Publication bias and p-hacking in the effect of COVID-19 on learning*

Martina Luskova, Nino Buliskeria, Ali Elminejad, Tomas Havranek, Zuzana Irsova, Stepan Jurajda, Marek Kapicka

June 13, 2026

Abstract

We revisit a central estimate in the economics of education: the human-capital loss associated with COVID-19 school closures. Estimates of pandemic learning loss may be affected by publication bias, p-hacking, and the mechanical correlation between standardized effect sizes and their standard errors. We conduct a comprehensive multi-method assessment of bias by applying a wide range of correction techniques – including PET-PEESE, three-parameter selection models (3PSM), Robust Bayesian Meta-Analysis (RoBMA), Meta-Analysis Instrumental Variable Estimation (MAIVE), Right-Truncated Meta-Analysis (RTMA), and multi-bias sensitivity analysis. Our preferred specifications, RoBMA and MAIVE, rely on different assumptions yet converge on an effect size of approximately −0.12 SD, equivalent to a learning loss of about 30% of a school year. Although some methods reveal signs of publication bias and selective reporting, these findings do not explain away the central finding: the COVID-19 learning deficit is economically meaningful and statistically robust.

KEYWORDS: Meta-analysis; Publication bias; p-Hacking; COVID-19; Learning loss; Economics of education; Human capital

JEL CODES: I21, I24, I28, C18

*Authors: Luskova: Institute of Economic Studies, Faculty of Social Sciences, Charles University. E-mail: martina.luskova@fsv.cuni.cz. Buliskeria, corresponding author: Department of Economics, Nazarbayev University. E-mail: nino.buliskeria@nu.edu.kz. Elminejad: Department of Economics, Nazarbayev University. E-mail: ali.elminejad@nu.edu.kz. Havranek: Institute of Economic Studies, Faculty of Social Sciences, Charles University; also affiliated with METRICS, Stanford University, and CEPR, London. E-mail: tomas.havranek@fsv.cuni.cz. Irsova: Institute of Economic Studies, Faculty of Social Sciences, Charles University; also affiliated with METRICS, Stanford University, and Anglo-American University, Prague. E-mail: zuzana.irsova@fsv.cuni.cz. Jurajda: CERGE-EI, a joint workplace of the Center for Economic Research and Graduate Education of Charles University and the Economics Institute of the Czech Academy of Sciences. E-mail: stepan.jurajda@cerge-ei.cz. Kapicka: CERGE-EI, a joint workplace of the Center for Economic Research and Graduate Education of Charles University and the Economics Institute of the Czech Academy of Sciences. E-mail: marek.kapicka@cerge-ei.cz. The authors declare no conflicts of interest related to this work.

Acknowledgment: We thank Bastian A. Betthäuser, Anders M. Bach-Mortensen, and Per Engzell for helpful comments. Tomas Havranek and Zuzana Irsova acknowledge support from the Czech Science Foundation, project 24-11583S. Martina Luskova acknowledges support from the Charles University Grant Agency, project No. 136724.

1 Introduction

The COVID-19 pandemic triggered an unprecedented disruption to formal education worldwide. School closures, shifts to remote learning, and broader social disruptions affected hundreds of millions of students across all income levels and regions. As schools reopened and assessment data accumulated, a large empirical literature emerged documenting what is now widely referred to as "pandemic learning loss" — declines in student achievement relative to pre-pandemic benchmarks or counterfactual trends (United Nations 2020, Engzell et al. 2021, Donnelly and Patrinos 2022). Understanding the magnitude of these losses directly informs remediation policy design, public expenditure decisions, and the longer-run assessment of how major systemic shocks accumulate into human-capital deficits.

For economists, the magnitude of pandemic learning loss matters because test-score deficits are widely used as proxies for losses in human-capital accumulation, with implications for inequality, remediation spending, and long-run productivity (Hanushek and Woessmann 2008, 2020). The policy question is therefore not only whether students learned less during the pandemic, but whether the estimated magnitude is robust enough to guide large-scale catch-up interventions. Answering this question requires assessing whether the reported effect sizes are robust to publication bias, selective reporting, and the statistical complications introduced by standardized effect-size measures.

Several meta-analyses have quantified the extent of pandemic learning loss (König and Frey 2022, Betthäuser et al. 2023a, Di Pietro 2023, Wisenöcker et al. 2025). These syntheses report average deficits of roughly -0.14 to -0.20 standard deviations (Cohen's d). Where moderator analyses are available, losses tend to be larger in mathematics, and several studies also report larger losses among disadvantaged students. We take as our empirical benchmark the most prominent and widely cited synthesis, Betthäuser et al. (2023a), which conducts a meta-analysis of 291 effect-size estimates from 42 studies across 15 countries and reports an average learning loss of approximately -0.14 standard deviations (Cohen's d). Following Betthäuser et al. (2023a), we interpret this magnitude

using the common benchmark that students typically gain about 0.40 standard deviations per school year under normal circumstances (Azevedo et al. 2021, Bloom et al. 2008, Hill et al. 2008) – this corresponds to roughly 35% of a school year. Existing assessments of publication bias in this literature, including Betthäuser et al. (2023a), have relied primarily on visual or conventional diagnostic tools, including funnel plots, Egger's test, Doi plots, distributions of z-statistics, and p-curves. These assessments generally conclude that publication bias is not a major concern.

Although these diagnostics are widely used and provide useful first-pass evidence, they remain limited in their ability to model selection directly or to quantify how sensitive the pooled estimate is to selective reporting. Moreover, expressing estimates as Cohen's d can mechanically induce a correlation between standardized effect sizes and their standard errors. This issue is not specific to any individual study, but arises from the construction of standardized effect-size measures, which is standard practice across the social sciences. We therefore extend the existing assessments by applying regression-based, selection-model, instrumental-variable, and sensitivity-analysis approaches that provide a more formal evaluation of publication bias, p-hacking, and their implications for the estimated magnitude of pandemic learning loss.

This paper provides a systematic econometric assessment of bias in the COVID-19 learning-loss literature. Using the dataset of Betthäuser et al. (2023a), we apply a wide range of state-of-the-art bias-detection and correction techniques: the Precision Effect Test and Precision Effect Estimate with Standard Errors (PET-PEESE), the Three-Parameter Selection Model (3PSM), Robust Bayesian Meta-Analysis (RoBMA), Meta-Analysis Instrumental Variable Estimation (MAIVE), Right-Truncated Meta-Analysis (RTMA), and multi-bias sensitivity analysis. These methods differ in their assumptions, their sensitivity to heterogeneity, and their ability to distinguish p-hacking from publication bias. Crucially, RoBMA and MAIVE are less reliant on funnel-based assumptions in settings where standardized effect sizes, such as Cohen's d, mechanically induce dependence between effect-size estimates and their standard errors: MAIVE addresses this dependence directly by instrumenting precision, while RoBMA avoids committing to the

funnel-asymmetry slope that this dependence distorts. The paper therefore contributes to the economics literature in two ways. Substantively, it reassesses a central estimate in the economics of education: the human-capital loss associated with COVID-19 school closures. Methodologically, it shows how publication-bias, p-hacking, and endogenous-precision corrections behave in a high-powered applied literature based on standardized effect sizes. Because we use the same dataset as Betthäuser et al. (2023a), computationally reproduce their main results — as documented in Appendix A — and reanalyze the same research question using additional bias-correction methods, the paper also constitutes a reproducibility and replication exercise in the spirit of the meta-science literature.

Our findings support the original conclusion, but on broader methodological grounds. Across correction methods, bias-adjusted estimates consistently point to a meaningful decline in student achievement. Our preferred estimates from RoBMA and MAIVE converge on an effect size of approximately -0.12, equivalent to a learning loss of roughly 30% of a school year. Although we find evidence consistent with selective reporting that is not apparent from graphical diagnostics alone, these biases do not overturn the central finding: COVID-19 learning losses are robust and substantively meaningful. The modest downward revision from approximately -0.14 to -0.12 does not materially alter the policy implications of the original estimate: a deficit of this magnitude still represents a non-trivial and potentially persistent loss in human-capital accumulation for the affected cohorts.

The remainder of the paper is organized as follows. Section 2 describes the dataset, summarizes key patterns of heterogeneity, and presents descriptive diagnostics that motivate the formal bias analysis. Section 3 applies publication-bias correction and sensitivity methods, including PET-PEESE, 3PSM, RoBMA, MAN, selection-ratio adjustments, and s-values. Section 4 turns to methods that additionally address within-study selection, endogenous precision, and multiple sources of bias, including MAIVE, RTMA, and multi-bias sensitivity analysis. Section 5 concludes by summarizing the evidence and discussing its implications for the robustness of estimated COVID-19 learning losses.

2 Background and Data

Our analysis builds on the dataset compiled by Betthäuser et al. (2023a). We use the same set of studies and effect-size estimates, so any difference between our findings and theirs reflects differences in aggregation and bias-correction methods rather than differences in study selection or effect-size construction. The dataset is publicly available via the Open Science Framework (Betthäuser et al. 2023b) at https://doi.org/10.17605/osf.io/u8gaz. Replication code for all analyses reported in this paper is available at https://meta-analysis.cz/learning. Throughout the analysis, we follow recent guidance for meta-analysis in economics provided by Irsova et al. (2024) and Cook et al. (2026a), as well as the reporting standards proposed by Havránek et al. (2020) and Cook et al. (2026b). Appendix D documents how the study addresses these recommendations item by item.

The dataset contains 291 effect-size estimates drawn from 42 studies across 15 countries (Figure 1). It is geographically concentrated and dominated by large administrative records: roughly half of all estimates come from the United States, while the United Kingdom and the Netherlands account for most of the remainder. Study sizes vary substantially, from entire national school populations numbering in the millions to samples of only a few hundred students.

The data reveal substantial heterogeneity across key dimensions (Tables 1 and 2): mathematics estimates are markedly larger in magnitude than reading estimates, consistent with evidence from other educational disruptions that mathematics learning depends more heavily on formal instruction (Figure 2, panel a). By contrast, there is little evidence of a systematic grade-level gradient: primary and secondary school estimates are nearly identical on average (Figure 2, panel b), and a finer-grained breakdown by individual grade confirms that effect sizes remain close to the overall mean across grades 1–9 (Figure 3). The more negative estimates in grades 11–13 should be interpreted cautiously because they are based on very few observations, and the confidence band is correspondingly wide. Each study receives equal total weight in the subgroup averages,

so that prolific studies, i.e., studies producing multiple effect estimates, do not dominate the pooled estimates, and study-level clustering accounts for the resulting within-study dependence throughout. Our corpus spans peer-reviewed articles, working papers, and institutional and commercial reports (Table 2); the relevant selection therefore operates across reporting outlets rather than through journal acceptance alone, and we use "publication bias" throughout in the broad sense of selective *availability* of results.

The estimates in the dataset are all expressed as Cohen's d, which is necessary to aggregate results across studies using different assessment instruments. This standardization has an important methodological consequence, to which we return in Section 3: because both the standardized effect size and its standard error are functions of the same underlying standard deviation, they may be mechanically correlated across studies. This dependence complicates the interpretation of funnel-plot-based publication-bias diagnostics, which rely precisely on the relationship between reported effects and their precision.

We begin with descriptive diagnostics of the structure of the data. Figure 4 presents a conventional funnel plot with effect sizes on the horizontal axis and estimated standard errors on the vertical axis. The dotted vertical line marks the pooled mean effect, while the dashed vertical line marks zero. Most estimates are negative and lie to the left of zero, consistent with an overall learning deficit. The most precise estimates, located near the bottom of the figure, cluster around modestly negative effects, while less precise estimates display substantially greater dispersion. The cluster of highly precise estimates near $SE \approx 0$ reflects the presence of very large studies in the dataset. Figure 5 separates statistically significant and non-significant estimates: significant estimates extend farther into the left tail, whereas non-significant estimates cluster closer to zero. Figure 6 presents the corresponding significance funnel, which we examine in Section 3. This visual pattern is consistent with the possibility of selective reporting favoring negative and statistically significant learning-loss estimates, but it does not provide a clean visual basis for distinguishing publication bias from genuine heterogeneity or from mechanical dependence introduced by standardizing estimates as Cohen's d. It instead motivates the formal bias-correction and sensitivity analyses reported below.

We therefore also examine threshold-based diagnostics. Figures 7 and 8 plot the distribution of z-statistics. The log-transformed distribution does not show clear clustering around the conventional 1.96 significance threshold, consistent with Betthäuser et al. (2023a). In the raw distribution, however, the z-statistics show visible mass near the negative 5% significance threshold and a long left tail. This pattern is consistent with a preference for negative statistically significant estimates, but it is not dispositive: bunching near critical values can also arise when many studies have similar power or sample sizes (Elliott et al. 2022b).

We additionally apply a caliper test and attempt a p-curve analysis to assess whether the data show evidence of local threshold manipulation. The caliper test finds no robust evidence of excess mass just inside the negative 5% significance threshold, $z = -1.96$, with only weak and non-monotonic significance at wider calipers. The p-curve is also not informative in this setting: the distribution is highly compressed near zero because the literature includes very large studies, with a median sample size of $n = 50{,}000$ and a maximum sample size of $n = 10{,}884{,}922$. These sample sizes generate extremely small standard errors, so even moderate effect sizes can produce very large z-statistics. Full results are reported in Appendix C. Thus, the threshold diagnostics provide little evidence of local threshold manipulation, even though some broader features of the distribution remain consistent with selective reporting.

As a further robustness check, we examine case-level influence using DFBETAS, which measures how much the pooled estimate changes when each observation is omitted in turn.[1] Under the sample-size-adjusted threshold $|\text{DFBETAS}| > 2/\sqrt{n} = 0.117$, eleven estimates are flagged: nine with negative DFBETAS values (range: -0.28 to -0.12), pulling the pooled estimate in the negative direction, and two with positive values ($+0.14$ and $+0.15$), pulling in the opposite direction. The net effect is negligible: excluding all eleven moves the pooled estimate from -0.124 to -0.116, confirming that no individual estimate materially drives the main result. The two most influential estimates — those from Ardington et al. (2021) (DFBETAS = -0.28 and -0.22) — report learning losses

[1] Influence diagnostics require a standard random-effects model; we use the REML estimate ($\hat{\mu} = -0.124$).

from South Africa, one of the few middle-income countries in the dataset and among those with the most severe school disruptions during the pandemic. Their influence is consistent with genuine heterogeneity rather than measurement error. The DFBETAS plot is presented in Figure 9.

Taken together, these figures motivate rather than settle the bias question. The estimates are predominantly negative, and some descriptive patterns are consistent with selective reporting, but visual and threshold-based diagnostics alone are insufficient in this setting. The following sections therefore apply regression-based, selection-model, instrumental-variable, right-truncated, and sensitivity-analysis methods that more directly evaluate publication bias and p-hacking.

3 Correcting for publication bias

As a benchmark, Panel A of Table 3 presents the standard uncorrected point estimates: an equal-weighted (unweighted) mean of -0.126 ($SE = 0.059$) and a robust variance estimation (RVE) estimate of -0.140 ($SE = 0.020$) (Hedges et al. 2010), which accounts for within-study dependence by clustering at the study level. These estimates align with Betthäuser et al. (2023a) precisely because they represent the uncorrected mean, assuming no selection bias. All correction methods below are assessed relative to this uncorrected mean of -0.140.

