_{\text{SD}}\Theta_{\text{new},d} - c_{PE})\,.$$

Dividing by K_{SD} gives the effect-size payoff above with $T = c_{PE}/K_{\text{SD}}$. This maps to

[19] This is a strong assumption, because we draw on the *association* between test scores and earnings in Ozawa et al. (2022)

MVPF in Hendren and Sprung-Keyser (2020), which is

$$MVPF_d = \frac{K_{\text{SD}}\mu_d}{c_{PE}}.$$

Under the assumed parameter values, an effect of 0.05 SD implies

$$MVPF = \frac{2,600 \times 0.05}{3} \approx 43.$$

In other words, even modest effects can generate very high value per dollar as the programs are inexpensive.

Second, we consider a portfolio-priority threshold. The relevant policy question may not be whether remote parental engagement beats a zero or financial break-even bar, but whether it should be prioritized over another effective education investment. Let Θ_{alt} denote the effect size of an alternative intervention and c_{alt} its per-student cost. Remote parental engagement is preferred on cost-effectiveness grounds when

$$\frac{\mathbb{E}[\Theta_{\text{new},d}]}{c_{PE}} > \frac{\Theta_{\text{alt}}}{c_{\text{alt}}}.$$

Equivalently, adoption requires

$$\mathbb{E}[\Theta_{\text{new},d}] > c_{PE}\frac{\Theta_{\text{alt}}}{c_{\text{alt}}} \equiv T_{\text{alt}}.$$

Suppose a policymaker is choosing between a remote parental engagement intervention and a scalable targeted-instruction program with $\Theta_{\text{alt}} = 0.12$ SD and $c_{\text{alt}} = \$7.20$ per student.[20] The threshold will be

$$T_{\text{alt}} = 3 \cdot \frac{0.12}{7.20} \approx 0.05.$$

We therefore treat $T = 0.05$ as the main portfolio-comparison threshold, while $T \approx 0$ captures the financial break-even decision.

7.3 Probability a New Intervention Surpasses a Threshold

We first translate the fitted model into the probability that a new implementation clears a threshold. For this, we use the plug-in latent-effect distribution

$$\Theta_{\text{new},d} \sim N\left(\widehat{\mu}_d, \widehat{\tau}_d^2\right). \tag{17}$$

[20] For instance, see Angrist and Meager (2023); Banerjee et al. (2007, 2016a); Duflo et al. (2024)

This distribution does not include sampling variance, as $\Theta_{\text{new},d}$ is the true effect in the new context. Sampling variance would enter only if we were modeling the estimate from a future study.

The probability that a new implementation exceeds threshold T is

$$\Pr\left(\Theta_{\text{new},d} > T\right) = 1 - \Phi\left(\frac{T - \widehat{\mu}_d}{\widehat{\tau}_d}\right),^{21} \tag{18}$$

Figure 10 plots this probability over a range of thresholds. The figure illustrates how likely a new implementation is to exceed a given effect-size benchmark under the estimated latent-effect distribution.

Grades has the highest probability of clearing most thresholds shown, reflecting its relatively large corrected mean and modest heterogeneity. Scores and attendance are closer to the portfolio threshold. Attendance has a high value at lower thresholds due to its small heterogeneity, while scores has more upper-tail upside due to its larger heterogeneity. Enrollment has a nontrivial probability of clearing modest thresholds, but the curve should still be interpreted with caution as the estimates are small and noisy.

This plot reiterates why average effects are not sufficient for policy interpretation: outcomes with similar means can imply different chances of clearing a benchmark once cross-context heterogeneity is taken into account.

7.4 Value of Perfect Information

The threshold curve describes the probability of success. We next ask how much it would be worth to a policymaker to resolve uncertainty before deciding. With current evidence, the policymaker uses the estimated mean and adopts when $\widehat{\mu}_d > T$:

$$a_d^{\text{now}} = \mathbf{1}\{\widehat{\mu}_d > T\}, \qquad V_d^{\text{now}} = \max\left(\widehat{\mu}_d - T, 0\right). \tag{19}$$

We show the Estimated Value of Perfect Information (EVPI) as follows (cf. Raiffa and Schlaifer (2000)). We include both uncertainty about the domain mean and cross-context heterogeneity. Combining Equations 13 and 14 yields

$$\Theta_{\text{new},d} \mid \mathcal{D} \sim \mathcal{N}\left(\widehat{\mu}_d, V_{\mu d} + \widehat{\tau}_d^2\right). \tag{20}$$

EVPI is the gain from learning $\Theta_{\text{new},d}$ before choosing whether to adopt. In other words,

[21] If $\widehat{\tau}_d = 0$, this probability is one when $\widehat{\mu}_d > T$ and zero otherwise.

it is the value of a *perfect signal* about the next implementation's true effect.

$$\mathrm{EVPI}_d(T) = \mathbb{E}\left[\max\left(\Theta_{\mathrm{new},d} - T, 0\right)\right] - \max\left(\widehat{\mu}_d - T, 0\right). \tag{21}$$

Because $\Theta_{\mathrm{new},d} \mid \mathcal{D}$ is normal, this expectation has a closed form. Let

$$s_d = \sqrt{V_{\mu d} + \widehat{\tau}_d^2}, \qquad z_d = \frac{\widehat{\mu}_d - T}{s_d}.$$

Then

$$\mathbb{E}\left[\max\left(\Theta_{\mathrm{new},d} - T, 0\right)\right] = \left(\widehat{\mu}_d - T\right)\Phi(z_d) + s_d\phi(z_d), \tag{22}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal CDF and PDF. EVPI is largest when the current evidence leaves substantial probability on both sides of the threshold.

7.5 Value of One Additional Study

How much value one additional RCT would provide? Let $x_{\mathrm{new},d}$ denote the estimate observed from a new study in domain d. Consider Equation 15,

$$x_{\mathrm{new},d} \mid \mu_d \sim \mathcal{N}\left(\mu_d, \widehat{\tau}_d^2 + \sigma_{\mathrm{new},d}^2\right). \tag{23}$$

Here $\sigma_{\mathrm{new},d}$ is the median standard error among written studies in domain d, used as the precision of one additional study. The observed variance includes both sampling error $\sigma_{\mathrm{new},d}^2$, and cross-context heterogeneity $\widehat{\tau}_d^2$, as a new study observes one draw from the distribution of true implementation effects.

After observing $x_{\mathrm{new},d}$, the policymaker updates beliefs about μ_d. Let

$$m'_d \equiv \mathbb{E}\left[\mu_d \mid x_{\mathrm{new},d}\right] \tag{24}$$

be the posterior mean of μ_d. Before seeing the new study, m'_d is itself random because the new study estimate is random. The pre-posterior variance, measures how much one additional study is expected to move beliefs about μ_d. Under normal-normal updating, this variance is

$$\mathrm{Var}(m'_d \mid \mathcal{D}) = V_{\mu d}^{pre} = \frac{V_{\mu d}^2}{V_{\mu d} + V_{d,1}^{obs}}. \tag{25}$$

where $\widehat{\tau}_d^2 + \sigma_{\mathrm{new},d}^2 = V_{d,1}^{obs}$.

The expected value of sample information from one additional study (EVSI) is

$$\mathrm{EVSI}_{d,1}(T) = \mathbb{E}\left[\max\left(m'_d - T, 0\right)\right] - \max\left(\widehat{\mu}_d - T, 0\right). \tag{26}$$

Because $m'_d \mid \mathcal{D}$ is normal, EVSI can be computed similarly as EVPI. Let

$$s_d^{SI} = \sqrt{V_{\mu d}^{pre}}, \qquad z_d^{SI} = \frac{\widehat{\mu}_d - T}{s_d^{SI}}.$$

Then

$$\mathbb{E}\left[\max\left(m'_d - T, 0\right)\right] = (\widehat{\mu}_d - T)\Phi(z_d^{SI}) + s_d^{SI}\phi(z_d^{SI}), \tag{27}$$

and therefore,

$$\text{EVSI}_{d,1}(T) = (\widehat{\mu}_d - T)\Phi(z_d^{SI}) + s_d^{SI}\phi(z_d^{SI}) - \max(\widehat{\mu}_d - T, 0). \tag{28}$$

The distinction between EVPI and EVSI is important. EVPI asks how valuable it would be to know the actual effect in the next implementation context. On the other hand, EVSI asks how valuable one more noisy study would be for learning about the mean.

When $\widehat{\tau}_d$ is large, perfect information (EVPI) about the next implementation can be valuable since many possible true effects lie on both sides of the threshold. But the same heterogeneity makes one generic noisy study less informative (EVSI) about μ_d, because the new study's estimate partly reflects sampling error and unexplained cross-context heterogeneity, summarized by τ_d. Thus heterogeneity can raise EVPI while lowering EVSI.

7.6 Policy Value Results

Table 9 reports EVPI and one-study EVSI in effect-size utility units.[22] At a financial break-even threshold near $T = 0$, information values are generally smaller than at $T = 0.05$, because the average adoption decision is largely resolved for the domains with positive corrected means. EVPI need not be zero at $T = 0$, however, because the calculation is for the new-context effect Θ_{new}. Residual cross-context heterogeneity leaves some value to knowing whether a particular implementation would have a low or negative effect. When the threshold is raised to $T = 0.05$, corresponding to comparison with an alternative effective intervention, information values become larger.

At $T = 0.05$, EVPIs are relatively high for test scores and attendance because their means are close to this threshold and their estimated heterogeneities are large. EVSIs, however, are small: a single additional study is a noisy signal about μ_d when true effects vary substantially across contexts. Enrollment has the largest one-study EVSI among the outcome domains because its mean is the most inconclusive. Grades has a high

[22] Entries are multiplied by 10^{-3}.

probability of clearing $T = 0.05$, and therefore lower EVPI, because the adoption decision is less ambiguous.

For test scores, where we have an established relationship between the outcome and earnings, Table 9 also translates the effect-size utility units into dollars for an illustrative rollout of $N = 100{,}000$ students. Dollar values are computed as

$$\text{VOI}_\$ = K_{\text{SD}}\, N\, \text{VOI}_{SD},$$

where K_{SD} is the present value of a one-standard-deviation gain for one student, N is the number of students affected, and VOI_{SD} is the value of information measured in effect-size units. We use $K_{\text{SD}} = \$2{,}600$, as calculated in Section 7.2.

To interpret this value relative to program cost, we scale the dollar value of information by the total rollout cost:

$$\frac{\text{VOI}_\$}{\text{Cost}_\$} = \frac{K_{\text{SD}} N \text{VOI}_{SD}}{c_{PE} N} = \frac{K_{\text{SD}} \text{VOI}_{SD}}{c_{PE}},$$

where c_{PE} is the per-student cost of the remote parental engagement intervention. Thus, the value of information relative to program cost depends on both the dollar valuation K_{SD} and the program cost c_{PE}.

