Buliskeria, Nino et al.

Working Paper

A Comment on "A Systematic Review and Meta-Analysis of the Evidence on Learning During the COVID-19 Pandemic"

I4R Discussion Paper Series, No. 223

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Buliskeria, Nino et al. (2025) : A Comment on "A Systematic Review and Meta-Analysis of the Evidence on Learning During the COVID-19 Pandemic", I4R Discussion Paper Series, No. 223, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/315577

INSTITUTE for REPLICATION

No. 223
I4R DISCUSSION PAPER SERIES

A Comment on "A Systematic Review and Meta-Analysis of the Evidence on Learning During the COVID-19 Pandemic"

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

April 2025

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 223

A Comment on "A Systematic Review and Meta-Analysis of the Evidence on Learning During the COVID-19 Pandemic"

Nino Buliskeria[1], Ali Elminejad[1], Tomas Havranek[2], Zuzana Irsova[2], Stepan Jurajda[3], Marek Kapicka[3], Martina Luskova[2]

[1]Nazarbayev University, Astana/Kazakhstan
[2]Charles University, Prague/Czech Republic
[3]Center for Economic Research and Graduate Education – Economics Institute (CERGE-EI), Prague/Czech Republic

APRIL 2025

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

A comment on "A Systematic Review and Meta-analysis of the Evidence on Learning During the COVID-19 Pandemic"*

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

March 25, 2025

Abstract

Betthäuser et al. (2023) examine the effects of the COVID-19 pandemic on the learning progress of school-aged children. They collect 291 estimates from 42 studies. Their meta-analysis-corrected estimate implies a substantial decline in students' learning (Cohen's $d = -0.14$, 95% confidence interval -0.17 to -0.10). First, we successfully reproduce the main results and the majority of supporting figures. Second, we provide additional analysis addressing publication bias by implementing correction techniques: PET-PEESE (funnel-based), 3PSM (selection model), and RoBMA (model averaging). Additionally, we implement novel approaches that account for the strength of biased selection favoring affirmative results in the sample of analyzed studies. Third, we use techniques that assume the presence of p-hacking (MAIVE, RTMA). Using these methods, the corrected effect ranges from -0.25 to -0.11 with high statistical significance. While our analysis does reveal some evidence of selection bias in underlying data (primary studies), these phenomena do not appear to systematically distort the overall findings of the original study.

KEYWORDS: Replication; Robustness; Meta-analysis; COVID-19; Education; Learning deficit

JEL CODES: I21, I24, I28, C68

*Authors: 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. E-mail: tomas.havranek@fsv.cuni.cz. Irsova: Institute of Economic Studies, Faculty of Social Sciences, Charles University. 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. Luskova: Institute of Economic Studies, Faculty of Social Sciences, Charles University. E-mail: martina.luskova@fsv.cuni.cz. The authors declare no conflicts of interest related to this work.

Acknowledgment: We thank the original authors, Bastian A. Betthäuser, Anders M. Bach-Mortensen, and Per Engzell, for their insightful comments and suggestions.

1 Introduction

Betthäuser et al. (2023) conduct a systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemic. The authors estimate the learning deficit using 291 estimates from 42 studies. As suggested by the authors, the pandemic led to a substantial decline in students' learning (Cohen's $d = -0.14$, 95% confidence interval -0.17 to -0.10), quantified using standardized test scores. Specifically, the study indicates that, during COVID-19, learning deficits were more pronounced in mathematics compared to reading within the same grade level. There were no significant differences in learning deficits across grade levels (mean difference $\delta = -0.01, t(41) = -0.59$, two-tailed p-value $= 0.556$, 95% CI -0.06 to 0.03). Moreover, the estimates of learning deficits are significantly higher in magnitude for middle-income countries than for high-income countries. The authors suggest on page 379 that "learning deficits opened up early in the pandemic and have neither closed nor substantially widened since then."

In the present report, we focus on three areas. (1) Narrow replication (reproduction). We check the reproducibility of all the results using the data and codes provided. (2) Publication bias. We use bias-correction techniques, such as PET-PEESE (funnel-based), 3PSM (selection model), and RoBMA (model average), and check if the results hold. (3) p-hacking. We use the techniques (RTMA, MAN) recently developed by Mathur (2024b) and MAIVE technique by Irsova et al. (2023). MAIVE is useful here since recomputing the estimates to Cohen's d introduces a mechanical correlation between estimates and standard errors, which breaks the assumptions of the study's meta-analysis model.

In section two, we successfully reproduce the findings of the original study and most of the figures using the code and data provided in the replication package. We do not reproduce the figures compiled manually by the original authors. In section three, we address publication bias. The original authors use only graphical tests (funnel plot, distribution of z-statistics, and p-curve) to test for publication bias. The authors conclude that there is no evidence of publication bias. We re-examine this conclusion using a variety of bias detection and correction methods such as PET-PEESE, 3PSM (a selection model as in Iyengar and Greenhouse (1988), Hedges (1992), Vevea and Hedges (1995)), and RoBMA (a model averaging approach de-

scribed in Bartoš et al. (2023), Maier et al. (2023)). We additionally corrected for bias by allowing for p-hacking using the MAIVE (meta-analysis instrumental variable estimation) by Irsova et al. (2023)), MAN (meta analysis of non-affirmative studies) and RTMA (right truncated meta-analysis) by Mathur (2024b). MAIVE also allows to address potential bias due to the mechanical correlation between estimates and standard errors introduced by standardization across studies using Cohen's *d*. Using PET, PEESE, 3PSM & RoBMA, the corrected effect size ranges from -0.245 to -0.118, indicating a statistically significant difference from zero. With 3PSM and RoBMA, we get results that are close to the original estimate. MAIVE yields an estimate of -0.119, which is highly statistically significant and is within the 95% confidence interval of the original result. Finally, we conduct sensitivity analysis following methods by Mathur (2024b) and get estimates between -0.206 and -0.039, which are highly statistically significant. While these results suggest the correct direction of the effect, the wide range of estimates indicates the need for further investigation into the model fit.

We consider the resulting estimates from RoBMA and MAIVE to be particularly reliable, since these methods are robust against p-hacking and the transformation to Cohens d. Therefore, we conclude that even though our analysis uncovers certain signs of publication bias and *p*-hacking, these phenomena do not seem to systematically affect the conclusions of the original study.

2 Computational Reproducibility

We used the replication package provided here: hyperlink[1]. 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 1 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 re-

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

produced the figure using the same code directly in R. In section 5, we present the reproduced figures; tables are presented in section 6. Table 2 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 3 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).

2.1 Discrepancies Between Pre-analysis Plan and Article

The authors registered a pre-analysis plan available here: hyperlink[2]. 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

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

p-value = 0.05). Additionally, the supplementary 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.

3 Robustness Reproduction and Replication Using Different Models

Betthäuser et al. (2023) utilized graphical tests to identify potential publication bias. Firstly, they use a graphical test based on the distribution of z-statistics. This approach is based on the assumption that in the presence of publication bias, there should be a notable jump in the distribution of z-statistics at the significance threshold where $z = -1.96$. However, recent work by Elliott et al. (2022b) argues that the presence of said jump does not always indicate publication bias. In reproduced Figure 1, the log-transformed z-statistics do not exhibit any clustering around the 1.96 significance level. Yet, it is important to note that in the distribution of the raw z-statistics, this clustering is clearly present. We have put these two plots together in Figure 8 for ease of comparison. In the (c) plot of Figure 8, the z-scores are transformed and displayed on a log scale, which changes their visual representation. In the plots (a) and (b), we present the distribution of raw absolute values of z-scores without any transformation reflecting their natural spread. Taking the logarithm of the z-scores significantly alters the distribution. Logarithmic transformation compresses higher values and spreads out smaller ones, making the distribution look different compared to the untransformed (or absolute) z-scores. Figure 8 (a) presents strong bunching at the 5% significance level and a long heavy left-skewed tail, suggesting a preference for negative significant results. However, upon closer examination, Figure 8(b) shows that the distribution of z-scores is somewhat symmetric around its peak at -1.96, and can also be caused, for example, by a relatively large share of studies being similarly powered or having a similar sample size. Therefore, the patterns observed in Figure 8(b) align with those in the original log-transformed z-score distribution (Figure 1), reinforcing the interpretation that the observed clustering around the -1.96 distribution may not

be attributed to publication bias.