We begin with the funnel-based precision effect test (PET-PEESE). PET is a weighted regression of the effect on its standard error, with weights $1/SE^2$; the significant coefficient on the standard error is consistent with funnel asymmetry (Table 3, Panel B, Column 1). PEESE regresses the effect on the squared standard error and provides a more accurate bias-corrected estimate (Stanley 2017, Bartoš et al. 2022), yielding a corrected estimate of -0.245, corresponding to $0.245/0.4 = 0.61$ school years of learning deficit (Table 3, Panel B, Column 2). However, as noted in Section 2, the standardization to Cohen's d mechanically links estimates and standard errors across studies, which means funnel-based results should be interpreted cautiously. We address this directly in Section 4.

A selection-model approach yields a more moderate correction. The three-parameter selection model (3PSM) applies maximum likelihood to correct for the preferential publication of p-values below the significance threshold (Iyengar and Greenhouse 1988, Hedges 1992, Vevea and Hedges 1995). The corrected estimate is -0.123, equivalent to $0.123/0.4 = 0.31$ school years of learning loss. The associated likelihood-ratio test does not reject the null hypothesis of no publication bias (Table 3, Panel B, Column 3). Relative to PEESE, this estimate is considerably closer to the uncorrected mean of -0.140, suggesting that conclusions about the magnitude of bias depend materially on the correction method.

Finally, we apply Robust Bayesian Meta-Analysis (RoBMA), which constructs a model-averaged estimate across multiple publication bias models, weighting by data fit and parsimony (Maier et al. 2023, Bartoš et al. 2023). The RoBMA estimate is -0.118, or $0.118/0.4 = 0.30$ school years (Table 3, Panel B, Column 4). Because RoBMA integrates over uncertainty about the appropriate bias model rather than committing to one, it is robust to misspecification and performs well under heterogeneity. We treat this as our preferred bias-corrected estimate from this section. The sensitivity of this estimate to increasingly severe forms of selective publication is assessed in the following paragraphs, and further corroborated by regression-based correction methods robust to p-hacking in Section 4.

Sensitivity to selective publication. The correction methods in Table 3 each rely on a specific model of the selection process. We complement them with a sensitivity-based approach that asks how robust the findings are to a range of assumed degrees of selective publication, following Mathur and VanderWeele (2020), Mathur (2024a,b). We implement these analyses using the `PublicationBias` and `phacking` packages available at metabias.io. These methods assess how sensitive conclusions are to the selective reporting of affirmative results — those that are statistically significant in the expected direction — relative to non-affirmative results, which include non-significant or unexpected-direction estimates. The degree of this asymmetry is quantified by the *selection ratio*: the relative likelihood of an affirmative results being published compared to a non-affirmative one. A selection ratio of one indicates no preference for significant results. In our setting, an

affirmative result is one that is negative and statistically significant, since learning deficit is measured as a negative value.[2] Because the true degree of publication bias is unknown, we assess robustness across a range of assumed selection ratios rather than committing to one value.

Meta-analysis of non-affirmative results (MAN). In the extreme case where affirmative studies are infinitely more likely to be published than non-affirmative ones, the worst-case bias-corrected estimate is obtained from a meta-analysis restricted to non-affirmative results only. As Mathur (2024a) notes, MAN does not measure the actual strength of publication bias; rather, it is a stress test that assesses how results hold up under the most extreme possible form of selective publication. Since it relies only on non-affirmative results, it is typically biased toward the null when affirmative results are selectively favored.[3] In our setting, MAN asks whether the negative learning-loss effect survives after placing all weight on the least favorable subset of estimates, and is therefore best interpreted as a deliberately conservative robustness check rather than as a preferred corrected estimate. If the worst-case estimate still aligns with the uncorrected mean in sign and magnitude, this provides strong evidence of robustness. Panel A, Column 1 of Table 4 reports a MAN estimate of 0.021, which is statistically significant at the 10% level but not at the 5% level. This estimate reverses sign relative to the uncorrected random-effects mean of -0.140, suggesting that the results are sensitive to worst-case selective publication and motivating the examination of less extreme scenarios.

Selection ratio sensitivity. Following Mathur and VanderWeele (2020), Mathur (2024b), we use a four-fold selection ratio as an illustrative benchmark, where affirmative studies are assumed to be four times more likely to be published than non-affirmative ones. Under this assumption, both the fixed- and random-effects estimates retain the negative sign of the uncorrected mean (Table 4, Panel A, Columns 2 and 3). The fixed-

[2] In Figures A2 through A4 and Figure 6, most coefficient estimates are negative, as expected. We account for the negative sign and adjust model specifications accordingly, assuming one-directional selection that favors negative, statistically significant estimates, using a two-sided significance threshold of $\alpha = 0.05$ (i.e. $z < -1.96$).

[3] Mathur (2024b) shows that MAN remains conservative under p-hacking when p-hacking favors affirmative outcomes, because it places no weight on affirmative estimates, which are the estimates favored by such selection.

effects estimate is larger in magnitude, at −0.206, while the robust random-effects estimate is smaller, at −0.068, reflecting its adjustment for heterogeneity and clustering of estimates within studies. Thus, even under a relatively strong assumed degree of selective publication, the corrected estimates continue to indicate learning losses, although their magnitude depends on how heterogeneity and within-study clustering are handled. The range between these estimates also contains our preferred bias-corrected estimates from RoBMA (−0.118) and MAIVE (−0.119, reported in Section 4), suggesting that those estimates are consistent with a moderate rather than worst-case degree of selective publication.

***s*-Value.** We next invert the sensitivity exercise and ask how strong selective publication would have to be to explain away the result. The *s*-value is the selection ratio required to shift the estimated effect or its confidence interval bound to a specified value. Small *s*-values indicate that a relatively modest degree of selective publication could explain away the result, whereas large *s*-values indicate greater robustness. In our case, affirmative studies would need to be 29.60 times more likely to be published than non-affirmative studies to shift the point estimate to zero, and 8.31 times more likely to shift the confidence interval bound to zero (Table 4, Panel B, Column 1). Moreover, no degree of selection bias under this model can shift either the point estimate or the confidence interval bound to +0.05 (Table 4, Panel B, Column 3). These results suggest that, although the MAN estimate indicates sensitivity to the most extreme form of selective publication, the main directional conclusion would require a very large degree of selective publication to be overturned and is therefore robust to plausible degrees of publication bias.

Significance funnel plot. As a supplement to the selection-ratio sensitivity analysis, we present the significance funnel plot in Figure 6. The figure separates affirmative estimates (orange points) from non-affirmative estimates (gray points) and visualizes the extent to which the mean estimate depends on affirmative results. The contrast between the two groups is substantial. The gray diamond shows the mean among non-affirmative estimates (0.021), corresponding to the MAN estimate reported in Table 4, Panel A,

Column 1. The black diamond[4] shows the precision-weighted (inverse-variance) pooled mean across all 291 estimates ($\hat{\mu}_{IVW} = -0.245$), which lies more negative than the equal-weighted mean of -0.126 reported in Panel A of Table 3 because inverse-variance weighting gives disproportionate influence to large, high-precision studies that tend to report larger learning losses. The gap between the black and gray diamonds is consistent with some degree of selective publication and suggests that the magnitude of the pooled effect may be exaggerated (Mathur and VanderWeele 2020). At the same time, the non-affirmative mean should not be interpreted as the most credible corrected estimate of the underlying effect. It is the visual counterpart to the MAN worst-case benchmark. The figure therefore reinforces the MAN result: the evidence is sensitive to worst-case selective publication, but this stress test is intentionally severe.

Taken together, the sensitivity analyses suggest that the results are sensitive to the most extreme form of selective publication, but that overturning the negative directional finding would require a very large degree of bias. The MAN estimate therefore serves as a severe stress test rather than as our preferred corrected estimate. By contrast, the preferred RoBMA and MAIVE estimates, reported above and in Section 4, respectively, are consistent with the range implied by finite selection-ratio adjustments, suggesting that the main conclusion is robust to plausible degrees of publication bias.

4 Accounting for *p*-hacking and multiple biases

The preceding section evaluates sensitivity to selective publication across studies. We now consider whether the conclusion changes once we account for biases that may arise within studies before estimates enter the meta-analysis. Following Mathur (2024b), we distinguish publication bias, or selection across studies, from p-hacking, or selection within

[4] The black diamond corresponds to the inverse-variance weighted mean from Mathur and VanderWeele (2020), estimated as a fixed-effects model ($\hat{\mu}_{IVW} = -0.245$, SE = 0.0001). The unrestricted weighted least squares (UWLS) estimator (Stanley et al. 2023, Stanley and Doucouliagos 2017), which uses the same inverse-variance weights but does not impose distributional assumptions on the residuals, yields an identical point estimate with a more conservative cluster-robust standard error (SE = 0.051, $p < 0.001$). The large discrepancy in standard errors reflects the high degree of heterogeneity in the dataset ($I^2 = 99.97\%$), which the fixed-effects model ignores but the cluster-robust UWLS standard error partially accounts for. Both estimates are statistically significant; we report the fixed-effects version following Mathur and VanderWeele (2020) and note the UWLS equivalence in the interest of transparency.

studies. In practice, p-hacking may involve searching across specifications, samples, covariate sets, or outcomes until an affirmative result is obtained. Because such decisions can affect both the measured underlying quantity and the reported precision, p-hacking can distort the primary-study estimate itself, rather than simply determining which otherwise unbiased estimate is written up or published.[5]

These concerns matter because conventional publication-bias corrections typically assume that primary-study estimates are, individually, unbiased for their corresponding study-specific population effects. In that framework, bias enters through the selection mechanism: some otherwise unbiased estimates are more likely than others to be written up, submitted, or accepted for publication. When p-hacking or spurious precision is present, however, the estimates, their reported standard errors, or both may already be distorted by within-study research choices.

We therefore proceed in three steps. First, we apply MAIVE, which instruments for potentially endogenous precision and is especially relevant when standardized effect sizes mechanically link estimates and standard errors (e.g., Cohen's d). Second, we apply RTMA, which directly models p-hacking as selection within studies. Third, we apply the multi-bias framework of Mathur (2024c), which allows publication bias to interact with internal study bias by allowing affirmative and non-affirmative estimates to differ in their probability of publication and in their average degree of internal bias.

MAIVE. We first apply the Meta-Analysis Instrumental Variable Estimator (MAIVE) of Irsova et al. (2025). MAIVE is designed for settings in which reported standard errors may themselves be endogenous. It has increasingly been used in recent meta-analyses in economics to address spurious precision, including applications in labor economics, education, and microeconomics (Havranek et al. 2024, Opatrny et al. 2025, Cala et al. 2026). Conventional funnel-based methods, such as PET and PEESE, treat precision as observed and place greater weight on estimates with smaller standard errors. In empirical research, however, precision is estimated by researchers and may be affected by specification choices, sample restrictions, or other research decisions. If these choices affect

[5]For detailed discussions of p-hacking, see Brodeur et al. (2020), Elliott et al. (2022a), Brodeur et al. (2023), Mathur and VanderWeele (2020), Mathur (2024b).

both the reported coefficient and its standard error, then the standard error is endogenous, and methods relying on the relationship between estimates and standard errors may themselves become biased (Brodeur et al. 2023, Irsova et al. 2025, Mathur 2024b).

MAIVE addresses this problem by instrumenting for the reported variance. Following Irsova et al. (2025), we use inverse sample size as an instrument for the squared standard error. The relevance condition is motivated by the mechanical relationship between sampling variance and sample size: all else equal, larger samples imply lower sampling variance. The exclusion restriction is motivated by the fact that sample size is generally harder to manipulate ex post than reported precision. Researchers may be able to affect standard errors through specification choices, sample restrictions, clustering decisions, or alternative outcome definitions, but the underlying sample size is often determined before estimation, especially in observational and administrative-data settings. Sample size is also directly observed and does not suffer from measurement error, whereas the reported standard error is itself an estimate and may vary with methodological choices. A possible limitation is that sample size may still be endogenous if researchers who anticipate smaller effects design larger studies. In our setting, however, many estimates come from observational or administrative data in which sample size is largely predetermined.

This approach is particularly relevant here because all estimates are standardized as Cohen's d. Such standardization can mechanically induce a correlation between effect sizes and their standard errors, violating the independence condition that underlies standard funnel-based meta-analytic models in the absence of publication bias. Applying MAIVE, we therefore regress the squared reported standard errors on inverse sample size and use the predicted values in place of the reported variance in the meta-regression. For our baseline MAIVE specification, we use the instrumented version of PET-PEESE without additional inverse-variance weighting and cluster standard errors at the study level, as recommended by Irsova et al. (2025).

The MAIVE estimate, reported in Column 1 of Panel C in Table 4, is -0.119 and statistically significant. The first stage confirms instrument relevance: inverse sample size strongly predicts reported squared standard errors ($F = 271.716$). The Anderson–Rubin

95% confidence interval, $[-0.137, -0.087]$, remains entirely below zero. The estimated publication-bias component (the coefficient on the fitted SE^2) is -0.245 and statistically significant at the 1% level, suggesting that publication bias remains detectable even after instrumenting for the potentially endogenous relationship between reported effects and precision. Substantively, the MAIVE estimate corresponds to a learning deficit of $0.119/0.4 = 0.30$ school years. This estimate is nearly identical to the RoBMA estimate of -0.118 reported in Table 3 and close to the uncorrected random-effects mean of -0.140. Because MAIVE directly addresses endogenous precision and the mechanical correlation introduced by standardization to Cohen's d, we treat it, together with RoBMA, as one of the most credible estimates in our analysis. Taken together, these estimates point to a learning deficit of approximately 30% of a school year.

RTMA. While MAIVE addresses endogenous precision that may arise from p-hacking or from the standardization to Cohen's d, it does not explicitly model p-hacking as a within-study search process. We therefore next apply the right-truncated meta-analysis (RTMA)[6] of Mathur (2024b), which is designed to account for p-hacking as selection within studies, while also allowing for selective publication across studies. RTMA is correctly specified when the favored estimates in p-hacked studies are affirmative, meaning that researchers search across estimates until an affirmative result is obtained, or when p-hacked studies with non-affirmative favored estimates are not published. The method also accounts for within-study heterogeneity, allowing both independent and autocorrelated estimates within studies and providing a framework for inference. The resulting RTMA estimate,[7] reported in Column 2 of Panel C in Table 4, is -0.039, corresponding to approximately $0.039/0.4 = 0.10$ school years of learning loss. The estimate remains negative but is much closer to zero than the MAIVE and RoBMA estimates.

Because RTMA is formulated as a random-effects model, it does not impose a single common true effect across studies; instead, it estimates an underlying distribution

[6] using `PublicationBias` and `phacking` packages available at metabias.io.