The policy conclusion is two-sided. Because these interventions are inexpensive, they almost certainly clear a financial break-even threshold in the domains with statistically precise positive corrected means. If the relevant decision is whether to prioritize remote parental engagement over other effective education programs, however, the answer depends on the threshold and on cross-context heterogeneity. One additional RCT effect estimate has limited value when heterogeneity is large because it does not resolve whether the next implementation will clear the policy threshold. Additional research is therefore most valuable when it explains or reduces uncertainty about *why* effects vary across settings.

8 Conclusion

In this paper, we synthesize what is known about low-cost remote parental engagement interventions. To do so, we aim to overcome several methodological challenges. We show how meta-analyses can use information on documented but unwritten studies to characterize the file drawer and correct for write-up selection. Our application is useful because remote parental engagement programs are inexpensive, scalable, and increasingly common, but the evidence base is fragmented across journal articles, working papers,

grant reports, registries, and studies that were conducted but never written up.

We identify 82 randomized controlled trials for remote parental engagements on test scores, grades, attendance, and enrollment, including 24 documented but unwritten studies. Incorporating these unwritten studies changes the publication-bias problem. Selection models using only written studies can identify how write-up probabilities vary with statistical significance up to scale. By observing studies that were conducted but not written up, we can estimate the levels of the conditional probabilities that a study enters the written evidence base. This distinction is important for characterizing the file drawer: it separates selection-driven write up rates from the overall rate at which studies are written up.

We find that overall, remote parental engagement interventions have small, positive effects on test scores, grades, attendance, and weakly positive for enrollment. These effects are not uniform in magnitude. The corrected mean effects are approximately 0.05 standard deviations for test scores and attendance, 0.07 standard deviations for grades, and 0.03 standard deviations for enrollment, though the enrollment effect is imprecise. These estimates are nonetheless meaningful relative to the program's very low cost.

The estimates also show heterogeneity, especially for test scores. The estimated heterogeneity for test scores is roughly as large as the mean effect, implying that the likely effect of a new implementation depends importantly on context, design, and population. Grades, by contrast, show larger mean effects and less heterogeneity, while enrollment remains the weakest and least precisely identified domain.

To incorporate the unwritten studies, the joint-likelihood approach we develop here helps with the selection estimation. This appears particularly valuable for domains with smaller samples. Using only written study estimates in these domains can imply much more severe relative selection than is likely true, with statistically insignificant estimates appearing far less likely to be written up. Once unwritten studies are considered, the estimated selection is less extreme. We interpret this contrast as evidence that selection estimates can be unstable in small-N domains, and adding unwritten studies helps anchor the absolute scale of the file drawer.

We also document that study precision is related to latent effects in this setting. Many meta-analytic models assume that latent treatment effects are independent of standard errors or sample sizes, which holds well in many cases. However, this assumption may fail in the experimental setting as sample sizes are typically chosen through design decisions and power calculations. Using sample sizes observed for both written and unwritten studies together with the precision model, we find that larger studies tend to estimate smaller latent effects. This does not overturn the main results, but reinforces a broader point: evidence synthesis should treat study design, precision, selection, and effect heterogeneity

as jointly related.

Several limitations remain. We may not have recovered every conducted study. For unwritten studies, we observe their existence and sample sizes, but not their effect estimates or standard errors. We therefore rely on distributional assumptions, a precision model, and a selection function that is intentionally simple. Some domains are small, and inference for boundary-prone parameters, especially in enrollment, should be interpreted cautiously. We address this potential missingness with sensitivity analysis in Section 6.2.

Lastly, we illustrate several policy implications. First, remote parental engagement programs are likely cost-justified on a financial break-even basis. Their costs are so low that even small positive impacts can generate high marginal value of public funds. Second, we provide a framework to answer whether remote parental engagement should be prioritized over other effective education investments. The answer depends on a benchmark derived from the alternative program's cost-benefit ratio and heterogeneity. We show that compared to a program like differentiated instruction, parental engagement interventions have high probability of clearing the benchmark in most domains. The value-of-information calculations imply that perfect information is most valuable when uncertainty is substantial and centered near the policy threshold, but one additional generic RCT may capture only a small share of that value when cross-context heterogeneity is large.

References

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

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, 24(2):3–30.

Angrist, N., Ainomugisha, M., Bathena, S. P., Bergman, P., Crossley, C., Cullen, C., Letsomo, T., Matsheng, M., Panti, R. M., Sabarwal, S., et al. (2023). Building resilient education systems: Evidence from large-scale randomized trials in five countries. Technical report, National Bureau of Economic Research.

Angrist, N., Bergman, P., and Matsheng, M. (2022). Experimental evidence on learning using low-tech when school is out. *Nature human behaviour*, 6(7):941–950.

Angrist, N., Evans, D. K., Filmer, D., Glennerster, R., Rogers, H., and Sabarwal, S. (2025). How to improve education outcomes most efficiently? a review of the evidence using a unified metric. *Journal of Development Economics*, 172:103382.

Angrist, N. and Meager, R. (2023). Implementation matters: Generalizing treatment effects in education. Technical report, SSRN.

Avvisati, F., Gurgand, M., Guyon, N., and Maurin, E. (2014). Getting parents involved: A field experiment in deprived schools. *Review of Economic Studies*, 81(1):57–83.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., Shotland, M., and Walton, M. (2016a). Mainstreaming an effective intervention: Evidence from randomized evaluations of "teaching at the right level" in india. Technical report, National Bureau of Economic Research.

Banerjee, A. V., Cole, S., Duflo, E., and Linden, L. (2007). Remedying education: Evidence from two randomized experiments in india. *The quarterly journal of economics*, 122(3):1235–1264.

Banerjee, A. V., Duflo, E., Kremer, M., et al. (2016b). The influence of randomized controlled trials on development economics research and on development policy. *The state of economics, the state of the world*, pages 482–488.

Banerji, R., Berry, J., and Shotland, M. (2017). The impact of maternal literacy and participation programs: Evidence from a randomized evaluation in india. *American Economic Journal: Applied Economics*, 9(4):303–337.

Barrera-Osorio, F., Gonzalez, K., Lagos, F., and Deming, D. J. (2020). Providing performance information in education: An experimental evaluation in colombia. *Journal of Public Economics*, 186:104185.

Bergman, P. (2019). How behavioral science can empower parents to improve children's educational outcomes. *Behavioral Science & Policy*, 5(1):53–67.

Bergman, P. (2020). Nudging technology use: Descriptive and experimental evidence from school information systems. *Education Finance and Policy*, 15(4):623–647.

Bergman, P. (2021). Parent-child information frictions and human capital investment: Evidence from a field experiment. *Journal of political economy*, 129(1):286–322.

Bettinger, E., Cunha, N., Lichand, G., and Madeira, R. (2021). Are the effects of informational interventions driven by salience? Technical Report 350, University of Zurich, Department of Economics, Working Paper.

Brodeur, A., Carrell, S., Figlio, D., and Lusher, L. (2023). Unpacking p-hacking and publication bias. *American economic review*, 113(11):2974–3002.

Brodeur, A., Cook, N., and Heyes, A. (2020). Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11):3634–3660.

Conn, K. M. (2017). Identifying effective education interventions in sub-saharan africa: A meta-analysis of impact evaluations. *Review of educational research*, 87(5):863–898.

Copas, J. and Jackson, D. (2004). A bound for publication bias based on the fraction of unpublished studies. *Biometrics*, 60(1):146–153.

Cunha, F. and Heckman, J. (2007). The technology of skill formation. *American economic review*, 97(2):31–47.

Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *The Annals of Statistics*, 36(2):665–685.

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

Denning, J. T., Nesbit, R. L., Pope, N. G., and Warnick, M. (2026). Easy a's, less pay: The long-term effects of grade inflation. Technical report, National Bureau of Economic Research.

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

Dizon-Ross, R. (2019). Parents' beliefs about their children's academic ability: Implications for educational investments. *American Economic Review*, 109(8):2728–2765.

Dizon-Ross, R. (2021). Using randomized information shocks to understand how parents' investments depend on their children's ability'. Technical report, J-Pal.

Driessen, E., Hollon, S. D., Bockting, C. L., Cuijpers, P., and Turner, E. H. (2015). Does publication bias inflate the apparent efficacy of psychological treatment for major depressive disorder? a systematic review and meta-analysis of us national institutes of health-funded trials. *PloS one*, 10(9):e0137864.

Duflo, A., Kiessel, J., and Lucas, A. M. (2024). Experimental evidence on four policies to increase learning at scale. *The Economic Journal*, 134(661):1985–2008.

Escueta, M., Nickow, A. J., Oreopoulos, P., and Quan, V. (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature*, 58(4):897–996.

Evans, D. K. and Popova, A. (2016). What really works to improve learning in developing countries? an analysis of divergent findings in systematic reviews. *The World Bank Research Observer*, 31(2):242–270.

Evans, D. K. and Yuan, F. (2022). How big are effect sizes in international education studies? *Educational evaluation and policy analysis*, 44(3):532–540.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272.

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

Fujii, T., Ho, C., Ray, R., and Shonchoy, A. S. (2025). Boosting study habits with high-frequency information: A field experiment to aid disadvantaged students. Technical report, Florida International University.

Glewwe, P. and Muralidharan, K. (2016). Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In *Handbook of the Economics of Education*, volume 5, pages 653–743. Elsevier.

Goyal, N., Okuno, A. A., Singhvi, D., and Singhvi, S. (2024). Increasing parents' engagement on ed-tech platforms: Evidence from the field. Technical report, SSRN.

Greaves, E., Hussain, I., Rabe, B., and Rasul, I. (2023). Parental responses to information about school quality: Evidence from linked survey and administrative data. *The Economic Journal*, 133(654):2334–2402.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, 7(2):246–255.

Hendren, N. and Sprung-Keyser, B. (2020). A unified welfare analysis of government policies. *The Quarterly journal of economics*, 135(3):1209–1318.

Huang, A., Komukai, S., Friede, T., and Hattori, S. (2021). Using clinical trial registries to inform copas selection model for publication bias in meta-analysis. *Research Synthesis Methods*, 12(5):658–673.

Huang, A., Morikawa, K., Friede, T., and Hattori, S. (2023). Adjusting for publication bias in meta-analysis via inverse probability weighting using clinical trial registries. *Biometrics*, 79(3):2089–2102.

Hurwitz, L. B., Lauricella, A. R., Hanson, A., Raden, A., and Wartella, E. (2015). Supporting head start parents: Impact of a text message intervention on parent–child activity engagement. *Early Child Development and Care*, 185(9):1373–1389.

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

Jackson, C. K. and Mackevicius, C. L. (2024). What impacts can we expect from school spending policy? evidence from evaluations in the united states. *American Economic Journal: Applied Economics*, 16(1):412–446.

Kraft, M. A. and Bolves, A. J. (2022). Can technology transform communication between schools, teachers, and parents? evidence from a randomized field trial. *Education Finance and Policy*, 17(3):479–510.