Secondly, the reader can find two more visual tests in the Betthäuser et al. (2023) supplementary material. First, a test based on the p-curve, assuming that if there is publication bias or p-hacking, the distribution of p-values should be left-skewed with a large number of p-values right below the 0.05 significance threshold. Second, the authors include a funnel plot, showing the relationship between effects and their standard errors. Based on these three methods, the authors conclude that there is no evidence of publication bias. While these methods offer valuable initial insights, we sought to enhance the robustness of the analysis by employing more rigorous techniques. In our robustness analysis, we test for publication bias and assess whether the findings are consistent with the original study using both established workhorse meta-analysis methods and more novel techniques detailed below.

We start by estimating funnel-based precision effect test - precision effect estimate with standard errors (PET-PEESE). The precision effect test (PET) is based on the regression of the effect on its standard error, weighted by $1/SE^2$, the squared precision[3]. The coefficient of the standard error in our PET analysis is highly statistically significant (see Table 4, Column 1). In Column 2, we estimate the PEESE since it offers a more accurate effect-size approximation (Bartoš et al. 2022, Stanley 2017). To obtain the PEESE estimate, we regress the effect on its squared standard error weighted by the squared precision $1/SE^2$. Based on heteroskedasticity robust standard errors, the publication bias coefficient is statistically significant only at 10%. The corrected effect of -0.245 suggests that the learning deficit is 0.245/0.4 or 0.61 of a school year's learning. This estimate is much larger compared to the 0.35 reported in the article.

Column 3 in Table 4 shows the results from 3PSM, a bias correction method based on the publication selection model described in Iyengar and Greenhouse (1988), and Hedges (1992). Publication selection arises with a preference for p-values below the significance threshold. We estimate the effect beyond bias using the selection model that employs the maximum likelihood method described in Vevea and Hedges (1995). The corrected effect size -0.123 and a corresponding

[3]Under no publication bias, there should be no relationship between the two (Egger et al. 1997, Stanley and Doucouliagos 2014)

learning deficit expressed as a school year's worth of learning $0.123/0.4 = 0.31$ are close to the estimate reported in the article. The 3PSM features a likelihood ratio test for publication bias. According to the p-value, we cannot reject the null hypothesis of no publication bias.

Next, we apply Robust Bayesian Meta-Analysis (RoBMA), which estimates different publication bias models and constructs a weighted average over them, with the weights being proportional to data fit and model parsimony (Bartoš et al. 2023, Maier et al. 2023). The RoBMA estimate, shown in Table 4, Column 4, is -0.118, corresponding to $0.118/0.4 = 0.30$ school year's worth of learning deficit, slightly smaller than the original article's estimate of $0.14/0.4 = 0.35$. This methodology is robust to misspecifications and performs well under heterogeneity. Given that RoBMA integrates and balances multiple commonly used publication bias models, the resulting estimate is particularly reliable.

In addition to conventional meta-analysis techniques, we conduct sensitivity analyses offered by Mathur and VanderWeele for publication bias in metaanalyses, described in Mathur and VanderWeele (2020), Mathur (2024a,b). We utilize the `PublicationBias` & `phacking` packages available at metabias.io. These methods are designed to perform sensitivity analyses for publication bias, where affirmative studies – those showing statistically significant results in the expected direction – are more likely to be published than non-affirmative studies, which include those with non-significant results or results in an unexpected direction. This bias is quantified by the *selection ratio* (Mathur and VanderWeele 2020). To clarify, consider a scenario where affirmative studies are known to be published twice as often as non-affirmative ones. In this case, Mathur and VanderWeele adjust the meta-analytic estimate by giving each non-affirmative study twice the weight of an affirmative study in the analysis. This adjustment would neutralize the publication process's bias that favors affirmative studies by a factor of two. In practice, however, the exact degree of publication bias is usually unknown. To address this uncertainty, Mathur and VanderWeele propose conducting sensitivity analyses that assess how much publication bias (i.e., the selection ratio) would be necessary to negate the results of the meta-analysis, such as shifting the point estimate to a null effect.

In the most extreme case, where affirmative studies are infinitely more likely to

be published than non-affirmative studies, the corrected estimate would be achieved by assigning infinitely more weight to non-affirmative studies. For example, in estimating COVID-19's impact on learning and the resulting learning deficiency, an affirmative result would be one that is negative and statistically significant (as opposed to the conventional definition of the affirmative result being positive and significant).[4] This, in fact, corresponds to retaining only the non-affirmative studies in analysis. Thus, to account for worst-case publication bias, we can conduct a meta-analysis of non-affirmative results (MAN) on a subset of only positive coefficients. Mathur (2024a) notes that the MAN method does not measure the actual strength of publication bias; instead, it is used in sensitivity analysis to assess how results are influenced by an extreme, hypothetical level of publication bias. Although such severe bias is highly unlikely, if the worst-case estimate still aligns with the uncorrected estimate and remains of meaningful size, it strongly suggests that the results are robust against potential publication bias. This would imply that both affirmative and non-affirmative estimates predict the same mean outcome – whether we consider only affirmative (negative coefficients) or only non-affirmative (positive coefficients) results, the estimated intercepts (mean outcomes) from the meta-regressions of coefficients on their standard errors would be similar. However, if the worst-case estimate is near null or shifts in the opposite direction of the uncorrected estimate, the meta-analysis may not be resilient to worst-case publication bias (Mathur 2024a).

As a benchmark point, we present the standard uncorrected point estimates using a fixed-effects model (-0.126, $SE = 0.059$) and a robust random-effects model that accounts for heterogeneity and clustering (-0.140, $SE = 0.020$) in Table 5, Column 1 and Column 2 of Panel A. The estimates align with that of the original study primarily because they represent the uncorrected mean estimate, assuming no selection bias. We then examine the sensitivity of this uncorrected mean result – specifically, the original study's coefficient of -0.14 – to a hypothetical worst-case scenario of selection bias. In Panel B Column 1 of Table 5, the MAN suggests a positive point estimate corrected for bias of 0.021 significant at the 10% level.

[4]In Figures 1 through 5, it is clear that most coefficient estimates are negative, which is expected since learning deficiency is measured as a negative value. We account for the negative sign and adjust the model specifications accordingly, assuming a one-tailed model of publication bias and a significance threshold of $\alpha = 0.05$.

Hence, we have obtained an estimate that has the opposite sign (and is not different from 0 at 5% significance level) to the uncorrected original estimate, suggesting that results are not robust to the hypothetical, extreme case of selection bias. In this case, Mathur (2024a) suggests conducting additional sensitivity analyses focused on less severe publication bias to assess robustness, techniques presented in Mathur and VanderWeele (2020). Following Mathur and VanderWeele, next, we look at less severe publication bias cases, as well as estimating the extent of bias required for mean beyond bias to be zero or positive – the strength of publication bias required to explain away the original results.