[7] In our application, affirmative results are negative and statistically significant estimates of learning loss. Since the RTMA implementation is formulated for settings in which affirmative results are positive and statistically significant, we apply the method to sign-reversed estimates and then reverse the sign of the resulting estimate. See Appendix B for a detailed discussion of this implementation choice and a related coding issue in `phacking_meta`.

of population effects, summarized by a mean effect and a heterogeneity parameter. The estimated heterogeneity parameter is 0.072, which captures the estimated standard deviation of the underlying true effects around the corrected mean (Borenstein et al. 2021, Mathur 2024b). Thus, the fitted RTMA model implies a relatively small average learning-loss effect, centered at -0.039, together with meaningful dispersion in the underlying true effects. Because RTMA relies on assumptions about the distribution of published non-affirmative estimates and the selection process, we examine its diagnostic fit before interpreting the estimate substantively.

Diagnostic plots for RTMA. Following Mathur (2024b), we examine two diagnostic plots – the z-score distribution and the Quantile-Quantile (Q-Q) plot showing the fitted versus the empirical CDF functions. The z-score distribution helps assess whether the observed pattern of selective reporting is consistent with the one-directional selection assumed by RTMA. In our setting, affirmative estimates are negative and statistically significant, so concentration around $z = -1.96$ is consistent with selection favoring negative significant learning-loss estimates. By contrast, comparable concentration around both -1.96 and $+1.96$, or concentration favoring non-affirmative significant estimates, would raise concerns about violations of RTMA's selection assumptions. The RTMA diagnostic Q-Q plot assesses whether the published non-affirmative estimates are well described by the fitted truncated distribution implied by the RTMA estimates of the mean effect and between-study heterogeneity.

Figure 7 displays the distribution of z-scores in the original data. The distribution is left-skewed, with visible mass near the negative critical threshold $z = -1.96$, consistent with a preference for negative statistically significant results. At first glance, this pattern could be interpreted as evidence of p-hacking. However, Figure 8 (b) shows mass concentrated near $|z| = 1.96$; because it is computed on absolute values, this does not by itself distinguish directional selective reporting from features of the power or sample-size distribution (Elliott et al. 2022b). The caliper test in Table C1 points in the same direction: we find no significant excess of estimates just inside conventional significance thresholds.

Figure 10 presents the RTMA diagnostic Q-Q plot, which compares the fitted CDF

of the published non-affirmative estimates with their empirical CDF. The points adhere closely to the 45-degree line in the lower quantiles, but deviations become more pronounced toward the right-hand side of the distribution, with dispersion increasing at higher quantiles. This pattern suggests that RTMA captures the lower part of the distribution reasonably well but fits the upper tail less accurately. One possible explanation is that the mechanical relationship between Cohen's d and its standard error distorts the distribution of non-affirmative estimates in ways not fully captured by the truncated normal model, as discussed in Sections 2 and 3.

Taken together, the diagnostics do not provide strong evidence of p-hacking, and the imperfect tail fit in the RTMA diagnostic plot suggests that the RTMA point estimate should be interpreted cautiously. We therefore view RTMA as a useful robustness check rather than a preferred correction.

Multiple biases. The preceding analyses separately address publication bias, spurious precision, and p-hacking. A remaining concern is that selective publication may interact with internal study bias. For example, if internally biased studies are more likely to produce large negative estimates, and if large negative estimates are more likely to be published, then publication bias may disproportionately select studies with greater internal bias. In this case, correcting only for selective publication, while assuming that all studies have the same average internal bias, may understate the combined distortion.

To address this possibility, we apply the multi-bias sensitivity framework of Mathur (2024c). The method asks how the meta-analytic estimate would change under specified assumptions about both publication bias and average internal bias.[8] It can also ask how large internal bias would need to be, given a specified degree of publication bias, to attenuate the result to the null or to another benchmark value.

The framework can allow internal bias to affect all studies or only selected subsets, such as non-randomized studies. Ideally, one would therefore allow internal bias to differ by study design, for example by assigning different bias parameters to randomized

[8](1) *"For a given severity of internal bias across studies and of publication bias, how much could the results change?"*; and (2) *"For a given severity of publication bias, how severe would internal bias have to be, hypothetically, to attenuate the results to the null or by a given amount?"*

and non-randomized studies. Because we do not have complete information on which estimates come from randomized designs, we take a more aggregate approach and allow average internal bias to differ between affirmative and non-affirmative estimates. This specification directly addresses the concern that affirmative estimates may differ from non-affirmative estimates not only in their probability of publication, but also in their average degree of internal bias.

Columns 3 and 4 of Panel C in Table 4 report the multi-bias results. As a benchmark, we use the robust random-effects selection-ratio estimate with a four-fold preference for affirmative results, reported in Panel A, Column 3. Under this specification, and assuming no internal bias, the corrected estimate is -0.068 with a 95% confidence interval from -0.101 to -0.036. We then impose the same four-fold selection ratio but allow affirmative studies to have greater average internal bias than non-affirmative studies – the mean internal bias is set to 0.05 for affirmative studies and 0.01 for non-affirmative studies.[9] Because the estimates in our application are measured as Cohen's d, the internal-bias parameters are also measured in standard-deviation units. Thus, a value of 0.05 means that affirmative estimates are assumed to contain, on average, internal bias of 0.05 standard deviations, while a value of 0.01 corresponds to very small average internal bias among non-affirmative estimates. These values should be interpreted as sensitivity parameters that quantify how much the corrected meta-analytic estimate would change under alternative assumptions about study-level bias.

Under these assumptions, allowing for differential internal bias moves the corrected estimate from the publication-bias-only benchmark of -0.068 to -0.097, and then to -0.111 when the assumed internal bias among affirmative estimates is increased to 0.08. The multi-bias exercise therefore clarifies that the smaller RE-SR4 estimate reflects the assumption that affirmative and non-affirmative estimates have the same average internal bias. Once affirmative studies are allowed to have somewhat greater internal bias, the corrected estimates move toward the MAIVE and RoBMA estimates rather than toward

[9]Here we assume: (1) affirmative studies are four times more likely to be published than non-affirmative studies; and (2) affirmative studies have a mean internal bias of 0.05, and (3) nonaffirmative studies have a mean internal bias of 0.01, which indicates very little bias.

the smaller RTMA estimate. We therefore interpret the multi-bias results as a sensitivity analysis rather than as a standalone correction. Its main implication is that plausible differences in internal bias across affirmative and non-affirmative estimates do not overturn the conclusion of a substantial negative effect of COVID-19 on learning.

Overall, the methods in this section reinforce the interpretation from Section 3. MAIVE yields an estimate almost identical to RoBMA, while RTMA produces a smaller estimate but fits the upper tail of the distribution less accurately. The multi-bias analysis shows that allowing affirmative studies to differ modestly in internal bias moves the selection-adjusted estimates back toward the RoBMA–MAIVE range. We therefore place greatest weight on RoBMA and MAIVE, which point to a learning-loss effect of approximately -0.12, or about 30% of a school year.

5 Conclusion

This paper builds on the meta-analysis of Betthäuser et al. (2023a) on the effect of the COVID-19 pandemic on the learning progress of school-aged children, extending their analysis by applying a comprehensive set of formal bias correction methods — spanning funnel-based tests, selection models, Bayesian meta-analysis, and instrumental variable estimation — to assess the robustness of the underlying effect size estimate against publication bias and p-hacking.

Our graphical analysis reveals that the evidence from visual diagnostics alone is ambiguous. The log-transformed z-score distribution shows no clustering around the significance threshold. However, the raw z-score distribution exhibits pronounced bunching at the 5% level, and our funnel plot displays some asymmetry, with a wider spread of estimates on the negative side. Taken together, these patterns are consistent with some degree of selective publication, though not conclusive on their own, and motivate the formal correction methods we employ.

The formal correction methods tell a consistent story. PET-PEESE yields a corrected estimate of -0.245, almost double the uncorrected estimate in magnitude, and is consistent with publication bias. The 3PSM and RoBMA estimates, at -0.123 and -0.118

respectively, are closer to the uncorrected mean of −0.14, but still suggest modest downward bias. The sensitivity analyses of Mathur and VanderWeele (2020) show that while results are sensitive to a worst-case selection scenario — MAN yields a sign reversal to 0.021 — an implausibly large degree of bias, 29.60 times stronger publication of affirmative over non-affirmative studies, would be required to drive the point estimate to zero. No degree of selection bias under this model can shift the estimate to +0.05.

Accounting additionally for p-hacking and the spurious correlation between estimates and standard errors introduced by normalization to Cohen's d, MAIVE yields an estimate of −0.119, highly statistically significant and closely consistent with RoBMA. The RTMA estimate of −0.039, while directionally consistent, is discounted due to imperfect upper-tail fit indicated by the diagnostic Q-Q plot. The multi-bias analysis, which allows for differential internal bias between affirmative and non-affirmative studies, further corroborates the RoBMA and MAIVE estimates: allowing affirmative estimates to carry somewhat larger internal bias moves the corrected estimates toward the RoBMA–MAIVE range, rather than estimating such differential bias from the data.

The key finding is the near-perfect consistency between RoBMA and MAIVE — both yielding estimates around −0.119 — and their close alignment with the uncorrected benchmark estimate of −0.14. RoBMA model-averages across publication-bias models, while MAIVE addresses endogenous precision and the mechanical link between Cohen's d and its standard error. This consistency across methodologically distinct approaches indicates that COVID-19 is associated with a learning deficit of approximately 30% of a school year based on the preferred RoBMA and MAIVE estimates, compared with about 35% using the uncorrected benchmark. This magnitude remains economically meaningful: the corrections change the estimated size of the learning deficit, but they do not change the conclusion that the shock was large enough to matter for human-capital policy. While our analysis reveals evidence of selection bias in the underlying data, this bias does not systematically distort the overall findings of Betthäuser et al. (2023a). The conclusion of a substantial and persistent COVID-19 learning deficit is robust to a wide range of correction methods.

More broadly, this robustness exercise highlights the growing importance of systematic comparison across meta-analytic techniques. The rapid expansion of bias-correction methods in recent years has substantially enriched the toolkit available to researchers, while also making the choice among alternative approaches increasingly consequential. These methods address different forms of bias, rest on different assumptions, and can yield meaningfully different corrected estimates. Credible evidence synthesis therefore requires applying bias-correction methods and assessing how conclusions vary across them. In this sense, our analysis compares alternative meta-analytic approaches applied to the same dataset and research question. Without such comparisons, it is difficult to know whether a corrected estimate reflects a robust empirical conclusion or primarily the assumptions of the chosen method.

Data Availability Statement

The effect-size dataset analyzed in this paper was compiled by Betthäuser et al. (2023a) and is publicly available at https://doi.org/10.17605/osf.io/u8gaz. Replication code for all analyses reported in this paper is available at https://meta-analysis.cz/learning.

Use of Generative AI

The authors used generative AI tools, including Claude Sonnet (Anthropic) and ChatGPT (OpenAI), during manuscript preparation to improve language and readability and to assist with formatting and debugging statistical code in `R`. The tools were accessed through their standard web interfaces during 2025–2026. AI tools were not used for study selection, data coding, estimation choices, analytical decisions, interpretation of results, or the formulation of conclusions. AI assistance with code was limited to formatting, debugging, and implementation support; all statistical analyses were run and verified by the authors. All results reported in the paper are fully reproducible without AI tools. The authors reviewed and verified all AI-assisted output and take full responsibility for the content of the manuscript. AI use follows the guiding principles of Cook

et al. (2026a) and the reporting standards of Cook et al. (2026b).

References

Ardington, C., Wills, G. and Kotze, J.: 2021, Covid-19 learning losses: Early grade reading in south africa, *International Journal of Educational Development* **86**, 102480.

Azevedo, J. P., Hasan, A., Goldemberg, D., Geven, K. and Iqbal, S. A.: 2021, Simulating the potential impacts of covid-19 school closures on schooling and learning outcomes: A set of global estimates, *The World Bank Research Observer* **36**(1), 1–40.

Bartoš, F., Maier, M., Quintana, D. S. and Wagenmakers, E.-J.: 2022, Adjusting for publication bias in JASP and R: Selection models, PET-PEESE, and robust bayesian meta-analysis, *Advances in Methods and Practices in Psychological Science* **5**(3), 25152459221109259.

Bartoš, F., Maier, M., Wagenmakers, E.-J., Doucouliagos, H. and Stanley, T.: 2023, Robust bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods, *Research Synthesis Methods* **14**(1), 99–116.

Betthäuser, B. A., Bach-Mortensen, A. M. and Engzell, P.: 2023a, A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemic, *Nature human behaviour* **7**(3), 375–385.

Betthäuser, B. A., Bach-Mortensen, A. M. and Engzell, P.: 2023b, A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemic. Dataset. https://doi.org/10.17605/osf.io/u8gaz.

Bloom, H. S., Hill, C. J., Black, A. R. and Lipsey, M. W.: 2008, Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions, *Journal of research on educational effectiveness* **1**(4), 289–328.

Borenstein, M., Hedges, L. V., Higgins, J. P. and Rothstein, H. R.: 2021, *Introduction to meta-analysis*, John wiley & sons.

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.

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y.: 2016, Star wars: The empirics strike back, *American Economic Journal: Applied Economics* **8**(1), 1–32.

Cala, P., Havranek, T., Irsova, Z., Luskova, M., Matousek, J. and Novak, J.: 2026, Financial incentives and performance: a meta-analysis of experiments in economics, *Journal of Political Economy Microeconomics forthcoming* .

Cook, N., Bartoš, F., Bom, P. R., Gechert, S., Kantová, K., Geyer-Klingeberg, J., Havránek, T., Irsova, Z., Luskova, M., Opatrnỳ, M. et al.: 2026a, Guidance for the use of ai in the meta-analysis of economics research, *Journal of Economic Surveys* .

Cook, N., Bartoš, F., Bom, P. R., Gechert, S., Kantová, K., Geyer-Klingeberg, J., Havránek, T., Irsova, Z., Luskova, M., Opatrnỳ, M. et al.: 2026b, Reporting guidelines for meta-analysis in economics—updated for ai, *Journal of Economic Surveys* .

Di Pietro, G.: 2023, The impact of COVID-19 on student achievement: Evidence from a recent meta-analysis, *Educational research review* **39**, 100530.

Donnelly, R. and Patrinos, H. A.: 2022, Learning loss during COVID-19: An early systematic review, *Prospects* **51**(4), 601–609.

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

Elliott, G., Kudrin, N. and Wüthrich, K.: 2022b, The power of tests for detecting p-hacking, *arXiv preprint arXiv:2205.07950* .

Engzell, P., Frey, A. and Verhagen, M. D.: 2021, Learning loss due to school closures during the COVID-19 pandemic, *Proceedings of the national academy of sciences* **118**(17), e2022376118.

Gerber, A. and Malhotra, N.: 2008, Do statistical reporting standards affect what is published? publication bias in two leading political science journals, *Quarterly Journal of Political Science* **3**(3), 313–326.

Hanushek, E. A. and Woessmann, L.: 2008, The role of cognitive skills in economic development, *Journal of economic literature* **46**(3), 607–668.

Hanushek, E. A. and Woessmann, L.: 2020, The economic impacts of learning losses, *OECD Education Working Papers 225*, OECD Publishing, Paris.