Kraft, M. A., Schueler, B. E., and Falken, G. (2024). What impacts should we expect from tutoring at scale? exploring meta-analytic generalizability. Technical Report 24-1031, EdWorking Paper.

Kremer, M., Brannen, C., and Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, 340(6130):297–300.

Leight, J., Asri, V., and Imai, T. (2025). Pathways from registration to publication: Evidence from the aea rct registry. Working paper, available at `https://www.jessicaleight.com/uploads/1/3/2/3/13234647/rct_registry_merged.pdf`.

Lichand, G., Christen, J., and Egeraat, E. V. (2024). Neglecting students' socio-emotional skills magnified learning losses during the pandemic. *npj Science of Learning*, 9(1):28.

Lichand, G., Christen, J., and van Egeraat, E. (2023). Behavioral nudges reduced dropout risk among vulnerable students during the pandemic: experimental evidence from brazil. In *AEA papers and proceedings*, volume 113, pages 494–497. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Mani, A., Mullainathan, S., Shafir, E., and Zhao, J. (2013). Poverty impedes cognitive function. *science*, 341(6149):976–980.

Mayer, S. E., Kalil, A., Oreopoulos, P., and Gallegos, S. (2019). Using behavioral insights to increase parental engagement: The parents and children together intervention. *Journal of Human Resources*, 54(4):900–925.

McEwan, P. J. (2015). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of educational research*, 85(3):353–394.

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.

Mullainathan, S. and Shafir, E. (2013). *Scarcity: Why having too little means so much*. Macmillan.

Nickow, A., Oreopoulos, P., and Quan, V. (2024). The promise of tutoring for prek–12 learning: A systematic review and meta-analysis of the experimental evidence. *American Educational Research Journal*, 61(1):74–107.

Ome, A. and Menendez, A. (2022). Using sms and parental outreach to improve early reading skills in zambia. *Education Economics*, 30(4):384–398.

Ozawa, S., Laing, S. K., Higgins, C. R., Yemeke, T. T., Park, C. C., Carlson, R., Ko, Y. E., Guterman, L. B., and Omer, S. B. (2022). Educational and economic returns to cognitive ability in low-and middle-income countries: A systematic review. *World development*, 149:105668.

Raiffa, H. and Schlaifer, R. (2000). *Applied statistical decision theory*. John Wiley & Sons.

Robinson, C. D., Lee, M. G., Dearing, E., and Rogers, T. (2018). Reducing student absenteeism in the early grades by targeting parental beliefs. *American educational research journal*, 55(6):1163–1192.

Rogers, T., Duncan, T., Wolford, T., Ternovski, J., Subramanyam, S., and Reitano, A. (2017). A randomized experiment using absenteeism information to" nudge" attendance. rel 2017-252. *Regional Educational Laboratory Mid-Atlantic*.

Rogers, T. and Feller, A. (2018). Reducing student absences at scale by targeting parents' misbeliefs. *Nature Human Behaviour*, 2(5):335–342.

Sirvani, H. (2007). The effect of teacher communication with parents on students' mathematics achievement. *American secondary education*, pages 31–46.

Snell, E. K., Wasik, B. A., and Hindman, A. H. (2022). Text to talk: Effects of a home-school vocabulary texting intervention on prekindergarten vocabulary. *Early Childhood Research Quarterly*, 60:67–79.

Snilstveit, B., Gallagher, E., Phillips, D., Vojtkova, M., Eyers, J., Skaldiou, D., Stevenson, J., Bhavsar, A., and Davies, P. (2017). Protocol: Interventions for improving learning outcomes and access to education in low-and middle-income countries: A systematic review. *Campbell Systematic Reviews*, 13(1):CL2–176.

Todd, P. E. and Wolpin, K. I. (2007). The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human capital*, 1(1):91–136.

Vevea, J. L. and Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60(3):419–435.

Vivalt, E. (2020). How much can we generalize from impact evaluations? *Journal of the European Economic Association*, 18(6):3045–3089.

Wang, X.-F. and Wang, B. (2011). Deconvolution estimation in measurement error models: the r package decon. *Journal of statistical software*, 39:1–24.

Wolf, S. and Lichand, G. (2023). Nudging parents and teachers to improve learning and reduce child labor in cote d'ivoire. *npj Science of Learning*, 8(1):37.

World Bank (2022). 70% of 10-year-olds now in learning poverty, unable to read and understand a simple text. Press release, World Bank.

York, B. N., Loeb, S., and Doss, C. (2019). One step at a time: The effects of an early literacy text-messaging program for parents of preschoolers. *Journal of Human Resources*, 54(3):537–566.

Tables

Table 1: Summary of included studies

Domain	Total Studies		Write-up Rate	Total Estimates		Median Observations	
	Written	Unwritten		Written	Unwritten	Written	Unwritten
Scores	40	19	0.68	64	19	3148.5	1657.0
Grade	11	5	0.69	18	5	2094.0	1656.5
Attendance	21	4	0.84	32	4	7656.0	2950.0
Enrollment	8	4	0.67	9	4	1412.0	2306.5

Notes: Total studies are the number of academic projects by their write-up status. Total estimates are the number of estimates across domains. One paper can have multiple estimates in multiple domains, thus this is a conservative estimate of unwritten estimates.

Table 2: Main Results by Domain

	Scores	Grades	Attendance	Enrollment
$\hat{\mu}_d$	0.05	0.07	0.05	0.03
	(0.01)	(0.01)	(0.01)	(0.04)
$\hat{\tau}_d$	0.06	0.03	0.03	0.04
	(0.01)	(0.01)	(0.01)	(0.04)
N total	81	20	36	13
N written	64	18	32	9
N unwritten	17	2	4	4

Notes: Primary model includes bounded, monotone two-bin selection function. Standard errors in parentheses, computed from profile-likelihood confidence intervals using a χ^2_1 reference distribution. N refers to the number of estimates in each domain by the written status.

Table 3: Heterogeneity for Binary Covariates

				$\hat{\tau}_d$ by Domain			
Covariate (W)	$\hat{\beta}$	95% CI	p-value	Scores	Grades	Attendance	Enrollment
LMIC	-0.01	[-0.03, 0.02]	0.69	0.06	0.03	0.03	0.04
Younger student	0.03	[-0.01, 0.05]	0.12	0.06	0.03	0.03	0.04
Remote instruction	-0.01	[-0.03, 0.02]	0.65	0.06	0.03	0.04	0.04

Notes: Each row is a separate analysis on covariate W. For each binary covariate, $\hat{\beta}$ is the latent mean-effect difference for $W = 1$ relative to $W = 0$, while each domain retains its own baseline mean, heterogeneity, precision model, and selection probabilities. The $\hat{\tau}_d$ entries are domain-specific heterogeneity estimates from the same pooled-β fit.

Table 4: Heterogeneity by Sample Size

Panel A. Pooled coefficient on log sample size

	Average marginal effect $\hat{\beta}$	95% CI	p-value
Log sample size	-0.016	[-0.019, -0.014]	0.04

Panel B. Predicted effects at geometric-mean sample sizes

Domain	$\widetilde{n}_d$	$\hat{\mu}_d(\widetilde{n})$	$\hat{\mu}_d(\widetilde{n}_d)$	$\hat{\mu}_d$	N
Scores	1,704	0.06	0.06	0.05	81
Grades	2,694	0.08	0.08	0.07	20
Attendance	4,632	0.07	0.06	0.05	36
Enrollment	1,093	0.02	0.03	0.03	13

Notes: The sample-size covariate is $W_i = \log n_i - \log \widetilde{n}$, where $\widetilde{n} = \exp(\overline{\log n}) = 2,216$ is the global geometric mean sample size in the updated estimation sample. The parameter of interest the change in the latent mean effect associated with a one-log-point increase in sample size. Confidence intervals are estimated using likelihood-ratio approach described in the text. $\hat{\mu}_d(\widetilde{n})$ is the fitted domain mean at the global geometric mean sample size. $\hat{\mu}_d(\widetilde{n}_d)$ is the fitted domain mean at the domain-specific geometric mean sample size, equivalently the model-predicted mean averaged over that domain's empirical sample-size distribution. $\hat{\mu}_d$ is from the model without covariates.

Table 5: Empirical Bayes Summary by Domain

	Total		Share			
Domain	Estimates	Studies	$x > 0$	$\theta_{EB} > 0$	$E[\Pr(\theta_{EB} > 0)]$	$E[\theta_{EB}]$
Scores	64	40	0.77	0.89	0.84	0.05
Grades	18	11	1.00	1.00	1.00	0.08
Attendance	32	21	0.94	1.00	0.98	0.05
Enrollment	9	8	0.78	0.89	0.80	0.04

Notes: The table summarizes empirical Bayes (EB) posterior estimates for written observations ($D = 1$). For each observed estimate x_i with standard error σ_i, the calculation combines the study estimate with the domain-level fitted distribution, $\Theta_i \sim N(\widehat{\mu}_d, \widehat{\tau}_d^2)$, to form the posterior distribution $\Theta_i \mid x_i, \sigma_i, d$. Total estimate is the number of estimates in each domain. Total study is the number of distinct academic papers. Share of $x > 0$ and $\theta_{EB} > 0$ is the percent of estimates that are greater than zero in each category. $E[\Pr(\theta > 0)]$ is the average posterior probability that the latent true effect is positive and mean EB $\hat{\theta}$ is the average EB posterior mean across all estimates.

Table 6: Median Sample Sizes by Write-Up Status

Domain	Median Observations		Difference (%)
	Written	Unwritten	
Scores	3148.5	1657.0	90.01
Grades	2094.0	1656.5	26.41
Attendance	7656.0	2950.0	159.53
Enrollment	1412.0	2306.5	-38.78

Notes: Medians are computed over outcome-domain observations by write-up status in the estimation sample. Difference is $(n_{\text{Written}} - n_{\text{Unwritten}})/n_{\text{Written}}$, so positive values mean unwritten observations have smaller median sample sizes than written observations.

Table 7: Selection Probabilities and Observed Write-Up Rates

Domain	$\hat{q}_\ell$	$\hat{q}_h$	Pr(Written)
Scores	0.75	0.87	0.79
	(0.07)	(0.06)	(0.05)
Grades	0.72	0.99	0.90
	(0.14)	(0.05)	(0.07)
Attendance	0.74	0.99	0.89
	(0.11)	(0.04)	(0.05)
Enrollment	0.60	0.99	0.69
	(0.14)	(0.05)	(0.13)

Notes: Standard errors in parentheses. SEs for $\hat{q}_\ell$ and $\hat{q}_h$ are CI-equivalent SEs, computed from profile-likelihood confidence intervals using a χ_1^2 reference distribution. Pr(Written) is the observed write-up rate with binomial SE.