Selection ratio Rather than assuming a worst-case selection, we continue by introducing a selection ratio, allowing us to choose the severity of publication bias. This measure reflects the relative likelihood of studies with significant estimates being published compared to those with non-significant estimates. By definition, if there is no preference for significant results, the selection ratio equals one. Since the severity of publication bias is not known in advance, we can assume the magnitude of the selection ratio to perform the sensitivity analysis. We follow Mathur (2024b) & Mathur and VanderWeele (2020) and use fixed- & random-effects meta-analysis while assuming a four-fold preference (selection ratio = 4) for significant results (affirmative studies). Columns 2 & 3 of Panel B report the results from fixed- & random-effects specifications with a pre-defined four-fold selection ratio. The fixed-effect estimate indicates that if affirmative (in this case, significant and negative) studies were four times more likely to be published than non-affirmative ones, the meta-analytic point estimate corrected for publication bias would be −0.206 (95% CI: −0.206 to −0.206). The robust random-effect estimate is rather smaller, due to accounting for heterogeneity and clustering of point estimates within papers; it is also based on the four-fold preference for affirmative studies. It suggests that the meta-analytic point estimate corrected for publication bias is −0.068 (95% CI: −0.101 to −0.036). Both these estimates support the sign direction of the uncorrected mean; while the fixed-effect estimate is larger and the random-effect estimate is smaller than the original one, the mean of these two falls close to the original result, suggesting relative robustness towards four-fold selection bias.

s-Value that explains away results The next exercise explores how strong a

preference for selecting negative and significant results is required for the true effect, beyond bias, to be zero or even positive. In Panel C of Table 5, we calculate the selection ratio, or s-value, required to shift the estimate or its confidence interval bound to a specific value. Column 1 of Table 5 shows that affirmative studies would need to have a 29.60 times higher publication probability than non-affirmative studies for the point estimate to change to 0. Similarly, for the confidence interval bound to shift to 0, affirmative studies would require an s-value = 8.31 fold higher publication probability. When examining the severity of publication bias necessary to shift the point estimate or its confidence interval bound to 0.05, we find that no amount of publication bias can achieve that shift for either the estimate or the confidence interval (Table 5, Panel C, Column 3). These results suggest that, under this model's specifications, no degree of bias would shift the point estimate or the confidence interval bound to 0.05 or beyond.

Generally, a sufficiently small selection ratio, s-value, representing a plausible degree of publication bias, indicates that the meta-analysis is relatively sensitive to publication bias. Conversely, if the s-value corresponds to an implausibly large degree of publication bias, the meta-analysis can be considered relatively robust. Column 1 in Panel C of Table 5 shows that a very strong publication bias would be required to "explain away" the results of the original meta-analysis. Furthermore, Column 3 shows that no amount of publication bias, under the assumed model, can shift the results of the original meta-analysis "to a positive effect of 0.05", thereby providing strong evidence for the robustness of the findings.

Significance funnel plot As a supplement to the sensitivity analysis, we present "significance funnel plot" in Figure 5, which illustrates the difference between affirmative and non-affirmative point estimates. This figure generally aids in assessing the extent to which the point estimates from non-affirmative studies are systematically smaller (in absolute terms) than the entire set of point estimates. Since the sample consists of the point estimates from the studies that investigate at "what extent has the learning progress of school-aged children slowed down during the COVID-19 pandemic", the positive coefficients in Figure 5, shown in gray, correspond to non-affirmative results, as they indicate non-significant negative or a positive effect of COVID-19 on learning progress. In this case, negative point

estimates, shown in yellow, reflect the negative impact of COVID-19 on learning progress, with lower coefficients suggesting a stronger estimated negative impact of COVID-19 on learning progress. Therefore, this significance funnel plot aids us in understanding how much lower, thus more negative, point estimates from affirmative studies are compared to non-affirmative studies. The mean estimate based solely on non-affirmative studies (gray diamond) represents an estimate corrected for publication bias under the worst-case scenario. If this gray diamond shows a negligible effect size or is significantly smaller, in absolute value, than the pooled estimate across all studies (black diamond), it indicates that the meta-analysis may be vulnerable to severe publication bias. The corrected point estimate under the worst-case biased selections scenario is 0.021 (95% CI: −0.003 to 0.045) as shown in Table 5, Panel B, Column 1. Meanwhile, the pooled estimate (black diamond) is approximately −0.25. Since the black and gray diamonds are significantly different, and the mean of non-affirmative results is near zero, the observed pooled effect may be somewhat exaggerated due to potential selection bias Mathur and VanderWeele (2020). Note, however, that the mean of non-affirmative results is almost always a downward-biased estimate of the underlying mean effect. The significance funnel plot also reveals that the reported pooled mean is substantially larger than the mean of the most precisely estimated coefficients, which is also consistent with publication bias in most meta-analysis models.

Allowing for p-hacking Next, we look at the models that account for p-hacking and relax the assumption of the unbiasedness of point estimates, a key premise in conventional meta-analytic approaches. By doing so, these models control for the potential biases introduced by selective reporting and manipulation of statistical significance. p-Hacking is an umbrella term referring to practices of adjusting the p-values to achieve statistical significance.[5] p-Hacking practices include but are not limited to continuing data collection until a significance threshold is met, adjusting samples, re-selecting covariates, or fitting multiple models in an attempt to obtain affirmative results. Mathur (2024b) defines p-hacking as "selection within study" and publication bias as "selection across studies".[6] Conventional methods

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

[6]There are various incrementally different definitions of "publication bias" in the literature, but in this work, we adopt this definition by Mathur (2024b): **Publication bias** can encompass decisions

for detecting publication bias may produce biased results in either direction when p-hacking is present (Mathur 2024b). However, when p-hacking favors affirmative outcomes, Mathur asserts that a meta-analysis of non-affirmative results (MAN, Table 5, Panel B) still provides a conservative estimate, biased toward the null.

MAIVE In addition to MAN, we apply the Meta-Analysis Instrumental Variable Estimator (MAIVE) by Irsova et al. (2023) to measure for the extent of publication bias and estimate corrected mean effect of COVID-19 on learning corrected for bias while allowing for the existence of p-hacking. Contrary to most conventional methods that assume unbiasedness of point estimates, MAIVE relaxes this assumption. Meta-analysis methods, such as PET and PEESE techniques, often give more weight to studies with lower standard errors, implying greater precision. However, in empirical research, particularly in observational studies, precision is not directly observed but must be estimated by the researcher. As widely noted in the literature, this estimation process is susceptible to p-hacking, where precision is artificially manipulated to achieve statistically significant results (Brodeur et al. 2023, Irsova et al. 2023, Mathur 2024b). Irsova et al. (2023) describe how p-hacking practices can introduce biases in both the precision and the coefficient estimates of the original studies that serve as the sample for meta-analyses, introducing the problem of endogeneity. They further demonstrate through simulations that even a small degree of spurious precision can significantly undermine the effectiveness of inverse-variance weighting and bias-correction methods that rely on funnel plots. Selection models designed to address publication bias often fail to resolve this issue; in some cases, a simple average may outperform more complex estimators. As they put it, "Cures to publication bias may become worse than the disease" (Irsova et al. 2023, pg. 1).