Havranek, T., Irsova, Z., Laslopova, L. and Zeynalova, O.: 2024, Publication and attenuation biases in measuring skill substitution, *Review of Economics and Statistics* **106**(5), 1187–1200.

Havránek, T., Stanley, T. D., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W. R., Rost, K. and van Aert, R. C.: 2020, Reporting guidelines for meta-analysis in economics, *Journal of Economic Surveys* **34**(3), 469–475.

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

Hedges, L. V., Tipton, E. and Johnson, M. C.: 2010, Robust variance estimation in meta-regression with dependent effect size estimates, *Research synthesis methods* **1**(1), 39–65.

Hill, C. J., Bloom, H. S., Black, A. R. and Lipsey, M. W.: 2008, Empirical benchmarks for interpreting effect sizes in research, *Child development perspectives* **2**(3), 172–177.

Irsova, Z., Bom, P. R., Havranek, T. and Rachinger, H.: 2025, Spurious precision in meta-analysis of observational research, *Nature Communications* **16**(1), 8454.

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

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

König, C. and Frey, A.: 2022, The impact of COVID-19-related school closures on student achievement—a meta-analysis, *Educational Measurement: Issues and Practice* **41**(1), 16–22.

Maier, M., Bartoš, F. and Wagenmakers, E.-J.: 2023, Robust bayesian meta-analysis: Addressing publication bias with model-averaging., *Psychological Methods* **28**(1), 107.

Mathur, M. B.: 2024a, Assessing robustness to worst case publication bias using a simple subset meta-analysis, *bmj* **384**.

Mathur, M. B.: 2024b, P-hacking in meta-analyses: A formalization and new meta-analytic methods, *Research Synthesis Methods* **15**(3), 483–499.

Mathur, M. B.: 2024c, Sensitivity analysis for the interactive effects of internal bias and publication bias in meta-analyses, *Research synthesis methods* **15**(1), 21–43.

Mathur, M. B. and VanderWeele, T. J.: 2020, Sensitivity analysis for publication bias in meta-analyses, *Journal of the Royal Statistical Society Series C: Applied Statistics* **69**(5), 1091–1119.

Opatrny, M., Havranek, T., Irsova, Z. and Scasny, M.: 2025, Publication bias and model uncertainty in measuring the effect of class size on achievement, *Journal of Labor Economics forthcoming* .

Simonsohn, U., Nelson, L. D. and Simmons, J. P.: 2014, *p*-curve: a key to the file-drawer., *Journal of experimental psychology: General* **143**(2), 534.

Stanley, T. D.: 2017, Limitations of PET-PEESE and other meta-analysis methods, *Social Psychological and Personality Science* **8**(5), 581–591.

Stanley, T. D. and Doucouliagos, H.: 2017, Neither fixed nor random: weighted least squares meta-regression, *Research synthesis methods* **8**(1), 19–42.

Stanley, T. D., Ioannidis, J. P., Maier, M., Doucouliagos, H., Otte, W. M. and Bartoš, F.: 2023, Unrestricted weighted least squares represent medical research better than random effects in 67,308 cochrane meta-analyses, *Journal of Clinical Epidemiology* **157**, 53–58.

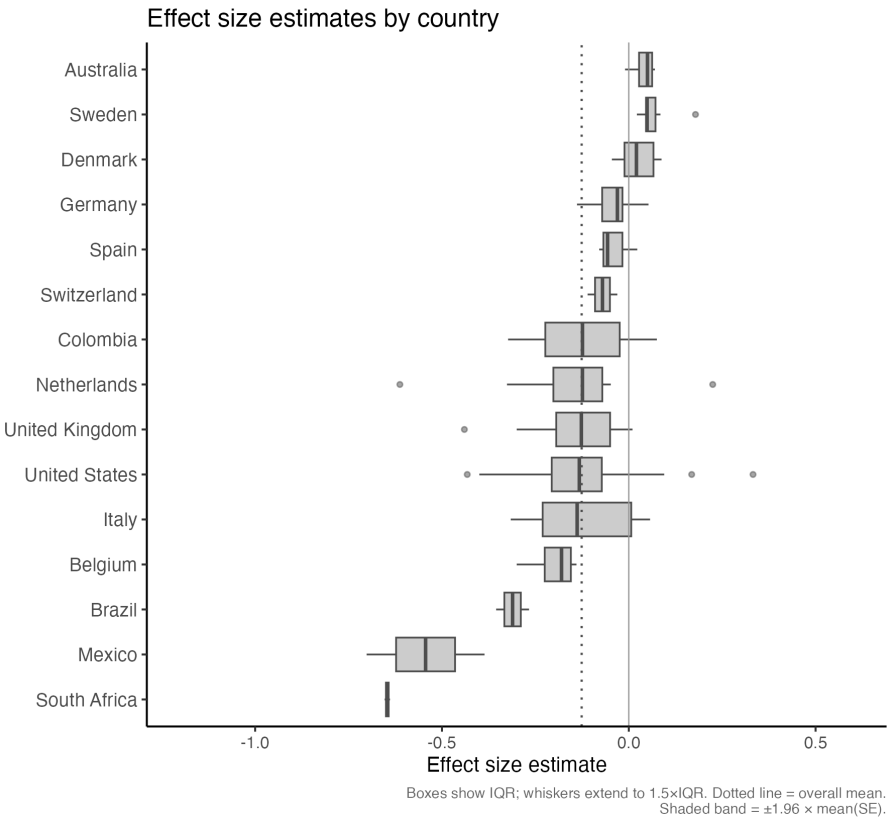
United Nations: 2020, The impact of COVID-19 on children. UN Policy Brief.

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

Wisenöcker, A. S., Helm, C., Große, C. S., Hübner, N. and Zitzmann, S.: 2025, A meta-analysis of students' academic learning losses over the course of the COVID-19 pandemic, *Learning and Instruction* **98**, 102111.

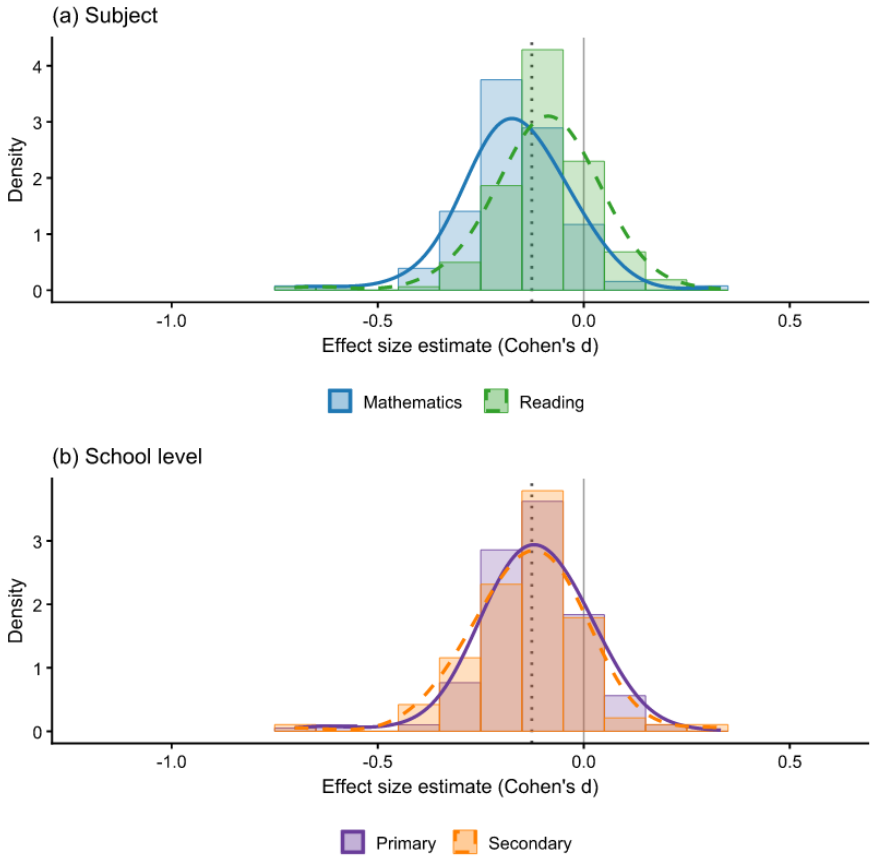
6 Figures

Figure 1: Effect size estimates by country, ordered by median.



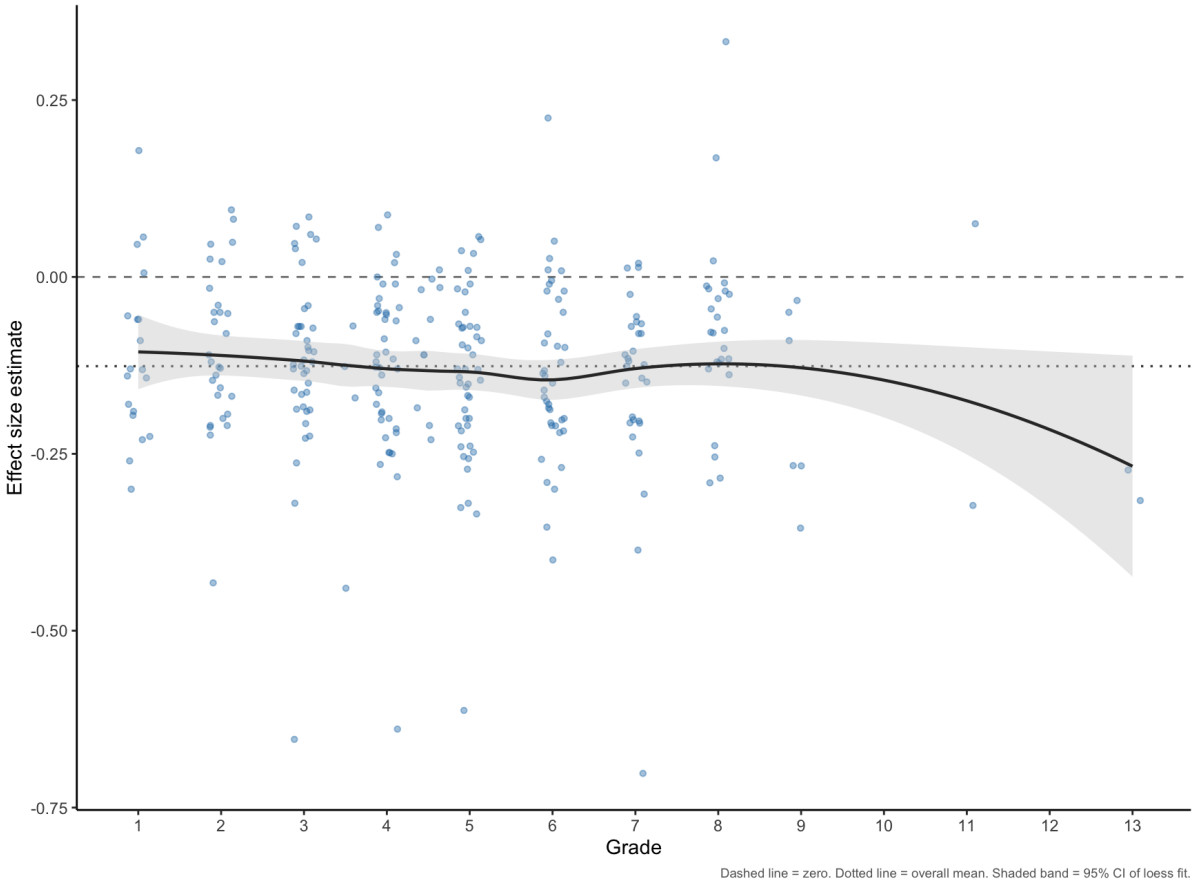
Notes: Boxes show the interquartile range of effect-size estimates within each country; whiskers extend to 1.5 × IQR, and dots denote outliers. The vertical dotted line marks the overall mean effect. Countries at the top of the figure, including Australia, Sweden, and Denmark, report near-zero or slightly positive effects, while Mexico and South Africa show the largest negative estimates. The three most represented countries—the United States, the United Kingdom, and the Netherlands—cluster close to the overall mean, though with substantial within-country dispersion.

Figure 2: Distribution of effect size estimates by subject area and school level.



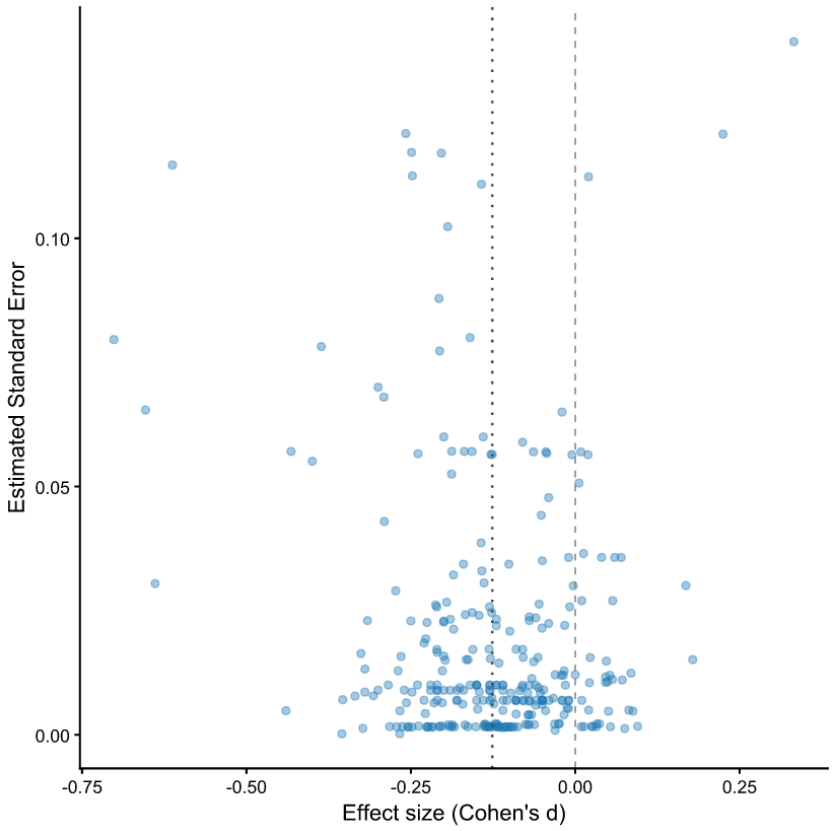
Note: Panel (a) shows the distribution of effect sizes for mathematics ($N = 128$) and reading ($N = 161$), and panel (b) for primary ($N = 196$) and secondary education ($N = 95$). Each panel overlays histograms and kernel density curves. The solid vertical line marks zero (no deficit); the dotted vertical line marks the overall pooled mean ($d = -0.126$). Learning deficits in mathematics tend to be larger on average than in reading (unweighted means of -0.167 vs. -0.095), whereas primary and secondary education show similar distributions (-0.122 vs. -0.135). Two estimates covering a composite of mathematics and reading are excluded from panel (a).

Figure 3: Effect size estimates by grade, with a loess smoother and 95% confidence band.