Table 8: Benchmark specifications by outcome domain

Specification	Data	N	N written	$\hat{\mu}$	$\hat{\tau}$	$\hat{q}_\ell/\hat{q}_h$	$\hat{q}_\ell$	$\hat{q}_h$
Scores								
(1) Unweighted mean	Written	64	64	0.06	–	–	–	–
(2) DL	Written	64	64	0.05	0.06	–	–	–
(3) AK, normal Θ	Written	64	64	0.05	0.06	1.00	–	–
(4) AK + precision	Written	64	64	0.05	0.06	1.00	–	–
(5) Joint likelihood	Written+unwritten	81	64	0.05	0.06	0.87	0.75	0.87
Grades								
(1) Unweighted mean	Written	18	18	0.12	–	–	–	–
(2) DL	Written	18	18	0.09	0.04	–	–	–
(3) AK, normal Θ	Written	18	18	0.04	0.02	0.06	–	–
(4) AK + precision	Written	18	18	0.03	0.02	0.06	–	–
(5) Joint likelihood	Written+unwritten	20	18	0.07	0.03	0.72	0.72	0.99
Attendance								
(1) Unweighted mean	Written	32	32	0.10	–	–	–	–
(2) DL	Written	32	32	0.04	0.01	–	–	–
(3) AK, normal Θ	Written	32	32	0.04	0.03	0.14	–	–
(4) AK + precision	Written	32	32	0.04	0.03	0.20	–	–
(5) Joint likelihood	Written+unwritten	36	32	0.05	0.03	0.74	0.74	0.99
Enrollment								
(1) Unweighted mean	Written	9	9	0.14	–	–	–	–
(2) DL	Written	9	9	0.06	0.07	–	–	–
(3) AK, normal Θ	Written	9	9	0.01	0.00	0.11	–	–
(4) AK + precision	Written	9	9	0.01	0.00	0.11	–	–
(5) Joint likelihood	Written+unwritten	13	9	0.03	0.04	0.60	0.60	0.99

Notes: N is the number of outcome-domain records in each specification. N written is the total number of written estimates. Specifications (1)–(4) estimate written-only models, while specification (5) uses audited written plus unwritten records. Specifications (3)–(5) use a normal random-effects distribution $\Theta \sim N(\mu, \tau^2)$ and the same two-bin selection function at $|z| = 1.96$. In the written-only AK rows, q_h is normalized to one; only the relative selection probability $\hat{q}_\ell/\hat{q}_h$ is identified, therefore absolute $\hat{q}_\ell$ and $\hat{q}_h$ are not reported. Specification (5) uses audited written and unwritten records with domain-specific absolute write-up probabilities.

Table 9: Value of Information of an Additional Study

	T	Scores	Grades	Attendance	Enrollment	Scores $
EVPI	0	6.76	0.19	1.23	9.62	1.76
EVSI	0	0.00	0.00	0.00	0.10	0.00
EVPI	0.05	23.56	4.80	13.58	15.93	6.13
EVSI	0.05	0.12	0.00	0.14	1.48	0.03
EVSI/EVPI (%)	0.05	0.49	0.00	1.03	9.29	–

Notes: EVPI is computed for perfect information on the new-context effect Θ_{new} under variance $V_\mu + \tau^2$ (see Equation 20). EVSI is the one-study value from learning μ via a new study with $V_{obs} = \tau^2 + \sigma^2_{new}$ (see Equation 23). Domain entries are in $\times 10^{-3}$ effect-size utility units. The Scores $ column reports dollar values in millions for test-score effects only, where we have an earnings conversion from achievement gains. Dollar values use $\text{VOI}_\$ = K_{\text{SD}} N \cdot \text{VOI}_{SD}$, with $K_{\text{SD}} = \$2{,}600$ per SD and $N = 100{,}000$ students. The threshold $T = 0.05$ SD corresponds to comparison between a remote parental engagement intervention and a differentiated instruction intervention.

Figures

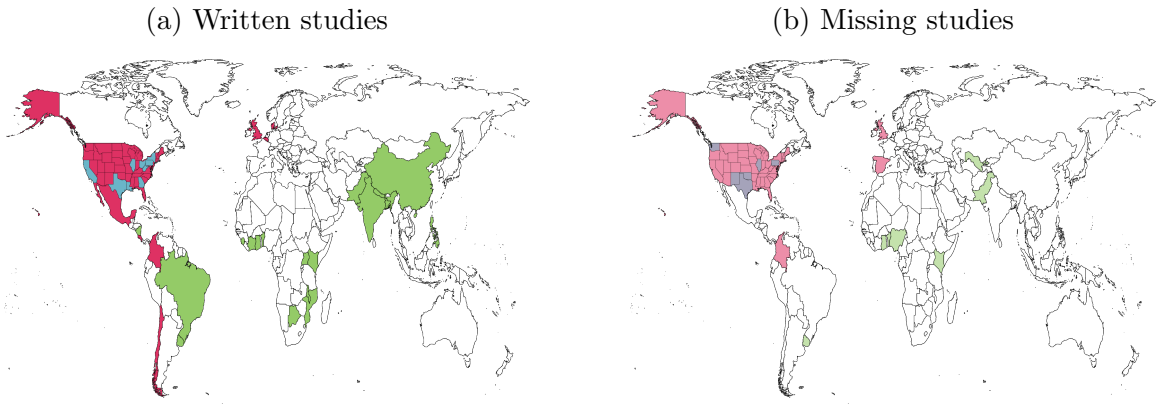


Figure 1: Geographic coverage

Notes: Highlighted countries are those where at least one experiment was conducted. Red indicates high-income countries and green indicates low-income countries, proxy by OECD membership. Blue highlights in the United States indicate states where experiments were conducted. The maps exclude Antarctica and no study was conducted there.

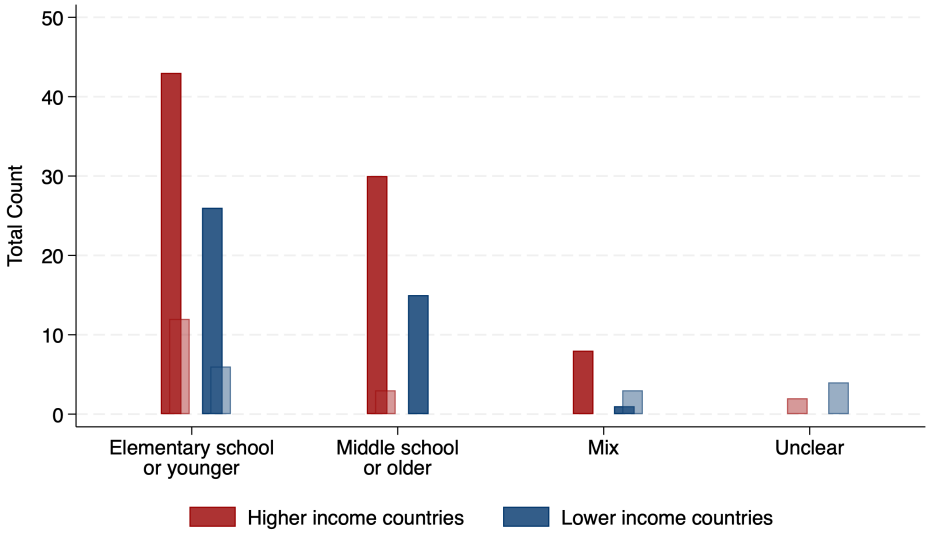


Figure 2: Number of studies by targeted age group

Notes: The faded bars indicate the intended target ages for the missing studies. High versus low income in our study is determined by OECD membership.

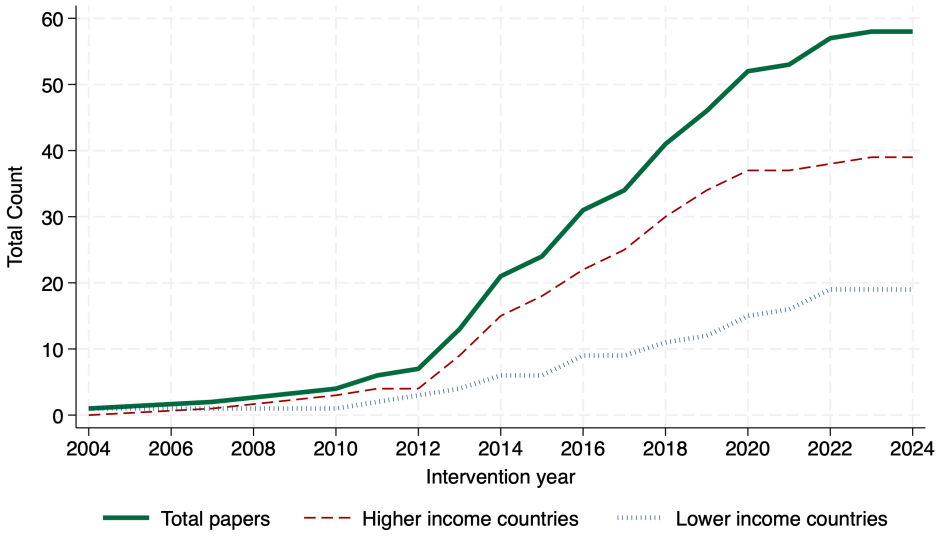


Figure 3: Number of studies by intervention year

Notes: The x-axis indicates the first year of the intervention reported in each paper, except for Snell et al. (2022), for which we use the year of the first draft submission because the intervention year is unavailable.

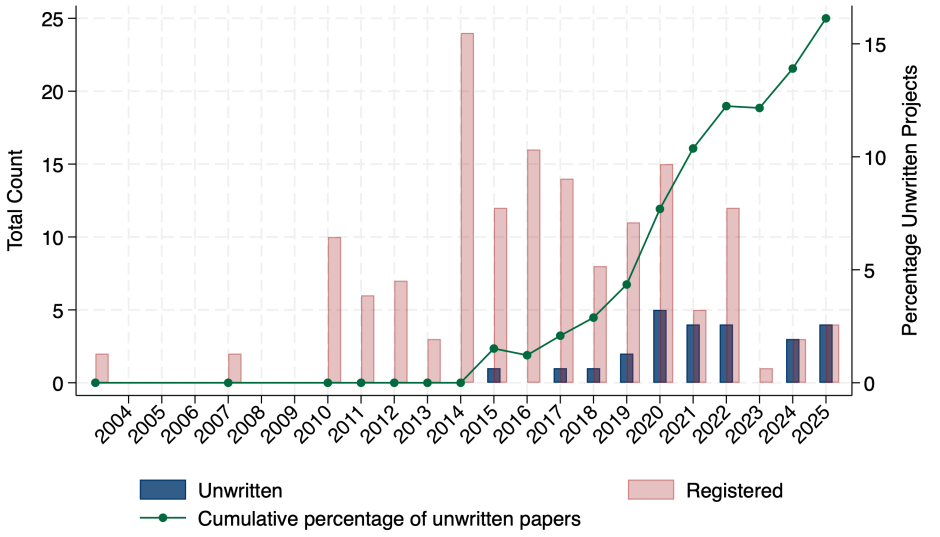


Figure 4: Number of unwritten studies compared to number of registered projects

Notes: The x-axis indicates the project registration year reported in each paper. The information is obtained by AEA RCT registry and grant reports. In case the registration year is missing, we use the first year of the intervention.