The statistical solution to this endogeneity issue is to identify an instrument for the standard error. A valid instrument must be correlated with the standard error but uncorrelated with the error term, making it independent of the sources of endogeneity. Irsova et al. suggest to instrument standard errors by the respective sample size. By definition, the squared standard error (SE^2) is a linear function

made by a study's investigators to withhold the study from submission to journals entirely, as well as decisions by journal editors and reviewers to reject the study, both based on whether affirmative results are present or absent. For a more in-depth discussion, please see Mathur and VanderWeele (2020), Mathur (2024b).

of the inverse of the sample size used in the primary study. Respective sample size is likely to be resistant to selection bias, as it is generally more difficult to increase sample size than to manipulate the standard error to achieve significance. Additionally, the sample size is free from measurement error since it is a direct measure, not an estimate. Unlike the standard error, the sample size is typically unaffected by changes in methodology. A potential weakness, as described by the authors, is that some endogeneity might persist if researchers anticipating smaller effects design larger studies. However, in observational research, the sample size, unlike the standard error, is often predetermined.

In addition to addressing the potential existence of p-hacking, MAIVE is useful here since the authors of the original paper recompute all results to Cohen's d. Doing so introduces a mechanical/spurious correlation between estimates and standard errors, so the core assumption of the basic meta-analysis model – independence of coefficient estimates and their standard errors in the absence of publication bias – is violated. Following MAIVE, we regress the squared reported standard errors on the inverse sample size and use the predicted values in place of the variance on the right-hand side of our meta-regression. For the baseline MAIVE, we select the instrumented version of PEESE (Precision-Effect Estimate with Standard Errors) without additional inverse-variance weighting and with clustering at the study level as is recommended in Irsova et al. (2023).

The statistically significant corrected estimate of -0.119, along with an estimate of the extent of bias at -0.245, suggests the presence of publication bias in the literature. Nevertheless, the MAIVE estimate of the mean beyond bias, presented in Column 1 of Panel D, Table 5, reinforces the findings of the original paper, closely aligning with the original mean estimate of -0.14. The MAIVE coefficient of -0.119 suggests a learning deficit equivalent to $0.119/0.4 = 0.30$ school years due to COVID-19. This result is consistent with RoBMA, and both should be emphasized as the main results due to their robustness against p-hacking and the transformation to Cohen's d. This near-perfect consistency among the original paper, RoBMA, and MAIVE indicates a $30 - 35\%$ school year learning deficit due to COVID-19.

RTMA To estimate the extent of p-hacking, we employ Mathur's right-truncated

meta-analysis (RTMA) method, which deals with both publication bias and p-hacking. We utilize the `PublicationBias` & `phacking` packages at metabias.io. As described in Mathur (2024b), RTMA is correctly specified if the favored estimates in hacked studies are always affirmative – meaning they yield significant, positive results as investigators continue generating estimates until the first affirmative result is obtained – or if hacked studies with non-affirmative favored estimates if any exist, are never published. Additionally, RTMA accounts for within-study heterogeneity, allows for both independent and autocorrelated estimates within studies, and provides a framework for inference. For our purposes, we need to adjust our data to fit RTMA assumptions, specifically affirmative data being positive and significant. For this, we first use the inverse sign of coefficient estimates in the data and then apply RTMA. This gives us an estimate for effect beyond bias with a reverse sign.[7] The resulting RTMA estimated mean beyond bias is reported in the second column of Panel D, Table 5. Accounting for potential p-hacking and publication bias, the estimated meta-analytic mean is -0.039, and the estimated standard deviation of the effects (i.e., heterogeneity) is 0.072.[8] The RTMA estimate is notably closer to zero than the estimate from the MAIVE approach. Before drawing any further conclusions, we evaluate the model fit of RTMA and examine its underlying distributional assumptions.

Diagnostic Plots To examine the fit of RTMA and determine if the distributional assumptions are appropriate, we plot the diagnostic q-q plot. We plot the fitted CDF of point estimates against the empirical CDF of point estimates in Fig-

[7]We are compelled to approach the exercise this way because adjusting the RTMA specification by setting `favor_positive = FALSE`, while technically feasible, effectively reverses the sign of the entire dataset and then continues to use this reversed data, leading to estimates with inverted signs (same exercise that we describe here). We reached this conclusion after obtaining identical results when using `favor_positive = FALSE` specification with original data y_i; and `favor_positive = TRUE` with original data but with reversed sign, $y_i^f = -y_i$, with `multibias_meta` & `phacking_meta`. We followed with thoroughly examining `phacking_meta`, `multibias_meta`, and `pubbias_meta` packages. After a detailed examination of the `phacking_meta`, `multibias_meta`, and `pubbias_meta` functions, we became even more confident in our conclusion. While we were able to correct this minor coding error in `multibias_meta`, our attempts were less successful with `phacking_meta`. This issue arises with the `phacking_meta` function, which estimates RTMA, as well as with `multibias_meta`, which estimates multiple biases, as described below. The function `pubbias_meta` did not suffer from this issue. We have reported this problem at https://github.com/mathurlabstanford/metabias-apps/issues/1, where you can see the detailed description of the issue.

[8]As described in footnote 6, we obtained positive coefficients and reversed the signs of mean estimates after.

ure 6. The points, particularly at the upper part of the distribution, do not adhere closely to the 45-degree line, suggesting that the RTMA model does not adequately fit the data. As a second diagnostic plot, we show the distribution of z-scores of all point estimates in the data. p-hacking, in our case, would favor negative over positive results, and as observed in Figure 7, z-scores disproportionately gather at a negative critical threshold, -1.95. Since here we are reversing the sign of our coefficient estimates before applying RTMA, bunching at the negative significance threshold indicates that RTMA assumptions of favoring positive and significant coefficients are satisfied in our sign-reversed data. In Figure 7, we show the z-score distribution, measured by the ratio of coefficient estimates and their standard errors, with the original dataset without sign reversal. The plot shows a distribution that is highly skewed, with a long tail to the left and a sharp peak just at z-score $= -1.96$, 5% significance level, which is consistent with a preference for negative significant results. However, upon closer examination, Figure 8 (b) shows that the distribution of z-scores is somewhat symmetric around its peak at -1.96, and can also be caused, for example, by a relatively large share of studies being similarly powered or having a similar sample size. Moreover, Elliott et al. (2022b) argues that a jump around 1.96 does not necessarily indicate publication bias, further supporting the view that the z-curve on the figure 7 & 8 provides no evidence of such bias.

Multiple Biases In addition to addressing publication bias & p-hacking, Mathur points out that meta-analysis can suffer from internal biases coming from individual studies, for example, confounding in non-randomized studies. The interaction of these biases, coming from individual studies and selection preferences, is often non-linear and non-additive. For example, publication bias that favors significant, positive outcomes may lead to selecting studies with greater internal bias. To address this issue, Mathur (2024c) proposes sensitivity analysis addressing two main questions: (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?" These methods, elaborated in Mathur, account for the average internal bias across studies, eliminating the need

to specify the bias in each individual study. The model specifications can be configured to either assume that internal bias affects all studies or, alternatively, that it only impacts a specific subset (e.g., nonrandomized studies). Additionally, the model can assume that internal biases are of unknown origin or, for certain types of bias in causal estimates, analytically bounded. The model can further specify the severity of publication bias or, alternatively, consider a worst-case form of publication bias. Robust estimation methods can handle non-normal effects, small sample sizes in meta-analyses, and clustered estimates. These methods developed by Mathur (2024c) can offer insights that are not revealed by addressing each bias individually. It is often assumed that randomized studies are free from bias, and different levels of bias could be assigned to various study types. However, due to the lack of information on which studies were randomized, we assume that internal bias uniformly affects all studies, regardless of the methodology used to obtain estimates.[9] Therefore, we only account for varying degrees of bias between affirmative and non-affirmative results.