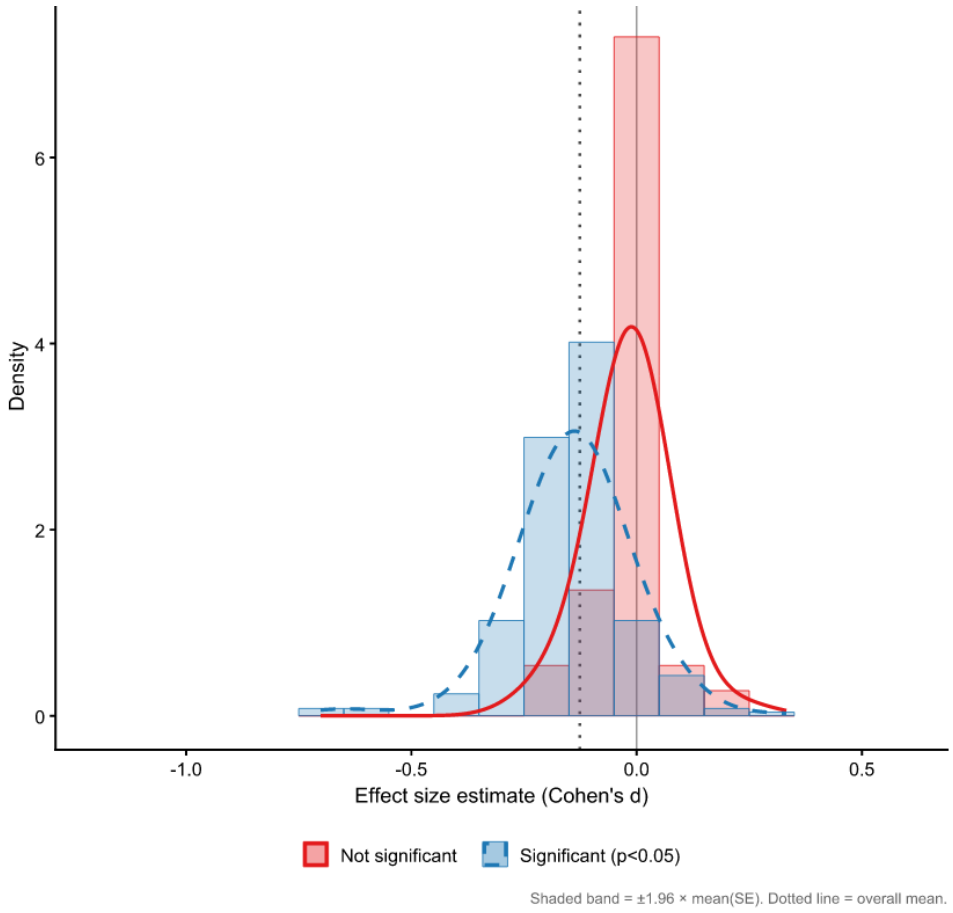
Note: Individual estimates are shown as jittered points. The dashed line marks zero; the dotted line marks the overall mean (-0.126). Effects are broadly stable across grades 1–9, consistent with the absence of a meaningful school-level gradient visible in panel (b) of Figure 2. The apparent decline in grades 11–13 should be interpreted cautiously: these grades are represented by very few estimates (two at grade 11, two at grade 13), and the wide confidence band reflects this uncertainty.

Figure 4: Conventional funnel plot



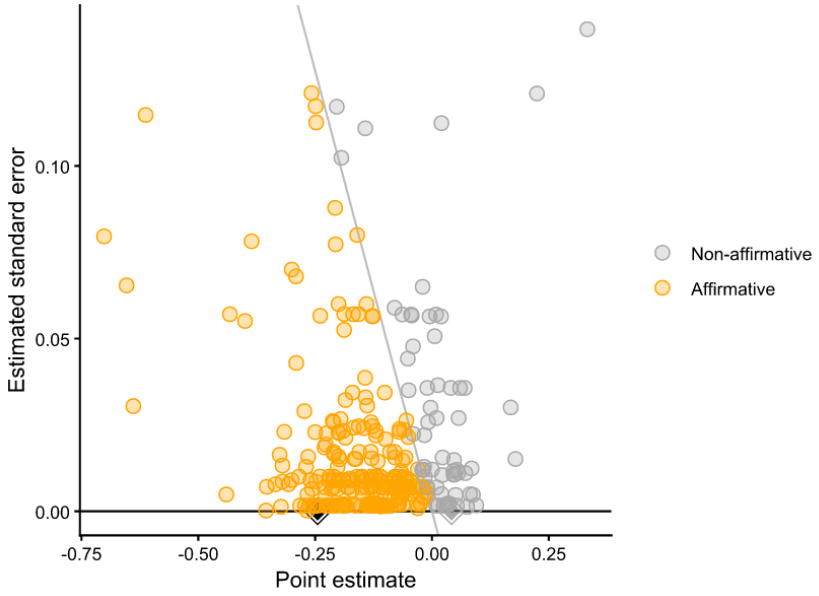
Notes: The figure shows a conventional funnel plot with effect sizes on the horizontal axis and their estimated standard errors on the vertical axis. The dotted vertical line marks the pooled mean effect. The dashed vertical line marks zero.

Figure 5: Distribution of effect size estimates by statistical significance



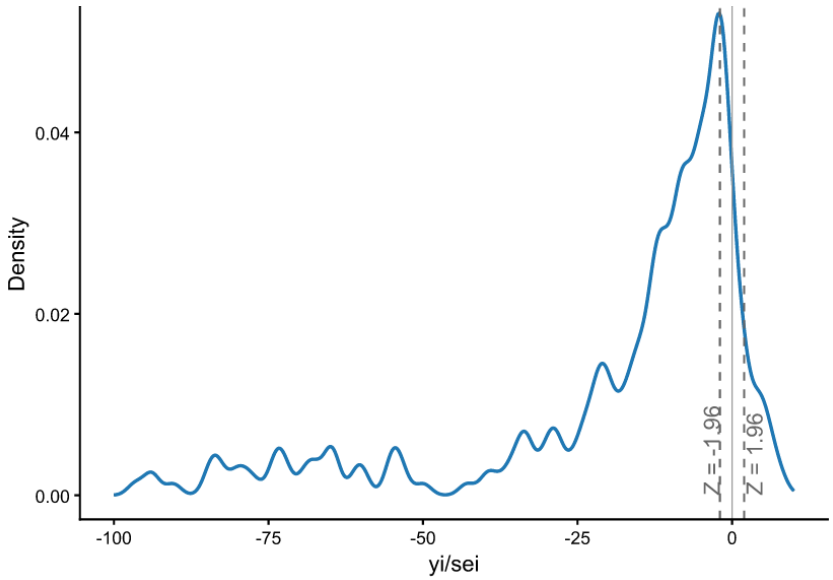
Notes: The figure shows the distribution of effect size estimates (Cohen's d), separately for statistically significant ($p < 0.05$, blue) and non-significant estimates (red). The dotted vertical line marks the overall mean effect size. The shaded band indicates the interval defined by $\pm 1.96\times$ the mean standard error around zero.

Figure 6: Significance funnel plot



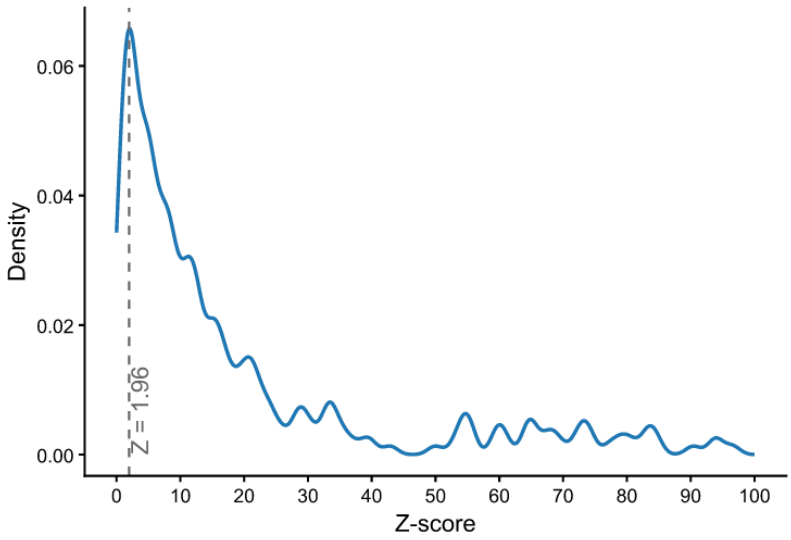
Notes: The figure displays a significance funnel plot following Mathur and VanderWeele (2020). Orange points are affirmative estimates; gray points are non-affirmative. The gray diamond is the fixed-effects mean among non-affirmative estimates (0.021), corresponding to the MAN worst-case benchmark reported in Panel A of Table 4. The black diamond is the precision-weighted pooled mean across all 291 estimates ($\hat{\mu}_{IVW} = -0.245$); it lies more negative than the Panel A equal-weighted mean of -0.126 because inverse-variance weighting gives disproportionate influence to large, high-precision studies.

Figure 7: Distribution of z-scores.

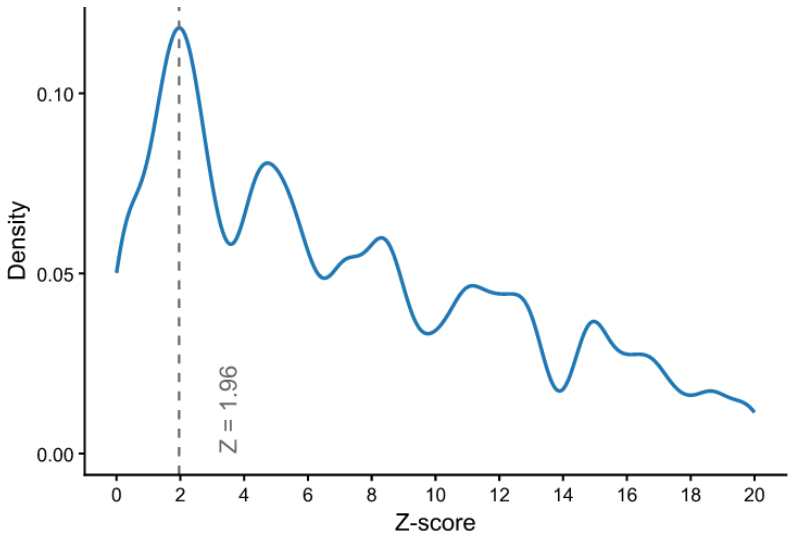


Note: The figure shows the distribution of z-scores, y_i/se_i, in the original dataset. The dashed vertical lines mark $z = -1.96$ and $z = +1.96$, the conventional 5% two-sided significance thresholds.

Figure 8: Comparison between the distribution of the z-scores in absolute values (a), (b), and in logs of absolute values (c).

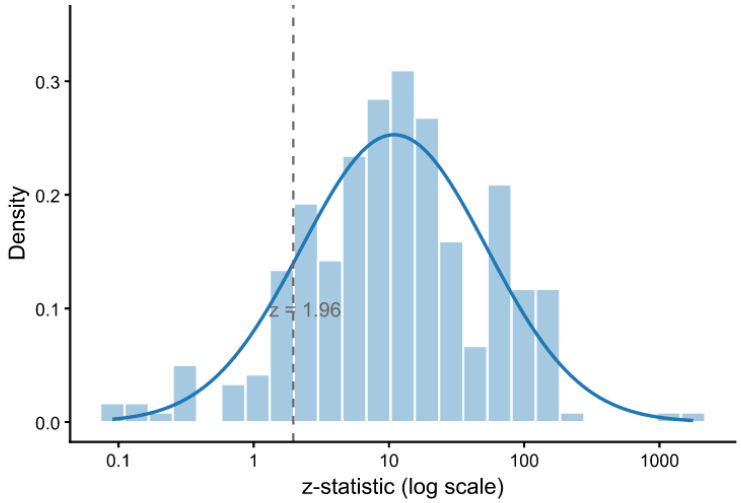


(a) The distribution of absolute z-scores, $abs(y_i/se_i)$. The dashed vertical line marks $z = 1.96$, the conventional 5% two-sided significance threshold.



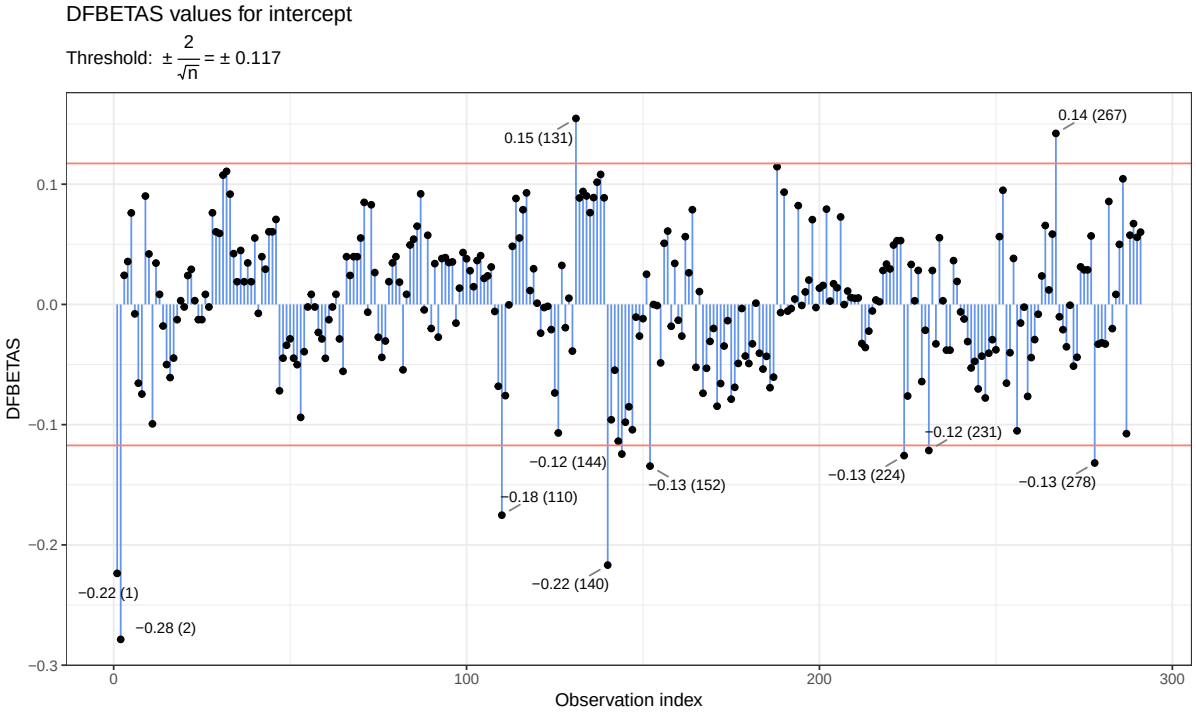
(b) The distribution of absolute z-scores, $(abs(y_i/se_i))$, zoomed to the (0, 20) neighborhood. The dashed vertical line marks $z = 1.96$, the conventional 5% two-sided significance threshold.