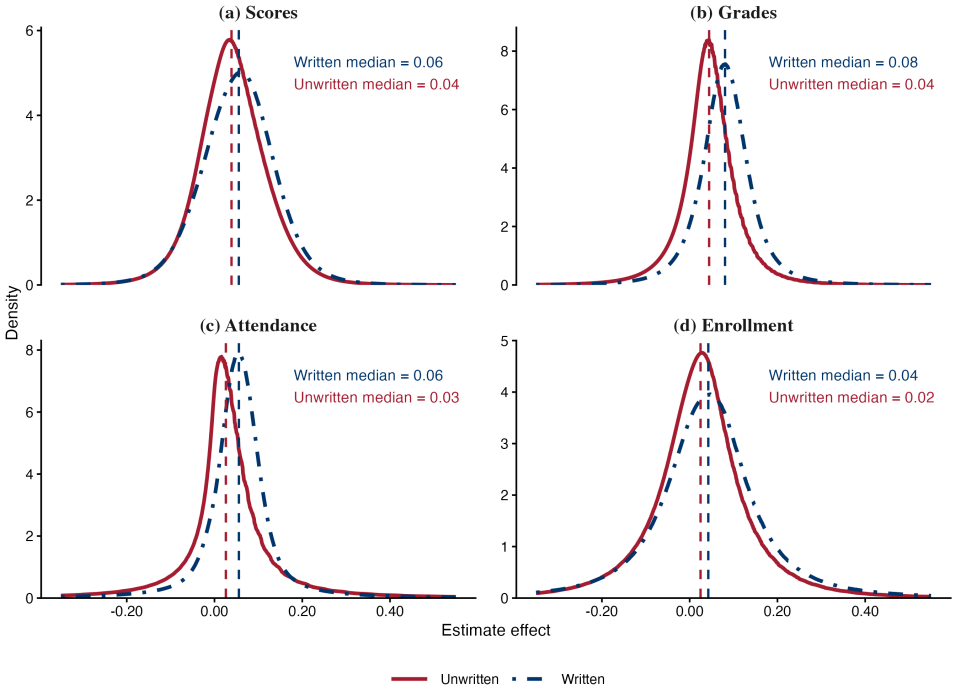


Figure 5: Distribution of Effects for Written and Unwritten Studies

Notes: The figure plots model-implied densities of observed effect estimates separately for written and unwritten studies. Written studies are those with reported estimates and standard errors; unwritten studies are documented studies without written-up estimates. The written curve is the fitted density $f_d(x \mid D = 1)$. The unwritten curve is the fitted counterfactual density $f_d(x \mid D = 0)$, i.e., the distribution of estimates that unwritten studies would have generated under the fitted model. For each domain, both curves average over the same empirical distribution of sample sizes and the fitted precision model. Thus differences between written and unwritten curves reflect the estimated selection mechanism rather than raw differences in sample-size composition. Dashed vertical lines mark the median of each fitted conditional density.

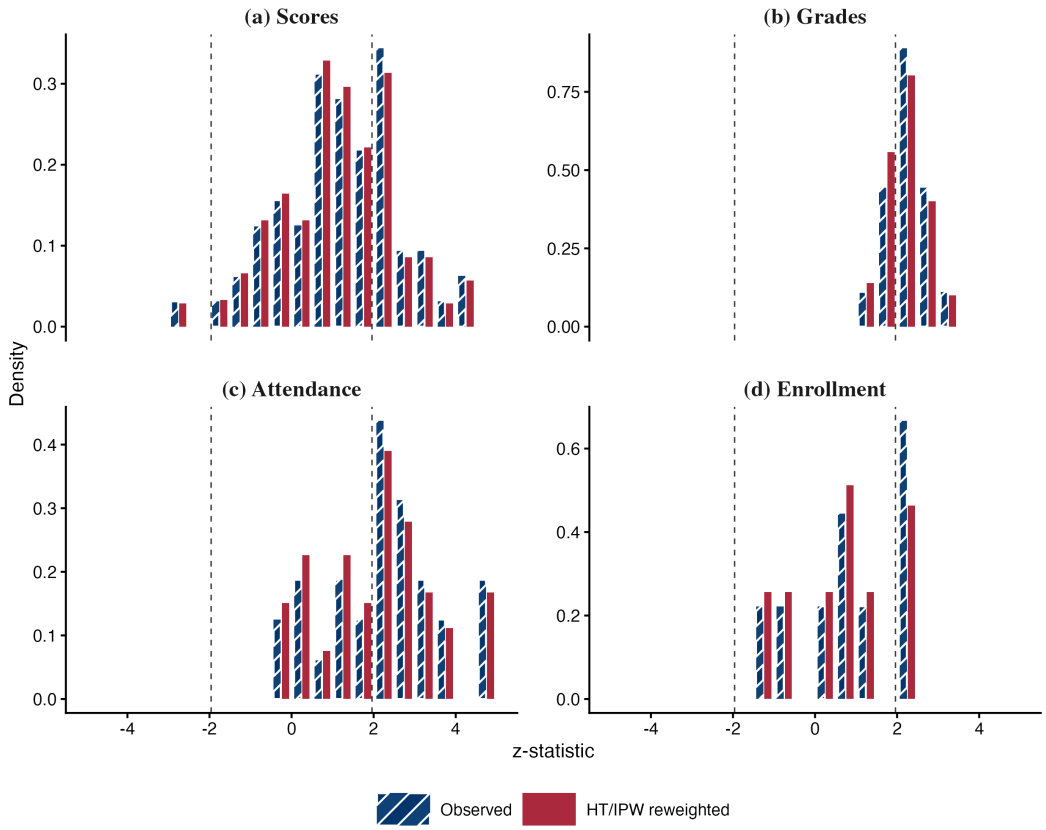


Figure 6: Reweighted histogram

Notes: This figure compares the observed histogram (stripped blue) to a reweighted one (red) using Horvitz-Thompson approach to account for selection-driven write up rates. The dashed vertical lines show the $z = |1.96|$.

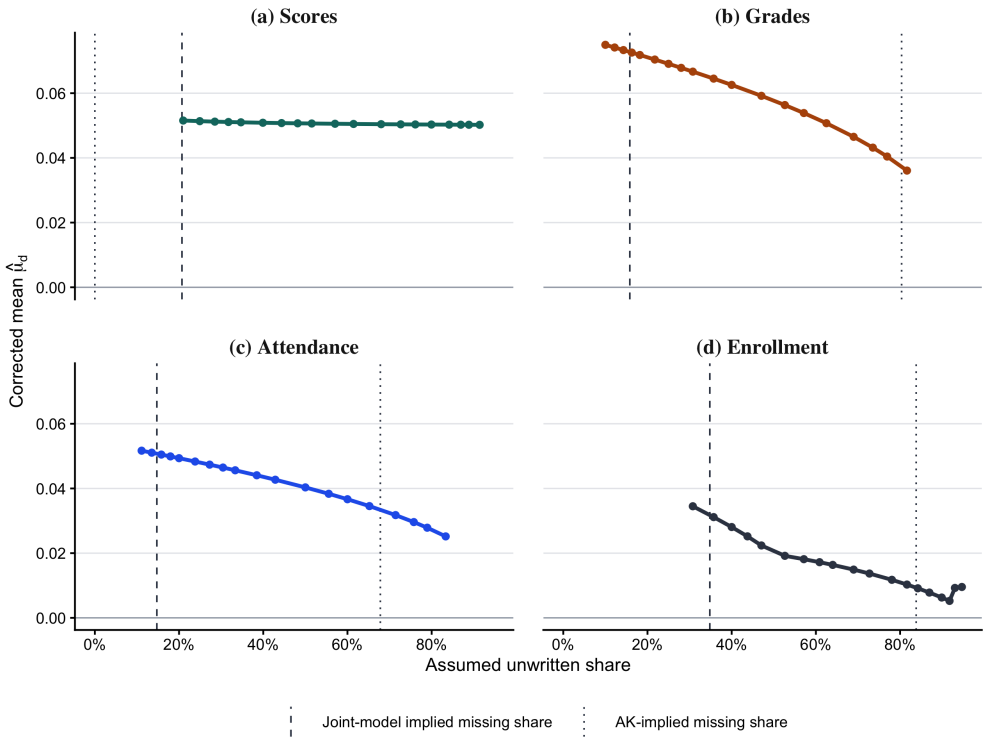


Figure 7: Robustness to additional missing studies

Notes: This figure plots the bias-corrected means as function of an increasing unwritten share, κ of studies by domain. The full model is re-estimated at a given level of κ. The dashed vertical line indicates our model's implied missing share, and the dotted vertical line indicates the AK-implied missing share.