Columns 3 & 4, Panel D in Table 5 present the estimates from the multi-bias analyses. We can take RE-SR4 (Column 3, Panel B) as a benchmark estimate that assumes no internal bias and four-fold selection bias, s-value = 4 – affirmative studies being four times more likely to be published than non-affirmative studies. Under these conditions, the corrected meta-analytic point estimate is −0.068 (95% CI: −0.101 to −0.036). However, if in addition to assumption *N1) affirmative studies are four times more likely to be published than non-affirmative studies*, we assume *N2) affirmative studies have a mean internal bias of* 0.05, and *N3) non-affirmative studies have a mean internal bias of* 0.01, which indicates very little bias,[10] corrected meta-analytic point estimate would be −0.097 (95% CI: −0.112 to −0.082; see Column 3, Panel D). This estimate is larger than the benchmark and is close to the RoBMA and MAIVE estimates reported in Table 4 and Table 5, respectively. The results align even more closely with our main findings from RoBMA and MAIVE when we assume a greater selection bias in affirmative studies, with a

[9]We plan to collect additional characteristics of primary studies in the data, including details on the use of randomization. This will allow us to conduct multi-bias analysis more thoroughly.

[10]Given Figure 7, we can confidently conclude that selection bias is either absent or minimal in non-affirmative, positive studies. The result of this multi-bias analysis closely aligns with the scenario where we assume minimal bias in non-affirmative studies, set at 0.01.

mean internal bias of 0.08. These findings indicate that the extent of mean internal bias differs between affirmative and non-affirmative studies, with affirmative studies experiencing significantly larger selection bias.

4 Conclusion

Betthäuser et al. (2023) perform a meta-analysis of the effect of the COVID-19 pandemic on the learning progress of school-aged children (also referred to as learning deficit). Their results suggest that the pandemic led to a substantial decline in students' learning deficits (Cohen's $d = -0.14$, 95% confidence interval -0.17 to -0.10). This report focuses on the narrow reproduction (replication) of the study's results and robustness reproduction. In the second part of the report, we employ various methods and new techniques to adjust the corrected estimate for publication bias and p-hacking. In the narrow reproduction, we reproduced all the main findings and the supporting figures as in the original study. We used the replication package containing both data and code.

To test for publication bias, the authors use a graphical test based on the distribution of z-statistics. Supplementary material features two more visual tests: a funnel plot and a test based on the p-curve. Based on these, the study concludes that there is no evidence of publication bias. Our PET-PPESE publication bias-corrected estimate -0.245 is highly statistically significant and almost double in magnitude compared to the study's own corrected estimate. For 3PSM, a selection model as in Iyengar and Greenhouse (1988), Hedges (1992), Vevea and Hedges (1995), the estimate equals -0.123 and is highly statistically significant. Next, we use RoBMA, which performs well under heterogeneity and yields an estimate of -0.118, which is highly statistically significant. These corrected effects are close to the original but slightly closer to zero.

Assuming a four-fold preference for studies with significant estimates, the corrected estimate is -0.206 under the fixed-effects specification and -0.068 under the random-effects specification, both of which are highly statistically significant. In contrast, the meta-analysis of non-affirmative studies (those with non-significant results or results in the undesirable direction), denoted as MAN, yields a positive

corrected estimate of 0.021, statistically significant at the 10% level. MAN indicates that this coefficient could represent the true mean beyond bias (corrected for selection bias) in a worst-case scenario, where only studies with affirmative (negative and significant) results are published. Reversing this selection means that MAN gives 100% weight to positive & non-significant values, i.e., non-affirmative studies. Next, we examine the degree of selection bias needed for the true mean beyond bias to be zero or positive, specifically testing for values of 0, 0.01, and 0.05. We find that the preference for publishing affirmative results would need to be thirty times stronger than for non-affirmative results for the mean beyond bias to reach zero. No amount of selection bias could push this estimate to 0.05.

To account for the spurious correlation between estimates and their standard errors, which may arise due to *p*-hacking and the normalization of estimates using Cohen's *d*, we apply the MAIVE technique (Irsova et al. 2023). This method controls for this endogeneity by using inverse sample size as an instrument for the standard errors. MAIVE yields an estimate of -0.119, which is highly statistically significant and suggests the existence of publication bias. To specifically correct for the *p*-hacking, we apply RTMA on original data with an inverse sign. We obtain a statistically significant corrected estimate of -0.039. While the direction of the effect aligns with that of the original paper, the small estimate is likely due to RTMA overcorrecting for p-hacking as a result of inadequate model fit.

Finally, if we assume different extents of biases in affirmative (0.05 or 0.08) and non-affirmative (0.01) studies, the estimate is similar to that of MAIVE & RoBMA. These findings indicate that the extent of mean internal bias differs between affirmative and non-affirmative studies, with affirmative studies experiencing a significantly larger selection bias.

The RoBMA integrates and balances multiple commonly used publication bias models. Both RoBMA and MAIVE are robust against *p*-hacking and the transformation to Cohen's *d*. Therefore, the resulting estimates from these methods are particularly reliable. The near-perfect consistency among the original paper, RoBMA, and MAIVE indicates robustness of the original conclusion that COVID-19 caused a $30 - 35\%$ school year learning deficit. Although the RoBMA and MAIVE estimates fall at the lower bound, the slight difference in the mean beyond

bias estimate does not significantly impact the final effect of COVID-19 on learning. Therefore, while our analysis does reveal some evidence of selection bias in the underlying data and predicts a somewhat smaller effect size, these phenomena do not appear to systematically distort the overall findings of the original study.

References

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.: 2023, A systematic review and meta-analysis of the evidence on learning during the covid-19 pandemic, *Nature human behaviour* **7**(3), 375–385.

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.

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

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* .

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

Irsova, Z., Bom, P. R., Havranek, T. and Rachinger, H.: 2023, Spurious precision in meta-analysis. Available at https://www.econstor.eu/handle/10419/286334.

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

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.

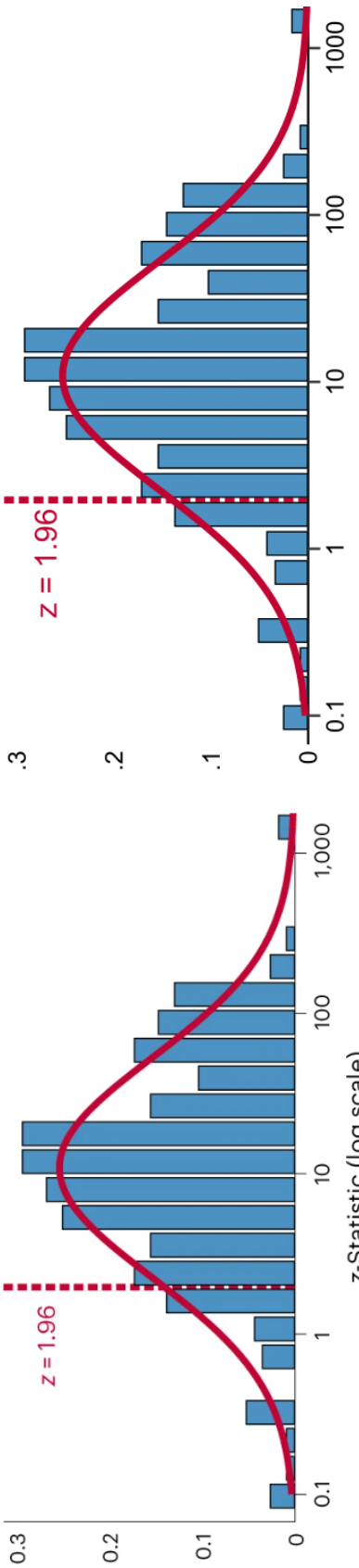
Stanley, T. D.: 2017, Limitations of pet-peese and other meta-analysis methods, *Social Psychological and Personality Science* **8**(5), 581–591.