Figure 8: Comparison between distribution of absolute value and log z-scores



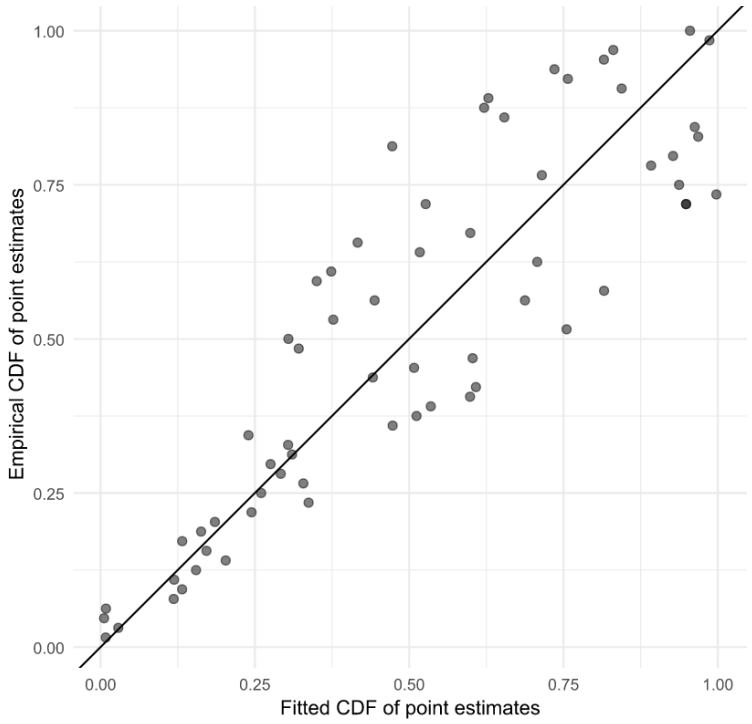
(c) The figure shows the distribution of log absolute z-scores, $ln(abs(y_i/se_i))$. The dashed vertical line marks $z = 1.96$, the conventional 5% two-sided significance threshold.

Figure 9: DFBETAS influence diagnostics for intercept.



Notes: The figure displays DFBETAS values for all 291 estimates, measuring the change in the pooled random-effects estimate when each observation is omitted in turn. The horizontal lines mark the sample-size-adjusted threshold $\pm 2/\sqrt{n} = \pm 0.117$. Eleven estimates exceed the threshold; excluding them moves the pooled estimate from -0.124 to -0.116.

Figure 10: Diagnostic q-q plot



Notes: The figure displays the RTMA diagnostic Q-Q plot, comparing the fitted CDF of published non-affirmative estimates (horizontal axis) with their empirical CDF (vertical axis). Points lying on the 45-degree line indicate good model fit. Deviations in the upper tail suggest the truncated normal model fits the lower quantiles well but captures the upper tail less accurately, consistent with the mechanical distortion introduced by standardizing estimates as Cohen's d.

7 Tables

Table 1: Descriptive statistics

Variable	N	Mean	SD	Min	Median	Max
Standardized mean difference	291	−0.126	0.127	−0.702	−0.124	0.333
Standard error	291	0.019	0.026	< 0.001	0.009	0.140
Sample size (n)	291	446,417	—	275	50,000	10,884,922
Absolute z-statistic	291	37.8	133.7	0.000	11.0	1,775
Composition						
Primary education	196	share = 67%				
Secondary education	95	share = 33%				
Mathematics	128	share = 44%				
Reading	161	share = 55%				
Mixed	2	share = 0.69%				

Notes: The dataset contains 291 effect size estimates from 42 studies across 15 countries, drawn from Betthäuser et al. (2023a). Effect sizes are expressed as standardized mean differences. Sample size refers to the number of students in the primary study from which each estimate is drawn. The mean and median sample size reflect the highly right-skewed distribution of study sizes; two estimates covering both mathematics and reading jointly are excluded from the subject composition rows.

Table 2: Summary statistics by subgroup

		Unweighted			Weighted (1/n per study)		
	N	Mean	95% CI		Mean	95% CI	
Subject							
Mathematics	128	-0.167	-0.188	-0.145	-0.164	-0.205	-0.122
Reading	161	-0.095	-0.113	-0.076	-0.13	-0.169	-0.09
Other/mixed	2	-0.07	-0.148	0.008	-0.07	-0.126	-0.015
School level							
Primary	196	-0.122	-0.139	-0.104	-0.138	-0.168	-0.107
Secondary	95	-0.135	-0.162	-0.109	-0.153	-0.213	-0.094
Grade group							
Early primary (1–4)	132	-0.119	-0.14	-0.098	-0.144	-0.186	-0.102
Middle (5–8)	149	-0.128	-0.149	-0.108	-0.132	-0.17	-0.093
Secondary (9+)	10	-0.19	-0.283	-0.097	-0.199	-0.312	-0.086
Sample size							
Small (<10k)	74	-0.151	-0.192	-0.111	-0.202	-0.267	-0.138
Medium (10k–100k)	118	-0.108	-0.126	-0.091	-0.094	-0.115	-0.074
Large (>100k)	99	-0.129	-0.15	-0.107	-0.131	-0.17	-0.092
Country							
United States	149	-0.136	-0.153	-0.12	-0.142	-0.166	-0.117
United Kingdom	58	-0.128	-0.151	-0.104	-0.136	-0.165	-0.107
Netherlands	27	-0.146	-0.198	-0.093	-0.14	-0.187	-0.093
Other countries	57	-0.089	-0.138	-0.039	-0.147	-0.207	-0.086
Source tier							
Tier 1: Peer-reviewed	62	-0.11	-0.157	-0.062	-0.165	-0.233	-0.098
Tier 2: Working paper	65	-0.132	-0.164	-0.1	-0.13	-0.164	-0.096
Tier 3: Gov't / institutional	34	-0.12	-0.156	-0.084	-0.122	-0.165	-0.079
Tier 4: Commercial	130	-0.133	-0.148	-0.118	-0.138	-0.163	-0.113
All estimates	291	-0.126	-0.141	-0.112	-0.142	-0.17	-0.115

Table 3: Publication bias: baseline and corrected estimates

Panel A: Baseline mean estimates

	Unweighted	**RVE**
Mean effect	−0.126**	−0.140***
	(0.059)	(0.020)
Observations	291	291

Panel B: Corrected estimates

	PET	**PEESE**	**3PSM**	**RoBMA**
Effect beyond bias	−0.271***	−0.245***	−0.123***	−0.118
	(0.040)	(0.051)	(0.009)	[−0.135, −0.094]
Publication bias	29.269***			
	(9.818)			
Publication bias		114.145		
Standard error^2		(58.836)		
Likelihood ratio test			$\chi^2 = 0.034$	
H0: no pub. bias			p-value = 0.854	
Observations	291	291	291	291

Notes: Unweighted = the equal-weighted (arithmetic) mean of all 291 estimates, i.e. the simple estimate-level mean, not the inverse-variance fixed-effect estimator. RVE = robust variance estimation mean following Hedges et al. (2010), clustering at the study level to account for within-study dependence; the RVE estimate of −0.140 coincides with the pooled mean reported by Bethäuser et al. (2023a). PET = precision effect test based on the estimates of weighted regression $estimate_{ij} = \beta_0 + \beta_1 * (SE_{estimate})_{ij} + u_{ij}$, where $estimate_{ij}$ is the i-th estimate from the j-th study, with $(SE_{estimate})_{ij}$ the respective standard error. PEESE = precision effect estimate with standard errors; for PEESE $(SE_{estimate})_{ij}$ is squared. 3PSM is a publication selection model as in Iyengar and Greenhouse (1988), Hedges (1992), Vevea and Hedges (1995). RoBMA = robust Bayesian model averaging as described in Bartoš et al. (2023), Maier et al. (2023). For PET & PEESE, we report heteroskedasticity robust standard errors clustered at the study level. Significant at ***[1%], **[5%], *[10%] level.

Table 4: Correcting for publication bias, additional specifications

Panel A: Worst-case and selection-ratio-adjusted estimates

	MAN	**FE-SR4**	**RE-SR4**
Effect beyond bias	0.021*	−0.206***	−0.068***
	(0.012)	(< 0.001)	(0.015)
Observations	41	291	291

Panel B: Publication bias required to explain away the results

	***coef*=0**	***coef*=0.01**	***coef*=0.05**
Estimate's *s*-value	29.60	60.65	no amount
CI *s*-value	8.31	10.46	no amount
Observations	291	291	291

Panel C: *p*-hacking & multibias

	MAIVE	**RTMA**	**Multi**$_{0.05;0.01}$	**Multi**$_{0.08;0.01}$
Effect beyond bias	−0.119***	−0.039***	−0.097***	−0.111***
	(0.012)	(0.003)	(0.008)	(0.008)
	[−0.137, −0.087]			
Coefficient on fitted SE^2	−0.245***			
	(0.051)			
First stage F-statistic	271.716			
Heterogeneity		0.072***		
		(0.001)		
Observations	291	291	291	291

Notes: FE = fixed effects mean, RE = mean effect estimated using robust random effects accounting for heterogeneity and clustering, MAN = meta-analysis of non-affirmative studies, FE-SR4 = fixed-effects meta-analysis with a 4-fold preference (selection ratio = 4) for affirmative studies, RE-SR4 = robust random-effects specification accounting for heterogeneity and clustering with the 4-fold preference for affirmative studies, RTMA = right-truncated meta-analysis, MAIVE = meta-analysis instrumental variable estimator, Anderson-Rubin 95% confidence interval is reported in square brackets, Multi$_{0.05;0.01}$ = affirmative studies bias is set to 0.05 and non-affirmative studies bias to 0.01, Multi$_{0.08;0.01}$ = affirmative studies bias is set to 0.08 and non-affirmative studies bias to 0.01. Standard errors are reported in parentheses. Significant at ***[1%], **[5%], *[10%] level. Mathur and VanderWeele (2020), Mathur (2024b,c), Irsova et al. (2025)

Online Appendix to:

Publication Bias and P-Hacking in the Effect of COVID-19 on Learning

Martina Luskova, Nino Buliskeria, Ali Elminejad, Tomas Havranek,

Zuzana Irsova, Stepan Jurajda, Marek Kapicka

A Computational Reproducibility of Betthäuser et al. (2023a)

In this section, we report on the computational reproducibility of the meta-analysis by Betthäuser et al. (2023a). Because our analysis relies on their study sample and dataset, rather than on an independent literature search, we do not provide a separate PRISMA flow diagram. Readers are referred to their Figure 1, which documents the study identification and selection process following PRISMA guidelines. Table A1 summarizes the contents of the replication package.

A.1 Computational Reproducibility

We used the replication package available on the Open Science Framework.[10] The replication package contains both code and data. The code incorporates the cleaning of the provided data. The final analysis data can be downloaded directly using the code in the replication package. See Table A1 for the description of replication package contents and reproducibility. We successfully computationally reproduced all the main results (*i.e.*, Figures 2b (pg. 377), 3 (pg. 378), 4 (pg.379), and 6 (pg.380)) from the raw data. The remaining figures were compiled manually by the original authors. Originally, Figure 4, pg. 379, is generated using an R extension in STATA. We reproduced the figure using the

[10] https://doi.org/10.17605/osf.io/u8gaz

same code directly in R. In subsection A.3, we present the reproduced figures; tables are presented in subsection A.4. Table A2 shows the original and replicated slope coefficient estimate, p-value, and 95% confidence interval for the learning deficits in time (mentioned in Figure 4, pg.379). Table A3 shows the original and the reproduced variation in estimates of learning deficit by school subject, level of education, and country income level (original results described in Figure 6, mean differences in text, pg. 380).

A.2 Discrepancies Between Pre-analysis Plan and Article

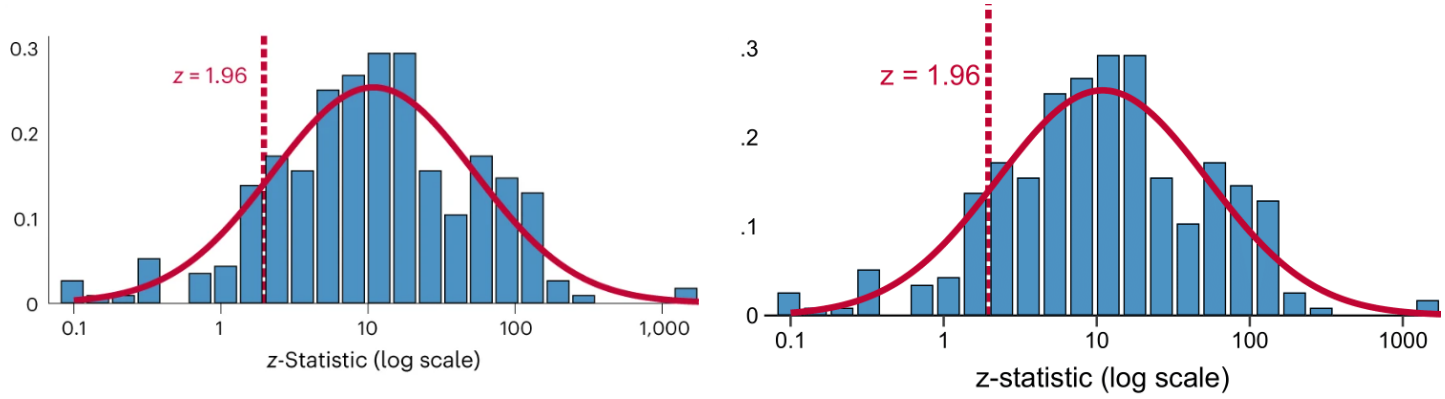
The authors registered a pre-analysis plan in the PROSPERO registry.[11] The paper follows the strategy specified in the pre-analysis plan, and relies on the pre-specified academic and pre-print databases. Regarding the data extraction, we find a minor deviation from the described plan. Despite the pre-analysis plan aiming to collect the key characteristics of the studied countries, the final data set only codes the country names. Moreover, the authors aimed to include the countries' income levels based on the World Bank's classification (low, lower-middle, upper-middle, and high-income). However, the majority of the dataset falls into the high-income category. The upper-middle-income is represented by less than 3% of the data. Low and lower-middle-income categories are not represented at all. Similarly, the final dataset does not include data on the funding source, sample restrictions, survey attrition, and follow-up period(s).

The pre-analysis plan describes the data synthesis strategy but does not specify the standardization measure (the article uses Cohen's d). The pre-analysis plan additionally aims to evaluate the learning differences between genders and varying exposure to school closures. These subgroup analyses were not performed due to the unavailability of the data. Lastly, there is no description of the specific tests the authors aimed to conduct. The article includes the following tests. To test for publication bias, the authors use a graphical test based on the distribution of z-statistics (assuming that the presence of publication bias can be seen in a notable jump in the distribution of z-statistics at the significance threshold, $z = 1.96$ or p-value $= 0.05$). Additionally, the supplementary

[11] https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42021249944

material features two more visual tests: a funnel plot and a test based on the p-curve. The article estimates the overall pooled effect size, focuses on the effect size in time, and performs sub-group analysis concerned with socio-economic inequality, school subjects (mathematics and reading), level of education, and country income level.