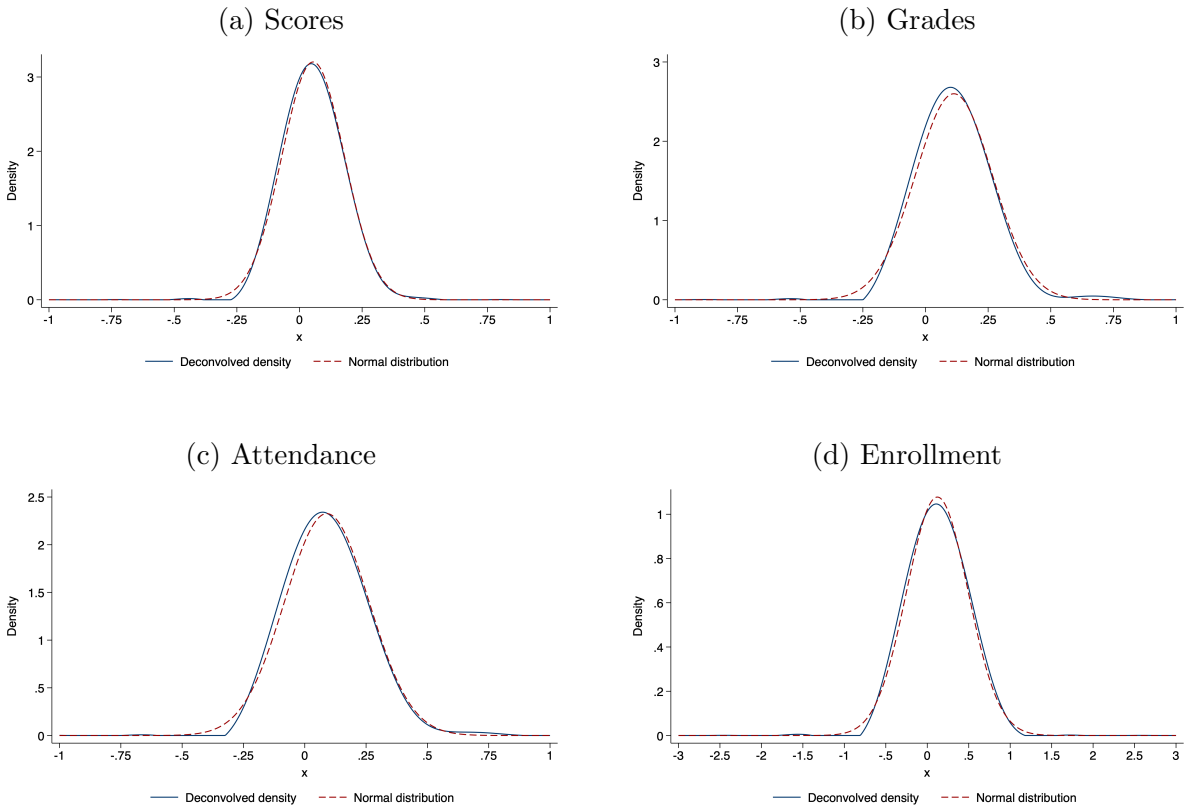
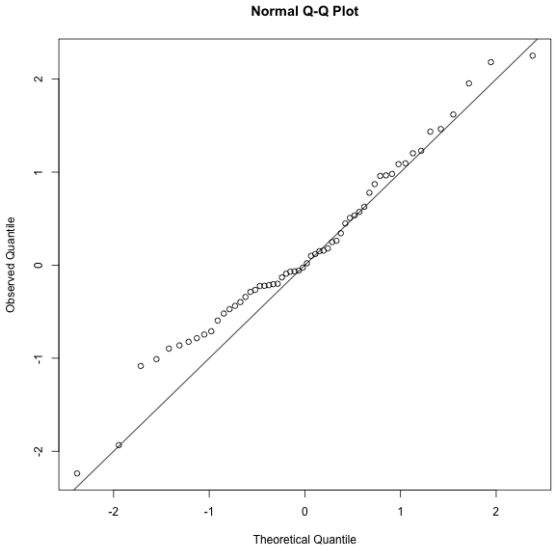


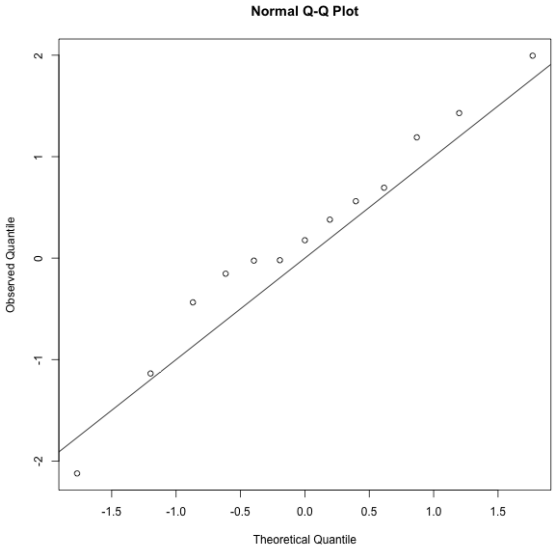
Figure 8: Deconvolved Density Estimates by Domain

Notes: This figure plots the deconvoled density and a normal distribution with means and standard deviations correspondent to the deconvolved values. This includes all estimates by domain. See methodological details in Delaigle et al. (2008); Fan (1991); Wang and Wang (2011).

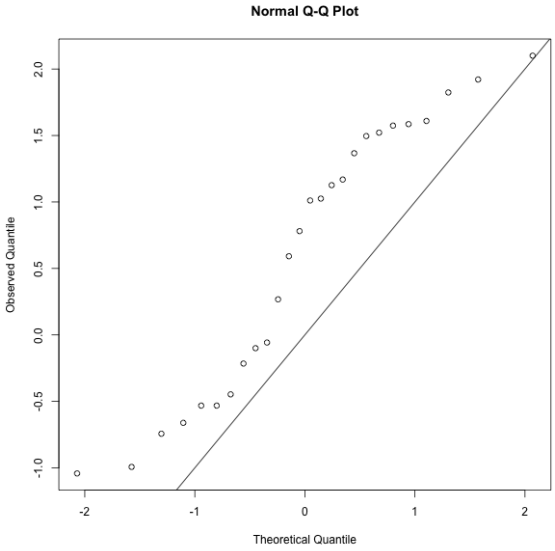
(a) Scores



(b) Grades



(c) Attendance



(d) Enrollment

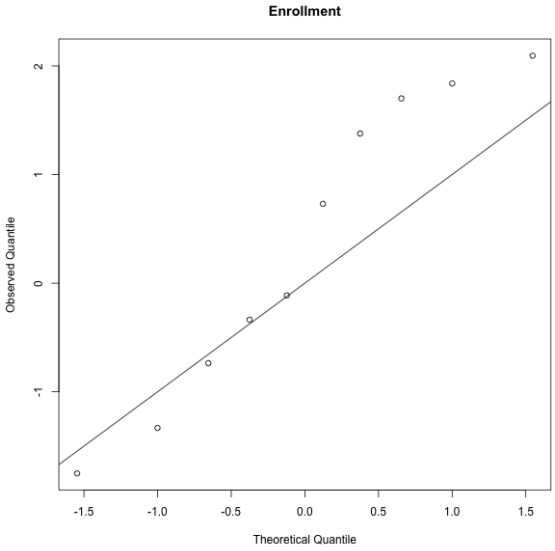


Figure 9: QQ plot for primary estimates

Notes: This figure displays a Quantile-Quantile plot comparing theoretical and observed quantiles for primary estimates within each domain. The straight line displays $x = y$. For multiple estimates, see Appendix Figure 12

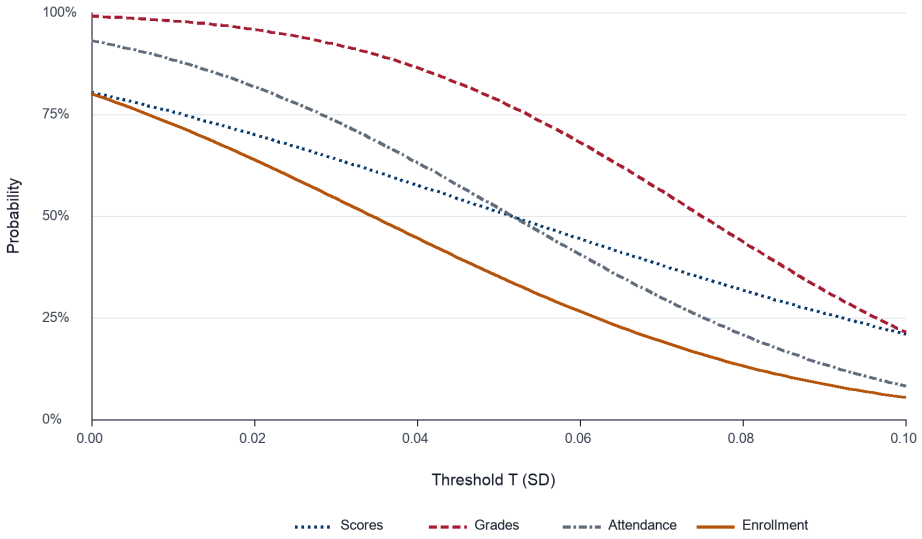


Figure 10: Probability that a new implementation effect exceeds threshold T

Notes: This figure plots a probability of an estimate surpassing a given threshold for each domain. For reference, T=0 is the financial break-even threshold, and T=0.05 is an alternative differentiated learning intervention threshold comparison.

A Examples of Excluded Studies

We excluded studies unrelated to remote interventions to boost parental engagement on student outcomes, and all those that did not satisfy our criteria in Section 2.1. Here we discuss a few well-known and seemingly related papers that we opted to exclude in detail.

Non-experimental design for the education outcomes (Condition a)

Although non-experimental designs can yield causal interpretation, we focused on the randomized controlled trial as it predominantly dominates the literature. For example, Greaves et al. (2023) study the impact of information via school ratings in England using difference-in-difference design. Kraft and Bolves (2022) use experimental design in providing application support, but the paper was excluded due to a weak first stage.

Interventions targeting students (Condition b)

Lichand et al. (2023) and Lichand et al. (2024) conduct an experiment in Brazil to prevent learning loss during the pandemic. However, text messages are sent to both high school students and their parents. As such, the two studies did not satisfy our criterion that interventions target parents exclusively.

Non-remote interventions (Condition c)

This condition restricted the sample to simple, low-cost, and remote interventions, such as text messages, phone calls, and report cards. We excluded studies that utilized more intensive methods such as parent-teacher meetings Avvisati et al. (2014) or monthly parent meetings and text messages Ome and Menendez (2022). studies in which staff members delivered information and instruction to parents at home, including Banerji et al. (2017).

Out-of-scope outcomes (Condition d)

Multiple studies use parental engagement initiatives to focus on other outcomes, such as children's socio-emotional development and parental investment. Nonetheless, we identified insufficient studies to conduct meta-analyses across these outcome categories. As such, we excluded studies with outcomes that we could not categorize, including education spending (Dizon-Ross, 2019, 2021), parent-child activities (Hurwitz et al., 2015; Mayer et al., 2019), and child labor reduction (Wolf and Lichand, 2023). Additionally, we excluded papers without the educational outcomes, such as Goyal et al. (2024).

Incomplete studies (Condition d)

In some cases, studies that met our inclusion criteria may nonetheless lack key information, such as standard deviation of outcomes necessary to compute standardized effect sizes or standard error of the effects. This criteria excludes several papers, for example Fujii et al. (2025); Robinson et al. (2018); Rogers et al. (2017); Sirvani (2007). The lack

of such details prevents us from correctly interpreting the precision or standardization of those estimates. Consequently, we did not include these studies in the analysis.