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

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.

5 Figures

Figure 1: Publication bias: distribution of z-scores

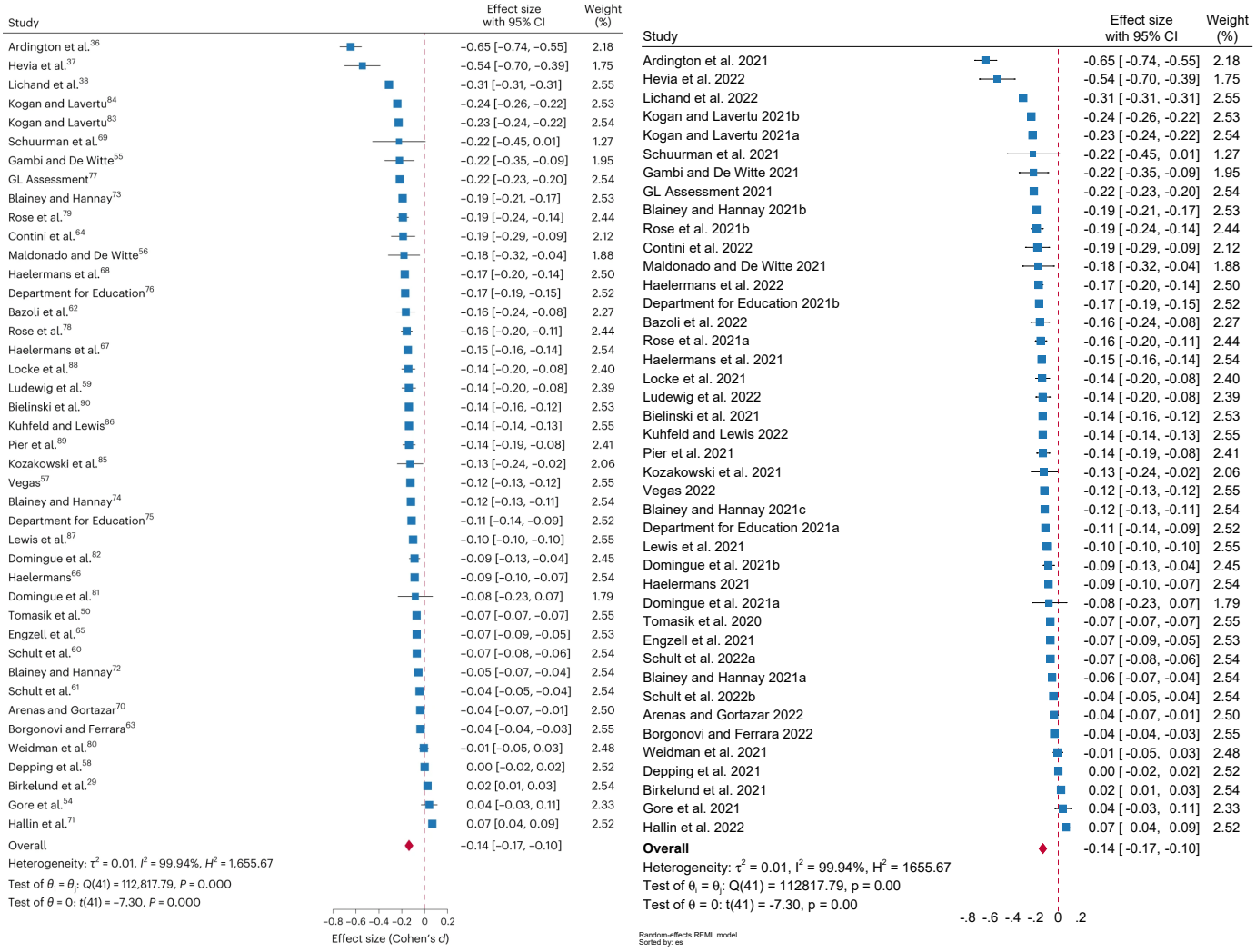


(a) Original Figure 2b

(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 reproduction, and (b) is the original. There is no difference between the two.

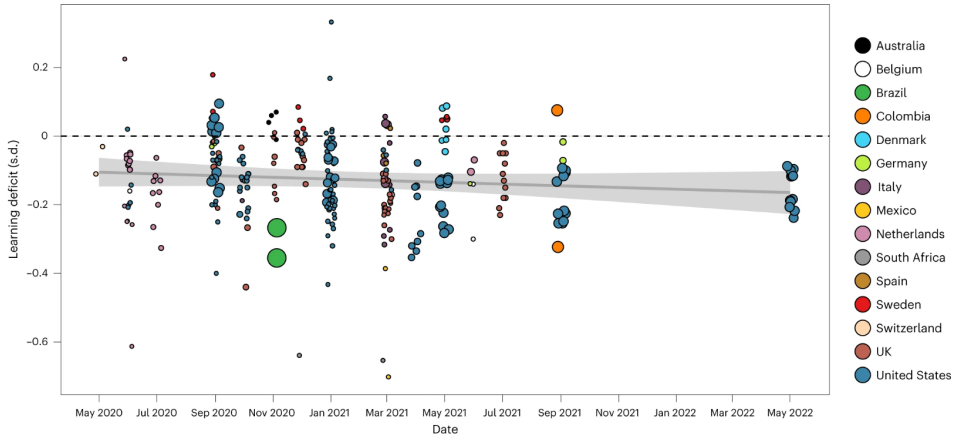
Figure 2: Forest plot



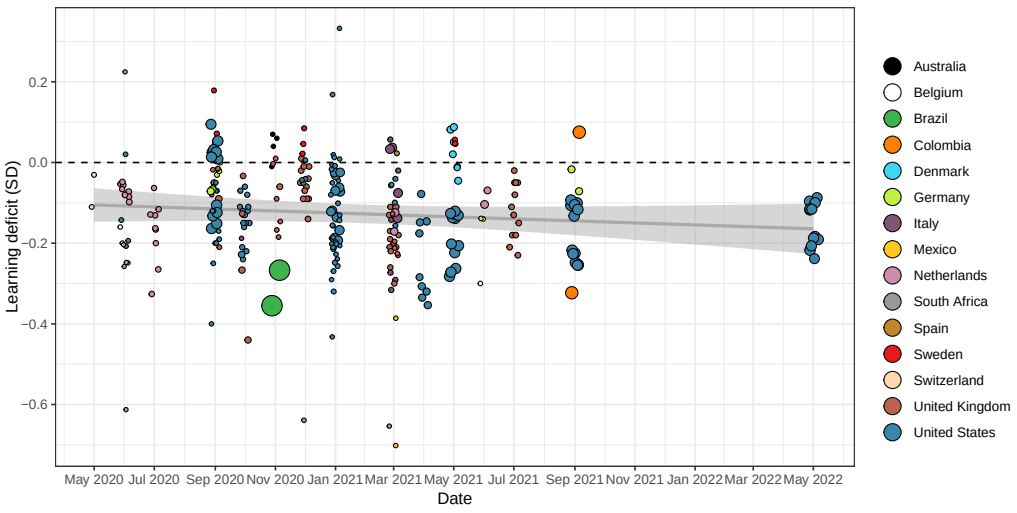
(a) Original Figure 3 (b) Reproduction of Figure 3

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 reproduction, and (b) is the original. 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 3: Estimates of COVID-19 learning deficits in time