Figure A1: Publication bias: distribution of z-scores

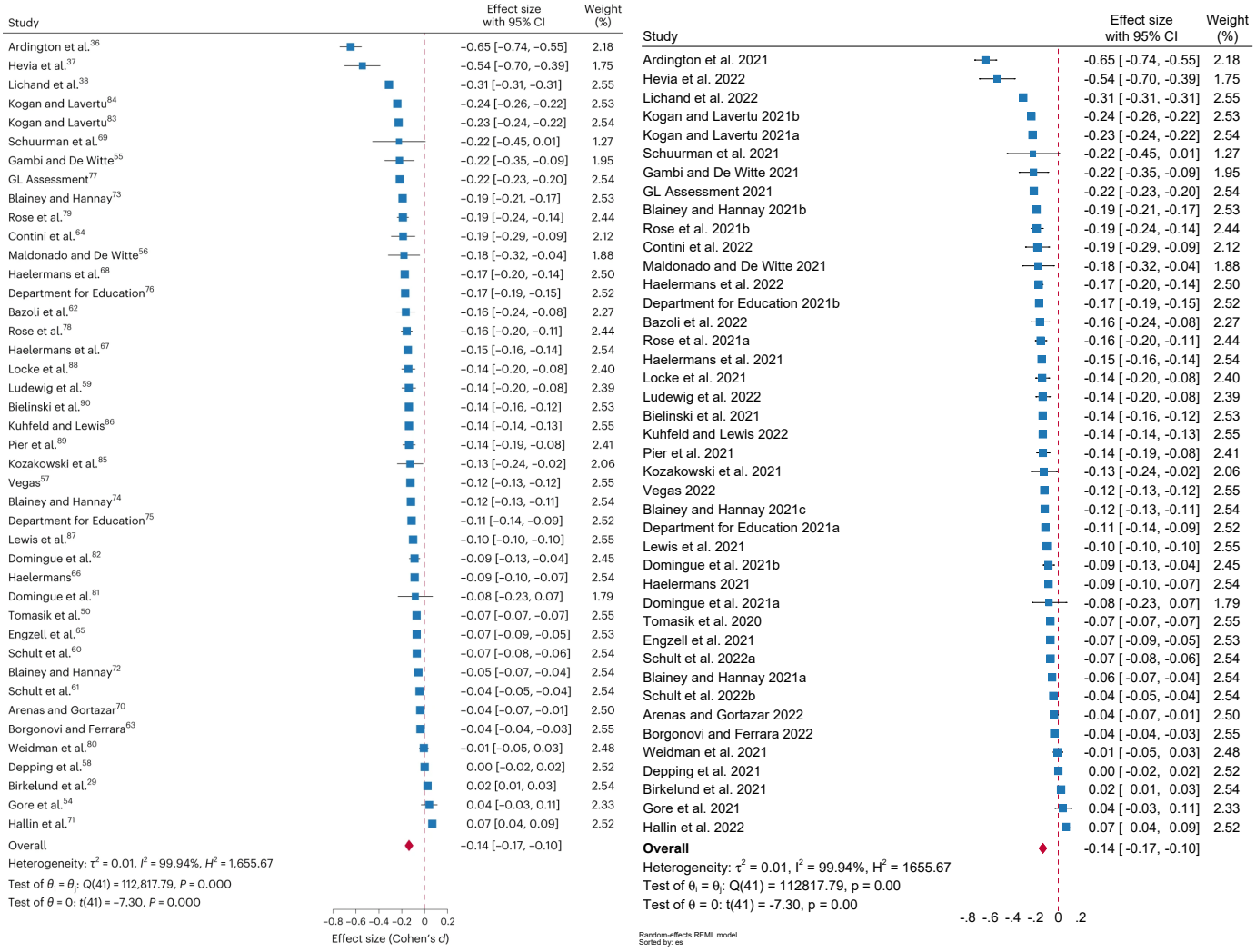


(a) Original Figure 2b

(b) Reproduction of Figure 2b

Notes: The figure displays a visual test for publication bias based on the distribution of the z-scores. (a) shows the original, and (b) is the reproduction. There is no difference between the two.

Figure A2: Forest plot

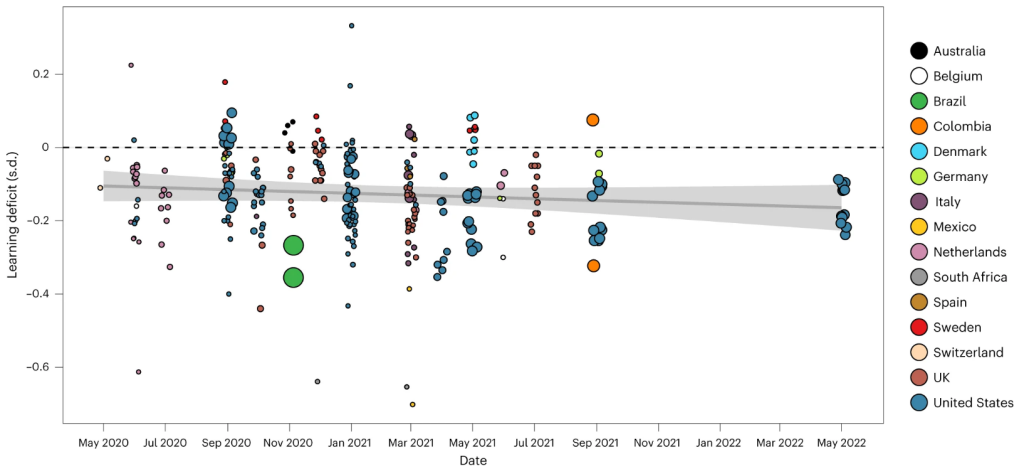


(a) Original Figure 3

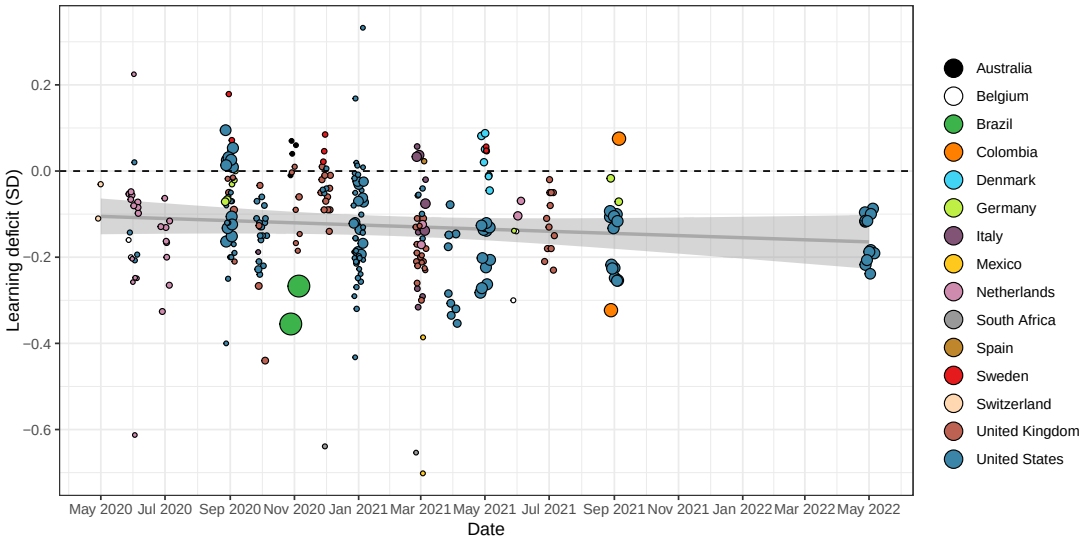
(b) Reproduction of Figure 3

Notes: The figure displays a forest plot of 42 included studies. The effects are expressed as Cohen's d weighted by the inverse of variance using the random effects model. (a) shows the original, and (b) is the reproduction. We reproduced the Blainey and Hannay 2021a effect size as −0.06, while the original article reports −0.05. The confidence intervals are the same, possibly due to rounding.

Figure A3: Estimates of COVID-19 learning deficits in time



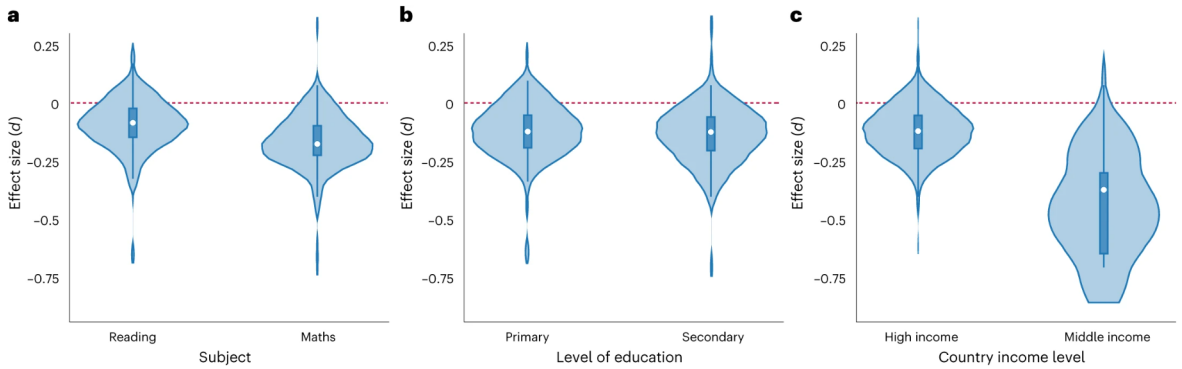
(a) Original Figure 4



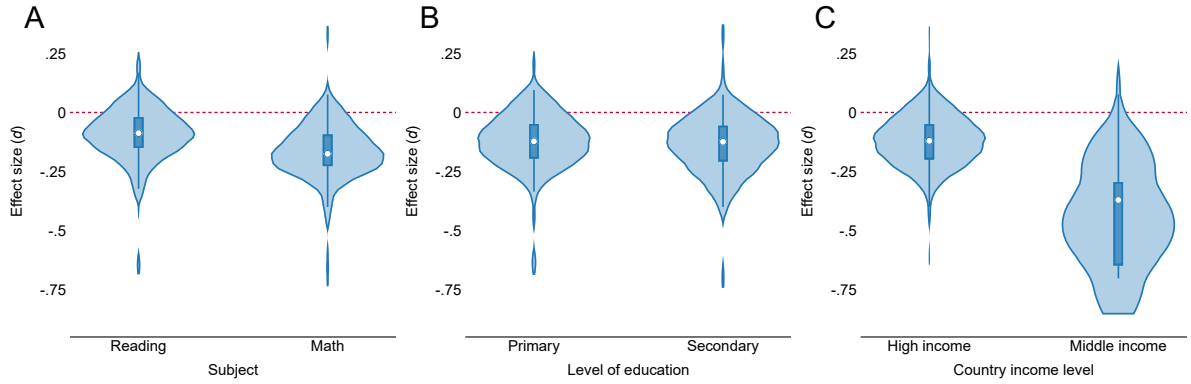
(b) Reproduction of Figure 4

Notes: The figure displays estimates of COVID-19 learning deficit. The horizontal axis shows the time of the estimate, and the vertical axis presents the estimates expressed as Cohen's *d*. Countries are in color scale. The slope coefficient of a trend line estimated using OLS with standard errors clustered at the study level is not statistically different from 0. (a) shows the original, and (b) is the reproduction. See Table A2 for details.

Figure A4: Variation in estimates of COVID-19 learning deficits



(a) Original of Figure 6



(b) Reproduction Figure 6

Notes: The figure displays variation in estimates of COVID-19 learning deficit for school subjects (mathematics and reading), level of education, and socio-economic inequality. (a) shows the original, and (b) is the reproduction. No differences between the two. See Table A3 for details.

A.4 Tables

Table A1: Replication Package Contents and Reproducibility

Replication Package Item	**Fully**	**Partial**	**No**
Raw data provided	✓		
Analysis data provided	✓		
Cleaning code provided	✓		
Analysis code provided	✓		
Reproducible from raw data		✓	
Reproducible from analysis data		✓	

Notes: This table summarises the replication package contents contained in Betthäuser et al. (2023a).

Table A2: Estimates of learning deficits in time

	Original Study	**Reproduction**
Slope coefficient: β_{months}	−0.00	−0.00
p-value	0.097	0.097
95% CI	[−0.01, 0.00]	[−0.01, 0.00]
Observations	291	291
Clusters	42	42

Notes: The table shows the comparison of the original and reproduced estimate of the slope coefficient obtained by regressing the estimates on months in which learning was measured. Standard errors are clustered at the study level. We report p-values and 95% confidence intervals (CI).

Table A3: Variation in estimates of learning deficits

	Original Study	**Reproduction**
School subject		
Reading	−0.09	−0.09
IQR	[−0.15, −0.02]	[−0.15, −0.02]
Mathematics	−0.18	−0.18
IQR	[−0.23, −0.09]	[−0.23, −0.09]
Mean difference	−0.07***	−0.07***
p-value	0.000	0.000
	[-0.11, -0.04]	[−0.11, −0.04]
Level of education		
Primary	−0.12	−0.12
IQR	[−0.19, −0.05]	[−0.19, −0.05]
Secondary	−0.12	−0.12
IQR	[−0.21, −0.06]	[−0.21, −0.06]
Mean difference	−0.01	−0.01
p-value	0.556	0.556
	[−0.06, 0.03]	[−0.06, 0.03]
Country income level		
High	−0.12	−0.12
IQR	[−0.20, −0.05]	[−0.20, −0.05]
Middle	−0.37	−0.37
IQR	[−0.65, −0.30]	[−0.65, −0.30]
Mean difference	−0.29***	−0.29***
p-value	0.008	0.008
	[−0.50, −0.08]	[−0.50, −0.08]

Notes: The table shows the comparison of the original and reproduced median learning deficit for school subjects, level of education, and country income level. IQR = Interquartile range as in the original paper. Significant at ***[1%], **[5%], *[10%] level.

B RTMA Implementation with Negative Affirmative Results

RTMA, as implemented in `phacking_meta` , assumes that p-hacking favors positive and significant results (`favor_positive = TRUE`). In our setting, affirmative results are negative and statistically significant since learning deficiency is measured as a negative value. We therefore apply RTMA to sign-reversed data ($y_i^f = -y_i$) with `favor_positive = TRUE` and reverse the sign of the resulting mean estimate.

We confirmed that this approach is equivalent to setting `favor_positive = FALSE` with the original data, as both specifications produce identical results. However, upon thorough examination of `phacking_meta`, `multibias_meta`, and `pubbias_meta`, we found that `favor_positive = FALSE` in `phacking_meta` effectively reverses the sign of the dataset internally before proceeding, producing estimates with inverted signs — which is identical to our manual sign-reversal approach. This behavior also affects `multibias_meta`. The function `pubbias_meta` does not suffer from this issue. We were able to correct this in `multibias_meta` but not in `phacking_meta`. We have reported this problem at https://github.com/mathurlabstanford/metabias-apps/issues/1, where a detailed description is available.

C Caliper Test and P-Curve Analysis

To formally test for p-hacking, we apply a caliper test following Gerber and Malhotra (2008) and Brodeur et al. (2016), which examines whether there is excess mass just inside the significance threshold relative to just outside it. Under p-hacking, researchers manipulate borderline non-significant estimates to cross the threshold, producing a spike just inside $z = -1.96$ and a corresponding gap just outside.