B Funnel Plot

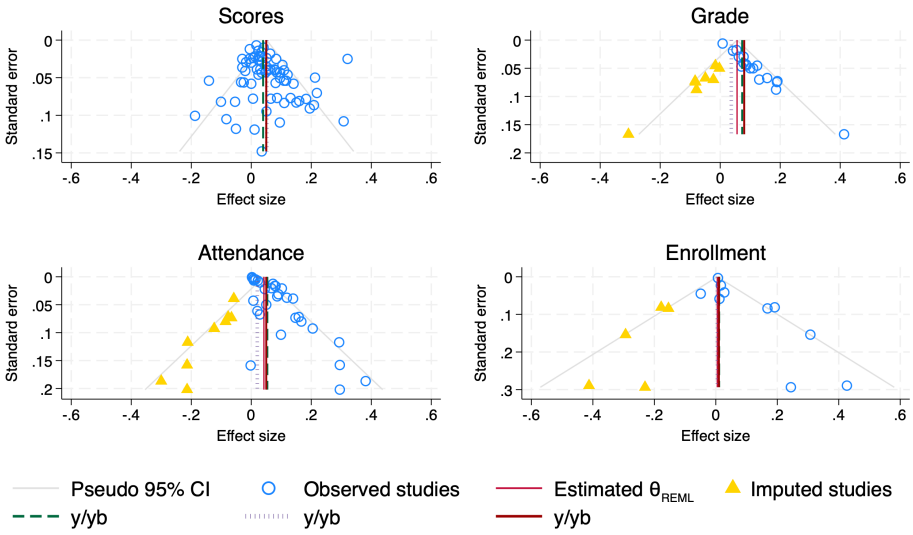
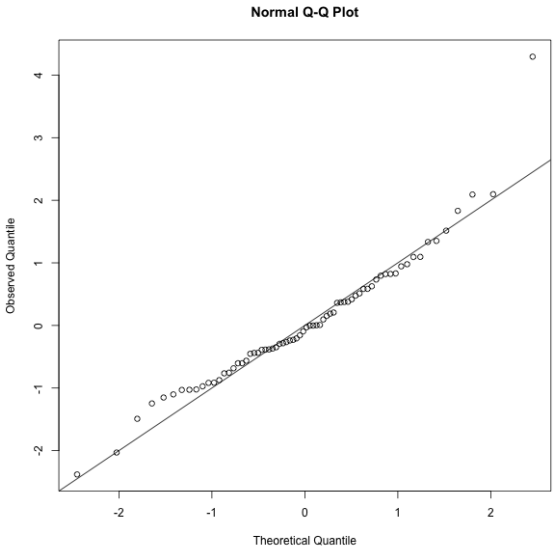


Figure 11: Funnel plot: Multiple estimates

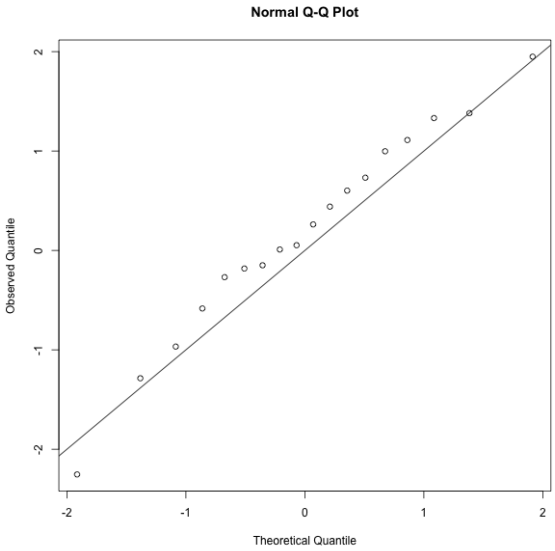
C Assessing Normality Assumption

Domain	Bandwidth
Scores	0.04
Grades	0.05
Attendance	0.06
Enrollment	0.14

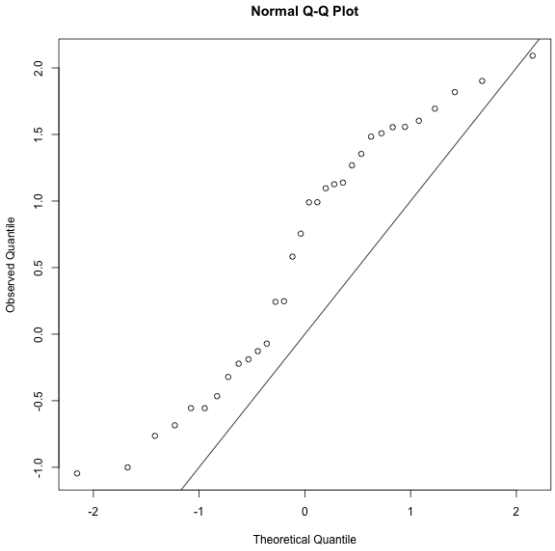
Table 10: Deconvolution bandwidth



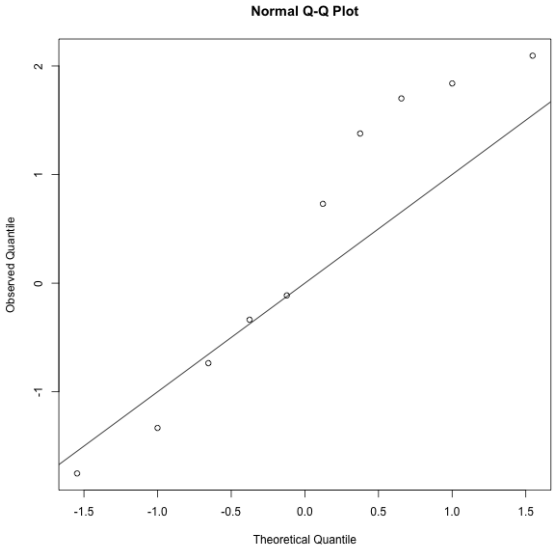
(a) Scores



(b) Grade



(c) Attendance



(d) Enrollment

Figure 12: QQ plot for multiple estimates

D Individual Bayes Estimate

Table 11: Empirical Bayes Estimates for Individual Observation

Study	n	x	σ	θ_{EB}	$\Pr(\theta_{EB} > 0)$
Scores $\hat{\mu}_d = 0.05$, $\hat{\tau}_d = 0.06$					
Afridia, Barooah, and Somanathan (2020)	1,338	0.31	0.11	0.11	0.98
Ajzenman et al. (2021)	2,727	0.03	0.04	0.04	0.87
Allende, Gallego, and Neilson (2019)	1,443	0.10	0.05	0.08	0.98
Allende, Gallego, and Neilson (2019)	776	0.22	0.07	0.12	1.00
Andrabi, Das, and Khwaja (2017)	112	0.11	0.04	0.09	1.00
Angrist, Bergman, and Matsheng (2022)	2,815	0.12	0.05	0.10	1.00
Angrist, Bergman, and Matsheng (2022)	2,815	0.02	0.05	0.03	0.83
Angrist et al. (2023)	8,902	0.32	0.03	0.28	1.00
Angrist et al. (2023)	9,148	0.04	0.01	0.04	1.00
Angrist et al. (2023)	8,902	0.08	0.03	0.08	1.00
Angrist et al. (2023)	9,148	-0.01	0.01	-0.00	0.40
Armstrong, Scherer, and Kim (2022)	3,961	0.03	0.02	0.04	0.99
Arteaga et al. (2024)	1,838	0.05	0.04	0.05	0.94
Barrera et al. (2020)	2,485	0.00	0.03	0.01	0.65
Barrera-Osorio et al. (2020a)	7,981	0.05	0.09	0.05	0.84
Barrera-Osorio et al. (2020b)	2,984	0.09	0.04	0.08	0.99
Barrera-Osorio et al. (2020b)	2,416	0.11	0.07	0.08	0.96
Barrera-Osorio et al. (2020b)	2,336	-0.08	0.07	-0.00	0.48
Barrera-Osorio et al. (2020b)	1,036	-0.02	0.10	0.03	0.74
Bergman (2021)	279	0.16	0.08	0.09	0.97
Bergman (2021)	256	0.11	0.08	0.07	0.93
Bergman and Chan (2021)	926	-0.03	0.04	-0.01	0.40
Bergman and Chan (2021)	70,076	0.06	0.02	0.06	1.00
Bergman and Chan (2021)	7,342	0.10	0.03	0.09	1.00
Bergman, Edmond-Verley, and Notario-Risk (2018)	703	-0.02	0.06	0.01	0.60
Bettinger et al. (2021)	12,577	0.08	0.04	0.07	0.98
Bettinger et al. (2021)	12,577	0.06	0.04	0.06	0.97
Bettinger et al. (2021)	12,577	0.14	0.06	0.10	0.99
Cabell et al. (2019)	174	-0.14	0.05	-0.05	0.09
Chamberlain et al. (2021)	275	0.21	0.09	0.10	0.98
Chinen and Bos (2016)	2,413	0.21	0.05	0.15	1.00
Cortes et al. (2023)	3,664	-0.01	0.03	0.00	0.54
Cortes et al. (2023)	3,664	0.03	0.02	0.03	0.92
Cortes et al. (2021)	2,920	0.02	0.04	0.03	0.81
Cortes et al. (2021)	2,920	-0.02	0.04	0.00	0.53
Crawfurd et al. (2023)	3,946	-0.02	0.03	-0.00	0.44
Doss et al. (2019)	578	0.18	0.08	0.10	0.98
Doss et al. (2019)	578	0.01	0.08	0.03	0.77
Doss et al. (2022)	1,336	0.00	0.06	0.02	0.72
Doss et al. (2022)	1,336	-0.03	0.06	0.01	0.56
Esposito and Sautmann (2022)	5,469	0.05	0.02	0.05	0.99
Gray-Lobe et al. (2024)	2,959	0.06	0.04	0.06	0.95
Gray-Lobe et al. (2024)	2,959	0.08	0.04	0.07	0.98
Gray-Lobe et al. (2024)	2,959	0.05	0.04	0.05	0.92
Heppen, Kurki, and Brown (2020)	2,393	0.03	0.03	0.03	0.89
Heppen, Kurki, and Brown (2020)	2,363	0.04	0.03	0.04	0.93
Heppen, Kurki, and Brown (2020)	2,436	0.03	0.03	0.03	0.88