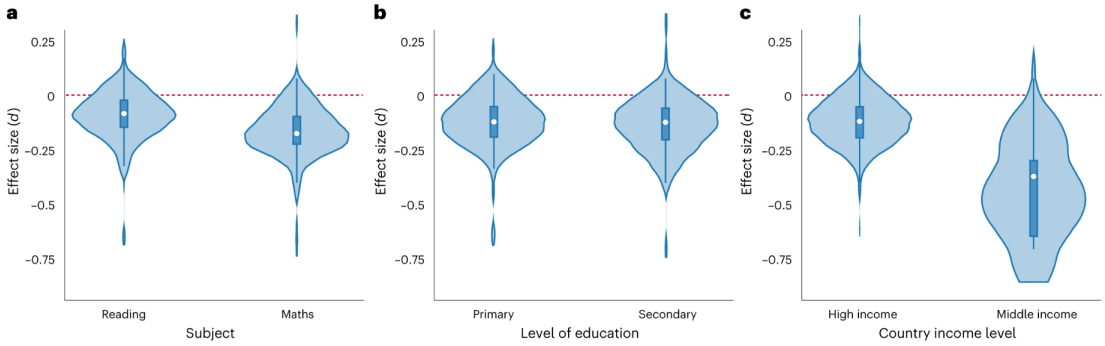
(a) Original Figure 4



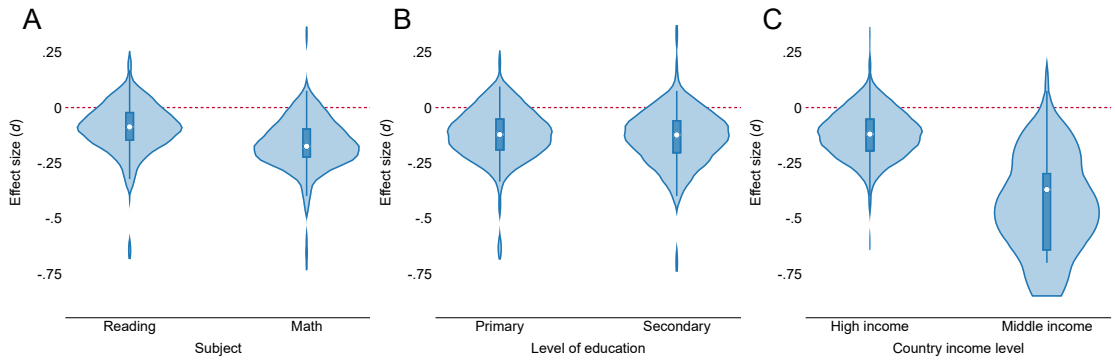
(b) Reproduction of Figure 4

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 reproduction, and (b) is the original. See Table 2 for details.

Figure 4: Variation in estimates of COVID-19 learning deficits



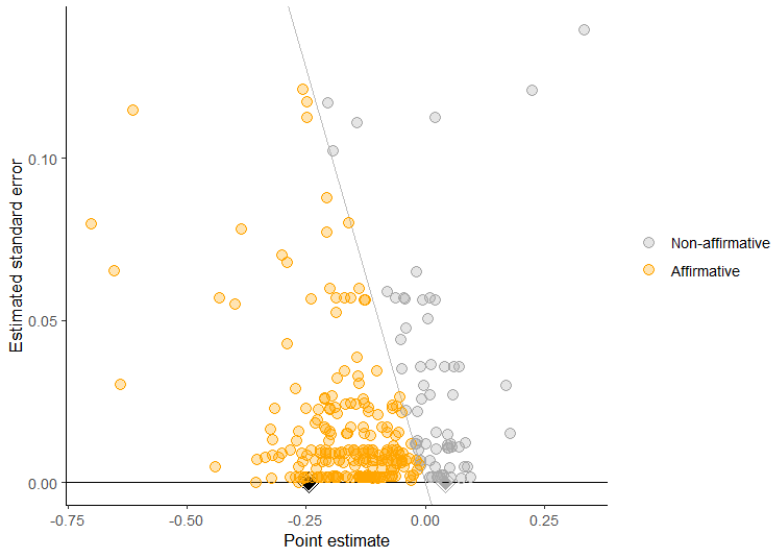
(a) Original of Figure 6



(b) Reproduction Figure 6

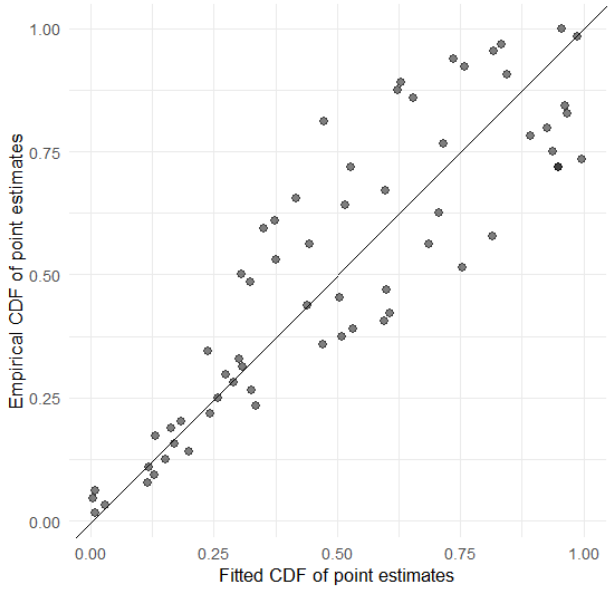
Notes: The figure displays variation in estimates of COVID-19 learning deficit for school subjects (mathematics and reading), level of education, and socio-economic inequality. (a) shows the reproduction, and (b) is the original. No differences between the two. See Table 3 for details.

Figure 5: Significance funnel plot



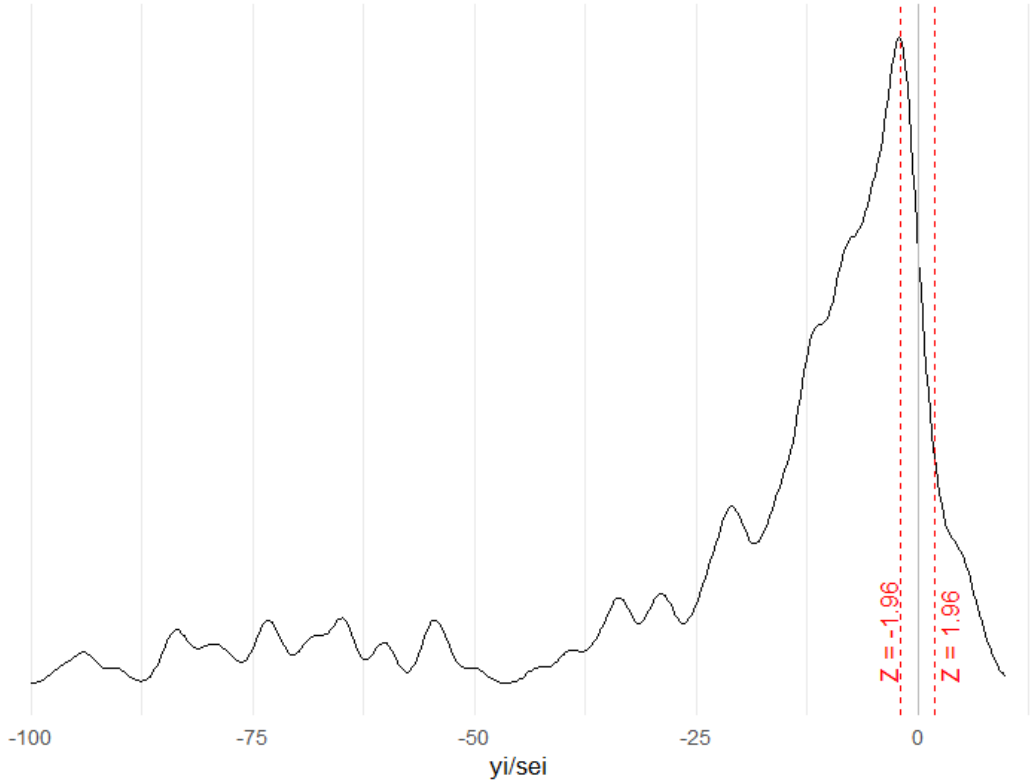
Notes: The figure displays a significance funnel plot. A gray diamond is the worst-case estimate. A black diamond is a pooled estimate for all the studies. See Table 5 for numerical values.