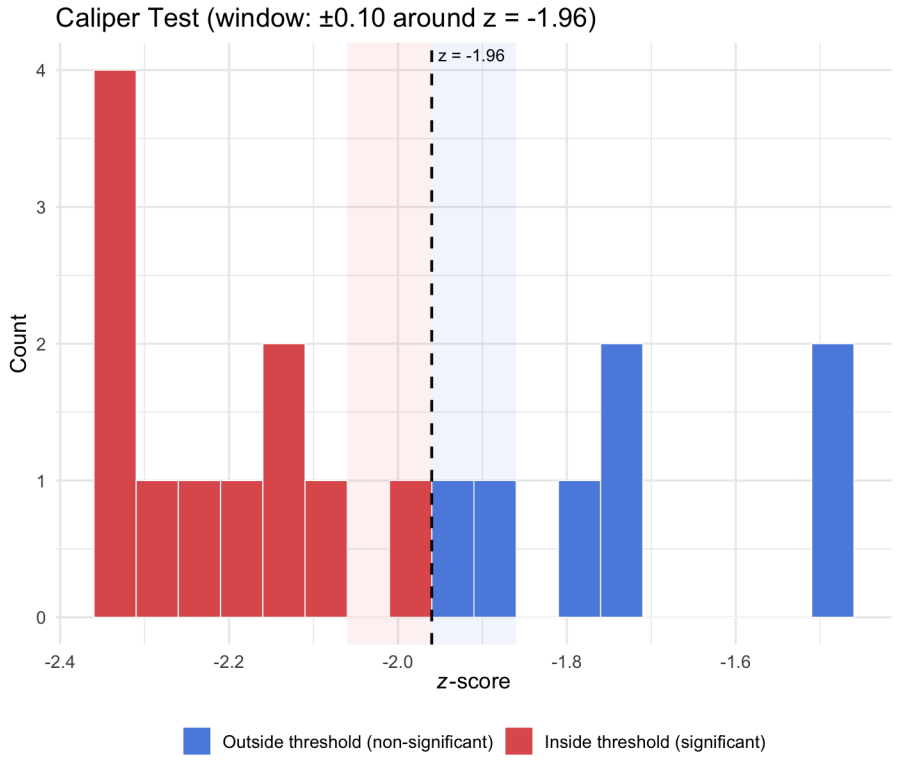
Across caliper widths from 0.05 to 0.50 and across the thresholds $t = 0$, $t = -1.96$, and $t = -2.58$, the caliper tests show little consistent evidence of bunching around reporting thresholds (Table C1, Figure C1). Because the relevant significance thresholds are negative, estimates below $t = -1.96$ are more negative and therefore lie on the statistically significant side of the threshold. At the standard caliper width of 0.10, at the $t = -1.96$ threshold is negative (-0.167, SE $= 0.333$, $N = 3$), indicating slightly more mass on the less negative, non-significant side of the threshold, opposite to what selective reporting of negative learning-loss estimates would predict. Only at wider calipers of 0.40 and 0.45 does the statistic become weakly significant at the 10% level. Overall, the caliper tests provide at most weak and non-robust evidence of bunching near the conventional significance cutoff.

We also attempted a p-curve analysis following Simonsohn et al. (2014), which tests whether the distribution of significant p-values is right-skewed as expected under genuine effects. However, the p-curve is structurally uninformative in this dataset. Sixty-one p-values underflow numerically to zero, and a further $\sim$175 produce $p < .001$, as a direct consequence of the extremely large sample sizes in the literature (median $n = 50{,}000$; max $n = 10{,}884{,}922$). Even modest effect sizes — such as $d = -0.27$ — generate z-scores as large as $|z| = 1{,}775$ when standard errors are of order 0.0002, collapsing virtually the entire p-value distribution into a narrow band indistinguishable from zero. The p-value distribution therefore carries no information about selective reporting, and we do not report formal p-curve test statistics (Figure C2).[12]

[12]This is a structural feature of the literature rather than a flaw in the analysis. Commercial platform studies with millions of observations have effectively infinite statistical power, making significance-threshold manipulation both unnecessary and undetectable through p-value-based diagnostics.

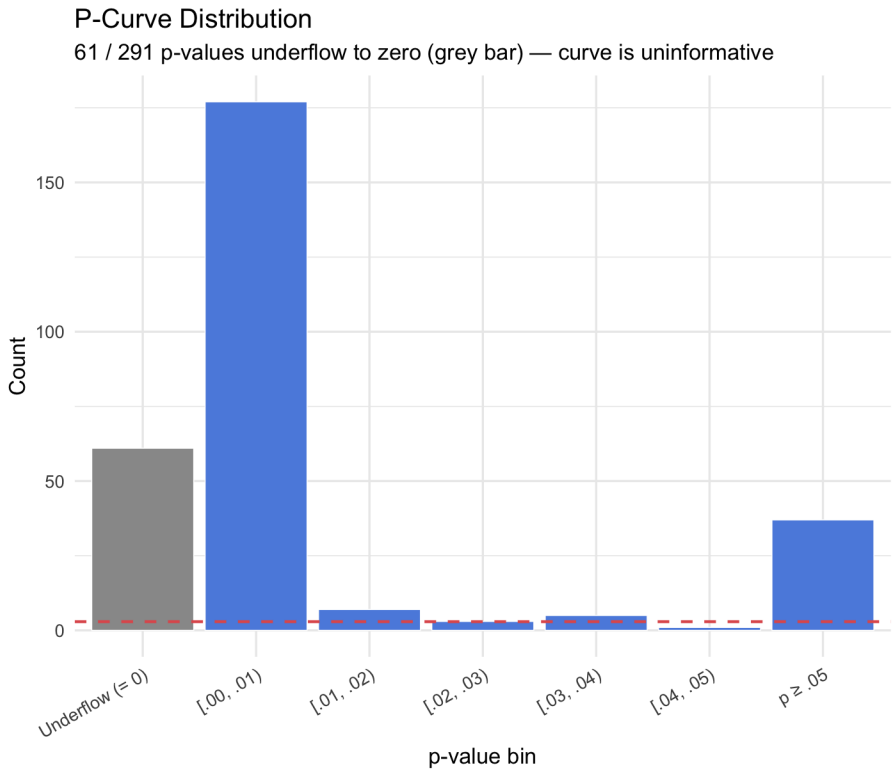
C.1 Figures

Figure C1: Caliper test for excess bunching around the significance threshold $z = -1.96$.



Notes: Shaded regions indicate the ±0.10 caliper window. Red bars denote estimates inside the threshold (significant); blue bars denote estimates outside (non-significant).

Figure C2: P-curve distribution.



Notes: The grey bar denotes p-values that underflow numerically to zero ($n = 61$). The dominance of the $[.00, .01)$ bin reflects the extremely large sample sizes in the literature rather than the evidential strength of individual studies.

C.2 Tables

Table C1: Caliper tests for selection around significance thresholds

	t-statistic = 0	t-statistic = −1.96	t-statistic = −2.58
Caliper .05	.	0.000	.
		(0.500)	
	N = 1	N = 2	N = 1
Caliper .1	0.000	-0.167	-0.167
	(0.500)	(0.333)	(0.333)
	N = 2	N = 3	N = 3
Caliper .15	0.000	0.000	-0.167
	(0.289)	(0.289)	(0.333)
	N = 4	N = 4	N = 3
Caliper .2	-0.167	0.071	0.000
	(0.203)	(0.202)	(0.289)
	N = 6	N = 7	N = 4
Caliper .25	-0.167	0.000	-0.250
	(0.203)	(0.155)	(0.143)
	N = 6	N = 10	N = 8
Caliper .3	-0.071	0.045	-0.250
	(0.202)	(0.138)	(0.143)
	N = 7	N = 11	N = 8
Caliper .35	-0.045	0.083	-0.167
	(0.141)	(0.138)	(0.110)
	N = 11	N = 12	N = 12
Caliper .4	-0.083	0.188*	-0.062
	(0.116)	(0.085)	(0.096)
	N = 12	N = 16	N = 16
Caliper .45	-0.083	0.188*	-0.062
	(0.116)	(0.085)	(0.096)
	N = 12	N = 16	N = 16
Caliper .5	-0.083	0.111	-0.100
	(0.116)	(0.108)	(0.110)
	N = 12	N = 18	N = 20

Notes: The table reports results for caliper tests (Gerber and Malhotra 2008). The tests compare the relative frequency of estimates above and below an important thresholds for the t-statistic; the rows show results for different caliper widths. A test statistic of −0.167, with $N = 3$ at −1.96 threshold means that 66.7% or 2 estimates are above the threshold and 1 estimates 33.3% is below. Figure C1 shows this graphically. N = number of estimates. Standard errors are reported in parentheses and clustered at the study level. Significant at ***[1%], **[5%], *[10%] level.

D Compliance with Meta-Analysis Guidelines

This appendix documents our adherence to the 18-item checklist for modern meta-analysis in economics proposed by Irsova et al. (2024). Because this paper reanalyzes an existing, publicly available dataset rather than conducting an independent literature search, several items in the checklist are not applicable by design (items 3–5, 9). Items 15–18, which concern multiple meta-regression and conditional implied estimates, are unfulfilled by design: the paper's contribution is confined to bias correction and reproducibility verification.

Table D1: Compliance with the Irsova et al. (2024) meta-analysis checklist

N	**Requirement**	**Status**	**Notes**
1	Topic known from own primary research	✓	Co-authors have published primary and meta-analytic research on publication bias, p-hacking, and the economics of education.
2	New meta-analysis justified by stronger methods	✓	Betthäuser et al. (2023a) do not apply formal bias-correction methods. Our *raison d'être* is the application of PET-PEESE, 3PSM, RoBMA, MAIVE, RTMA, and multi-bias sensitivity analysis to their dataset.
3	Google Scholar search, first 500 hits inspected	N/A	We build on the dataset of Betthäuser et al. (2023a) rather than conducting an independent literature search. Readers are referred to their Figure 1 and PROSPERO registration for the original PRISMA diagram (Appendix A).
4	Snowballing: 30 most-cited studies inspected	N/A	See item 3.
5	No study excluded *a priori* on quality grounds	N/A	See item 3. Study inclusion follows Betthäuser et al. (2023a).

Continued on next page

Table D1 continued

N	**Requirement**	**Status**	**Notes**
6	All estimates and standard errors collected	✓	All 291 effect-size estimates from 42 studies are included, as reported in Betthäuser et al. (2023a).
7	Data collected independently by two co-authors	✓*	The dataset is inherited from Betthäuser et al. (2023a) and publicly available; no independent re-coding of the primary effect-size estimates was performed.
8	Original effect-size measures used when comparable	✓	All estimates are standardized mean differences (Cohen's d), the measure used in the original literature.
9	Partial correlations used only as last resort	N/A	Cohen's d is used throughout; partial correlations are not employed.
10	Outliers and influence points inspected	✓	DFBETAS diagnostics are reported in Figure 9. Eleven influential observations are identified; excluding all eleven moves the pooled estimate from -0.124 to -0.116.
11	At least 10 heterogeneity variables coded	×	The paper's scope is confined to bias correction. Table 2 reports subgroup means by subject, school level, grade group, sample size, country, and source tier as descriptive context. Systematic heterogeneity analysis across moderator variables is available in Betthäuser et al. (2023a), who examine variation by subject, school level, socio-economic background, and country income level. A full multiple meta-regression is beyond the scope of the present paper.

Continued on next page

Table D1 continued

N	**Requirement**	**Status**	**Notes**
12	Simple summary statistic uses UWLS rather than FE/RE	✓*	The UWLS estimate (Stanley et al. 2023, Stanley and Doucouliagos 2017) yields the same inverse-variance weighted point estimate as the fixed-effect mean shown by the black diamond in the significance funnel plot ($\hat{\mu} = -0.245$). However, inference is based on a considerably more conservative cluster-robust standard error clustered at the study level (SE = 0.051), rather than the conventional fixed-effect standard error, which is implausibly small in this setting given the extreme heterogeneity ($I^2 = 99.97\%$). Panel A also reports unweighted and RVE means as uncorrected benchmarks, consistent with the paper's focus on bias correction rather than summary estimation. See footnote 4 in Section 3 for further discussion.
13	Publication bias corrected using methods from both model families	✓	Selection models: 3PSM and RoBMA, which averages across a wider class of selection models weighted by data fit and parsimony. Funnel-based models: PET-PEESE and MAIVE, the latter addressing the specific endogeneity problem arising from Cohen's d standardization and potential p-hacking. Additional sensitivity methods: MAN, selection-ratio adjustments, s-value, RTMA, and the multi-bias framework (Mathur 2024a,b,c).
14	Standard errors clustered at study level; wild bootstrap if < 40 studies	✓	Standard errors are clustered at the study level throughout (42 clusters). Wild bootstrap is recommended for fewer than 40 clusters and is not required here.

Continued on next page

Table D1 continued

N	**Requirement**	**Status**	**Notes**
15	Study-level dummies used in meta-regressions	×	Study-level fixed effects are not included in the main specifications. The principal estimators (PET-PEESE, MAIVE, 3PSM, RoBMA) do not incorporate study dummies; doing so would eliminate between-study variation that is the primary object of interest in a bias-correction exercise.
16	Multiple MRA estimated by BMA with dilution prior	×	A full multiple meta-regression with BMA is not conducted. The paper's contribution is confined to bias correction and sensitivity analysis.
17	Frequentist model averaging or general-to-specific as robustness check	×	See item 16.
18	Conditional means provided for different scenarios	×	Conditional implied estimates are not provided, consistent with the paper's focus on the mean bias-corrected effect rather than heterogeneity decomposition. Subgroup means by key moderators are reported in Table 2.

Notes: ✓ = satisfied; ✓* = substantially satisfied with acknowledged limitations; × = not satisfied; N/A = not applicable given the paper's design.

Reporting Standards (Havránek et al. 2020, Cook et al. 2026b,a)

The reporting standards of Havránek et al. (2020), updated for AI use by Cook et al. (2026b) following the guiding principles of Cook et al. (2026a), are addressed as follows. The research question and effect size measure (Cohen's d) are defined in Section 2. Study inclusion criteria and the PRISMA flow diagram follow Betthäuser et al. (2023a), to whose replication package we refer readers (Appendix A). Descriptive statistics are reported in Tables 1 and 2. Publication, selection, and misspecification biases are the central subject of Sections 3 and 4; bias-corrected estimates are reported in Tables 3 and 4. Within-study dependence is addressed via study-level clustering throughout. Effect sizes are displayed in Figures 1–6. Influence and leverage are assessed via DFBETAS (Figure 9); the sensitivity of the pooled estimate to excluding influential observations, to alternative correction methods, and to varying degrees of selective publication is documented across Sections 3 and 4. Economic significance is discussed using the benchmark of 0.40 SD per school year (Bloom et al. 2008). The underlying dataset is available at https://doi.org/10.17605/osf.io/u8gaz; replication code for all analyses is available at https://meta-analysis.cz. AI use is disclosed in the Use of Generative AI statement following the Data Availability Statement.