Continued on next page

Study	n	x	σ	θ_{EB}	$\Pr(\theta_{EB} > 0)$
Heppen, Kurki, and Brown (2020)	2,437	0.04	0.02	0.04	0.96
Hernandez-Agramonte et al. (2024)	1,877	0.12	0.05	0.09	0.99
Holtzman, Quick, and Keuter (2023)	346	0.10	0.11	0.06	0.88
Kalil et al. (2023)	257	-0.19	0.10	-0.01	0.41
Kalil et al. (2023)	257	-0.08	0.11	0.02	0.64
Kraft and Monti-Nussbaum (2017)	779	0.15	0.08	0.09	0.96
Miller et al. (2016)	5,825	0.01	0.01	0.01	0.76
Riis-Vestergaard (2021)	2,981	-0.03	0.03	-0.02	0.22
Robinson et al. (2022)	2,201	0.03	0.02	0.03	0.93
Robinson-Smith et al (2019)	1,128	0.01	0.02	0.02	0.79
Siebert et al. (2018)	1,942	0.04	0.04	0.04	0.88
Siebert et al. (2018)	1,891	0.07	0.08	0.06	0.88
Snell, Wasik, and Hindman (2022)	309	0.17	0.07	0.10	0.99
Wolf and Lichand (2023)	2,246	0.08	0.06	0.07	0.95
Yedomiffi (2025)	6,177	0.05	0.03	0.05	0.95
York, Loeb, and Doss (2019)	821	0.11	0.05	0.08	0.98
de Walque and Valente (2023)	173	0.20	0.09	0.10	0.97
Grades $\hat{\mu}_d = 0.07$, $\hat{\tau}_d = 0.03$					
Bergman (2020)	19,218	0.10	0.05	0.08	1.00
Bergman (2021)	279	0.19	0.07	0.09	1.00
Bergman (2021)	279	0.19	0.07	0.09	1.00
Bergman (2021)	27,297	0.12	0.05	0.09	1.00
Bergman and Chan (2021)	1,137	0.07	0.05	0.07	1.00
Bergman, Edmond-Verley, and Notario-Risk (2018)	1,120	0.13	0.07	0.08	1.00
Bergman, Edmond-Verley, and Notario-Risk (2018)	1,120	0.16	0.07	0.09	1.00
Bergman and Rogers (2017)	6,291	0.05	0.02	0.06	1.00
Bergman and Rogers (2017)	6,291	0.04	0.02	0.05	1.00
Berlinski et al. (2025)	2,011	0.09	0.04	0.08	1.00
Bettinger et al. (2021)	12,577	0.06	0.03	0.07	1.00
Bettinger et al. (2021)	12,577	0.08	0.03	0.08	1.00
Bettinger et al. (2021)	12,577	0.08	0.04	0.08	1.00
Kraft and Rogers (2015)	521	0.19	0.09	0.09	1.00
Mabel et al. (2020)	160,269	0.01	0.01	0.01	0.96
Santana et al. (2019)	51	0.41	0.17	0.09	1.00
Yedomiffi (2025)	2,056	0.09	0.04	0.08	1.00
Yedomiffi (2025)	2,056	0.11	0.05	0.09	1.00
Attendance $\hat{\mu}_d = 0.05$, $\hat{\tau}_d = 0.03$					
Ajzenman et al. (2021)	4,098	0.01	0.01	0.01	0.94
Berger et al (2025)	78,732	0.12	0.04	0.08	1.00
Bergman (2021)	278	0.30	0.20	0.06	0.96
Bergman (2021)	2,252	0.30	0.16	0.06	0.97
Bergman (2021)	2,252	0.20	0.09	0.07	0.98
Bergman and Chan (2021)	1,137	0.17	0.08	0.07	0.99
Bergman, Edmond-Verley, and Notario-Risk (2018)	1,120	0.03	0.07	0.05	0.93
Bergman, Edmond-Verley, and Notario-Risk (2018)	1,120	0.02	0.06	0.04	0.93
Berlinski et al. (2025)	2,011	0.16	0.07	0.07	0.99
Bettinger et al. (2021)	12,577	0.01	0.00	0.01	1.00
Bettinger et al. (2021)	12,577	0.01	0.00	0.01	1.00
Bettinger et al. (2021)	12,577	0.02	0.01	0.02	1.00
Chibwana et al. (2023)	2,073	0.01	0.04	0.03	0.89
Dizon-Ross (2019)	541	-0.01	0.06	0.04	0.88
Gold et al (2025)	71,916	0.00	0.00	0.00	0.98
Heppen, Kurki, and Brown (2020)	6,631	0.10	0.02	0.09	1.00
Heppen, Kurki, and Brown (2020)	6,591	0.07	0.02	0.06	1.00

Continued on next page

Study	n	x	σ	θ_{EB}	$\Pr(\theta_{EB} > 0)$
Heppen, Kurki, and Brown (2020)	6,583	0.08	0.02	0.07	1.00
Heppen, Kurki, and Brown (2020)	6,564	0.08	0.02	0.08	1.00
Himmelsbach et al. (2022)	5,552	0.03	0.01	0.03	1.00
Kalil, Mayer, and Gallegos (2021)	741	0.15	0.07	0.07	0.99
Kraft and Rogers (2015)	27,037	0.29	0.12	0.07	0.98
Miller et al. (2016)	7,434	0.00	0.00	0.00	1.00
Musaddiq, Prettyman, and Smith (2020)	22,883	0.09	0.03	0.07	1.00
Musaddiq, Prettyman, and Smith (2020)	20,508	0.05	0.05	0.05	0.96
Musaddiq, Prettyman, and Smith (2020)	22,883	0.09	0.04	0.07	1.00
Rogers and Feller (2018)	7,037	0.07	0.01	0.07	1.00
Swanson (2023)	7,656	0.10	0.10	0.06	0.96
Swanson (2023)	3,887	-0.00	0.16	0.05	0.93
Swanson et al. (2025)	41,468	0.04	0.02	0.05	0.99
Yedomiffi (2025)	2,085	0.14	0.04	0.09	1.00
de Walque and Valente (2023)	173	0.38	0.19	0.06	0.97
Enrollment $\hat{\mu}_d = 0.03$, $\hat{\tau}_d = 0.04$					
Ajayi, Friedman, and Lucas (2020)	7,087	-0.05	0.04	-0.00	0.46
Andrabi, Das, and Khwaja (2017)	112	0.19	0.08	0.07	0.97
Barrera-Osorio et al. (2020a)	180	0.31	0.15	0.05	0.91
Bergman, Edmond-Verley, and Notario-Risk (2018)	1,120	0.01	0.06	0.03	0.79
Dizon-Ross (2019)	579	-0.03	0.05	0.01	0.62
Kraft and Rogers (2015)	521	0.17	0.08	0.06	0.95
Weixler et al. (2020)	1,407	0.43	0.29	0.04	0.85
Weixler et al. (2020)	1,407	0.24	0.29	0.04	0.83
Yedomiffi (2025)	2,068	0.03	0.04	0.03	0.86

Notes: Empirical Bayes posterior means combine each written estimate x and standard error σ with the domain-level prior distribution from the model, $\theta_{id} \sim N(\hat{\mu}_d, \hat{\tau}_d^2)$. θ_{EB} is the posterior mean for the latent effect for that observation, and $\Pr(\theta_{EB} > 0)$ is the posterior probability that the latent effect is positive. Domain header rows report the plug-in $\hat{\mu}_d$ and $\hat{\tau}_d$ used for shrinkage.

E File Drawer Distribution

As the precision model is conditional on n, we must also marginalize over the sample-size distribution. In general, this would be

$$f_d(x, D = 1) = \int f_d(x, D = 1 \mid n)\, dF_d(n), \qquad f_d(x, D = 0) = \int f_d(x, D = 0 \mid n)\, dF_d(n).$$

We use the empirical distribution of sample sizes in the estimation sample. This keeps the calculation anchored to the observed sample-size composition within each outcome domain.

Let $n^{(m)}$, $m = 1, \ldots, M_d$, denote the distinct sample sizes observed in domain d. Let

$$w_m = \frac{\sum_i \mathbb{1}\{n_i = n^{(m)}\}}{N_d}$$

be the empirical frequency of sample size $n^{(m)}$, where N_d is the number of outcomes in domain d. Equivalently, this computation averages over the sample sizes n_i attached to the study-domain observations in the estimation sample. As the same study can contribute multiple outcomes with the same sample size, some n_i are repeated. This is algebraically to summing over observations one by one.

The empirical analogue of the marginalization over n is therefore

$$f_d(x, D=1) = \sum_{m=1}^{M_d} w_m f_d(x, D=1 \mid n^{(m)}), \qquad f_d(x, D=0) = \sum_{m=1}^{M_d} w_m f_d(x, D=0 \mid n^{(m)}).$$

In implementation, we evaluate the σ-integrals using the change of variables $u \sim N(0,1)$:

$$\sigma(u; n^{(m)}) = \exp\Big(\widehat{a}_d + \widehat{b}_d \log n^{(m)} + \widehat{s}_d u\Big).$$

Define

$$v(u; n^{(m)}) = \widehat{\tau}_d^2 + \sigma(u; n^{(m)})^2, \qquad z = \frac{x}{\sigma(u; n^{(m)})}.$$

Then the empirical domain-level joint densities can be written as

$$f_d(x, D=1) = \sum_{m=1}^{M_d} w_m \int p_d(z)\, \phi\big(x; \widehat{\mu}_d, v(u; n^{(m)})\big)\, \phi(u)\, du,$$

and

$$f_d(x, D=0) = \sum_{m=1}^{M_d} w_m \int [1 - p_d(z)]\, \phi\big(x; \widehat{\mu}_d, v(u; n^{(m)})\big)\, \phi(u)\, du.$$

The conditional densities plotted in the figure are obtained by normalizing these joint densities by the model-implied marginal write-up probabilities:

$$f_d(x \mid D=1) = \frac{f_d(x, D=1)}{\Pr_d(D=1)}, \qquad f_d(x \mid D=0) = \frac{f_d(x, D=0)}{\Pr_d(D=0)}.$$

Both curves are integrated over the same empirical distribution of sample sizes. This holds the baseline sample-size composition fixed, so differences between the written and unwritten curves reflect the fitted write-up selection mechanism and its interaction with precision, rather than raw differences in the observed sample-size composition of written and unwritten observations.

F Robustness to Additional Unwritten Studies

Joint Likelihood Model

Let the overall likelihood be:

$$\ell_d(\theta_d;\kappa) = \sum_{i:D_i=1} \log L_{id}^{(1)}(\theta_d) + \kappa \sum_{i:D_i=0} \log L_{id}^{(0)}(\theta_d).$$

For each value of κ, we re-estimate the full model:

$$\widehat{\theta}_d(\kappa) = \arg\max_{\theta_d \in \Theta_{S2}} \ell_d(\theta_d;\kappa),$$

All parameters are re-estimated at each value of κ, including the corrected mean μ_d, heterogeneity τ_d, the precision-model parameters, and the selection probabilities.

The implied unwritten share under each expansion factor is given by

$$m_d(\kappa) = \frac{\kappa N_{0d}}{N_{wd} + \kappa N_{0d}},$$

where N_{wd} is the number of written observations and N_{0d} is the number of unwritten observations in domain d. The plotted series is

$$\widehat{\mu}_d(\kappa),$$

the corrected mean effect obtained after re-estimating the model at that assumed unwritten share.

Andrews and Kasy model Benchmark

We calculate AK-implied missing-share as a benchmark. As this benchmark is based on a model with no information on the absolute write-up rate, we cannot interpret it as an estimate of the true missing share. Instead, it provides a benchmark for how large the file drawer would need to be for the AK selection ratio to hold.

Since the AK model identifies only relative write-up probabilities, we normalize $q_{hd} = 1$ and use the estimated relative selection probability to construct inverse-probability

weights for written studies:

$$\widehat{p}_{id}^{AK} = \begin{cases} \widehat{q}_{\ell d}^{AK}, & |z_i| < 1.96, \\ 1, & |z_i| \geq 1.96. \end{cases}$$

For a given domain, the AK-implied latent evidence-base size is

$$\widehat{N}_d^{AK} = \sum_{i:D_i=1} \frac{1}{\widehat{p}_{id}^{AK}},$$

and therefore the AK-implied missing share is

$$\widehat{m}_d^{AK} = \frac{\widehat{N}_d^{AK} - N_{wd}}{\widehat{N}_d^{AK}}.$$