Figure 6: Diagnostic *q*-*q* plot: fitted CDF vs. empirical CDF of point estimates



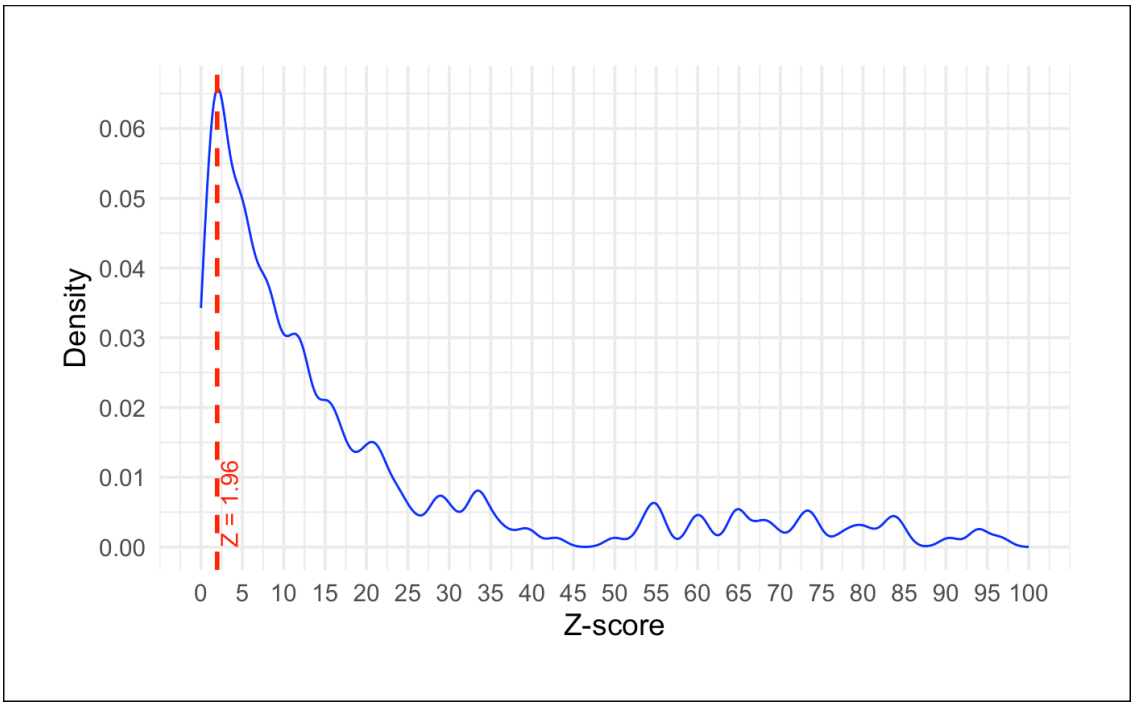
Notes: The figure displays a diagnostic *q*-*q* plot. We plot the fitted CDF of point estimates vs. the empirical CDF of point estimates.

Figure 7: Diagnostic plots: distribution of z-scores



Note: The figure displays the distribution of the z-scores in original data set.

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

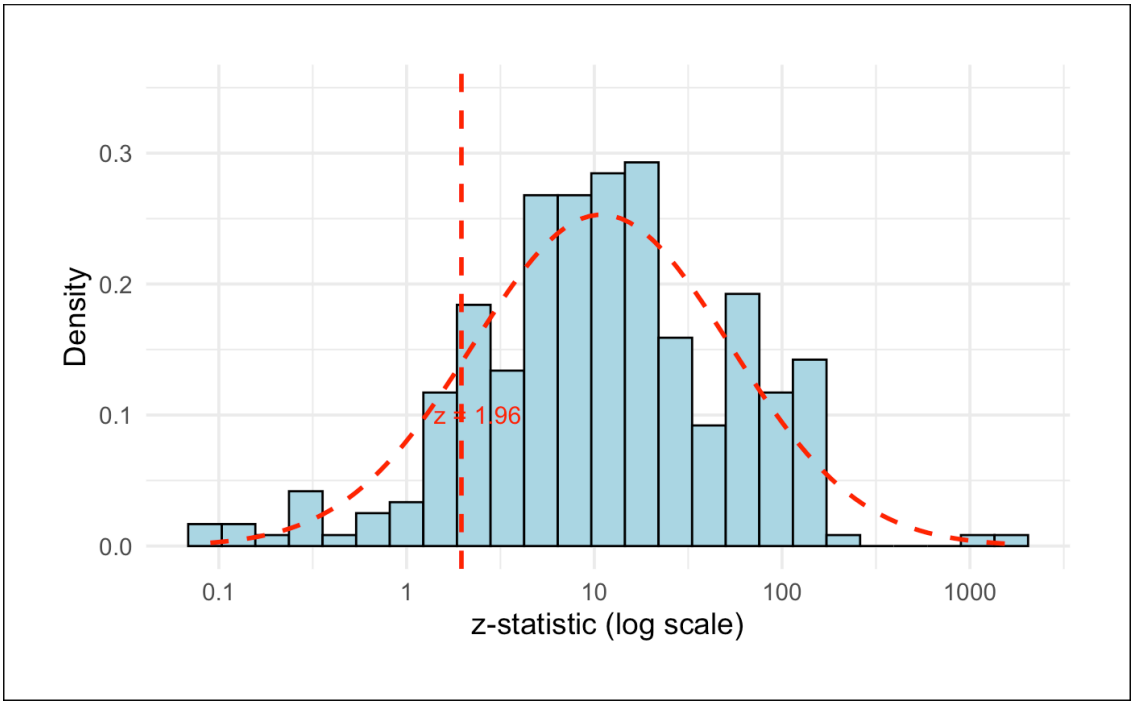


(a) $abs(y_i/se_i)$

Notes: The figure displays the distribution of the z-scores in absolute values (a) and in logs of absolute values (b).

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

(b) $abs(y_i/se_i)$ zoomed to the (0, 20) neighborhood



(c) $ln(abs(y_i/se_i))$

Notes: The figure displays the distribution of the z-scores in absolute values (a), (b), and in logs of absolute values (c).

6 Tables

Table 1: 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. (2023).

Table 2: 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 3: 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.

Table 4: Correcting for publication bias

Method				
	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: PET = precision effect test based on the estimates of 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 $(SE_{estimate})_{ij}$ respective standard error. PET = precision effect test. 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 = model averaging 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 5: Correcting for publication bias, additional specifications

Panel A: Correcting for publication bias

	FE	**RE**
Mean Effect	−0.126**	−0.140***
	(0.059)	(0.020)
Observations	291	291

Panel B: Correcting for publication bias, *s*-value

	MAN	**FE-SR4**	**RE-SR4**
Effect beyond bias	0.021*	−0.206***	−0.068***
	(0.012)	(0.000)	(0.015)
Observations	291	291	291

Panel C: 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 D: *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)
Publication Bias	−0.245***			
	(0.051)			
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, 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. (2023)