CENTER FOR GLOBAL DEVELOPMENT

The Returns to Education: A Comment

Lee Crawfurd

Abstract

Clark and Nielsen (2026) assemble a large dataset of quasi-experimental studies and argue that the causal return to an extra year of schooling is close to zero once one corrects for the selective publication of statistically significant findings. If correct, this would overturn a central pillar of human capital theory and weaken the empirical case for investing in education. This comment re-examines their own data with the standard battery of publication-bias corrections, including methods that guard against spuriously precise estimates and against p-hacking. Every method points to a return of around 5–6% per additional year of schooling—positive, statistically significant, and far from zero—and the estimate stays well above zero even under the most conservative variants. Clark and Nielsen's near-zero figure comes instead from a bespoke procedure that assumes large numbers of negative-return studies were suppressed; the distribution of published estimates does not support that assumption, and the same procedure understates even a genuine, positive return in simulation. Publication bias in this literature is real but modest: corrected for properly, the evidence is inconsistent with a near-zero causal return.

KEYWORDS
returns to education; meta-analysis; publication bias; PET-PEESE; selection models

This paper is forthcoming in the journal Kyklos.

The Returns to Education: A Comment

Lee Crawfurd
Center for Global Development
lcrawfurd@cgdev.org

Lee Crawfurd. 2026. "The Returns to Education: A Comment." CGD Working Paper **756**. Washington, DC: Center for Global Development. https://www.cgdev.org/publication/returns-education-comment

I thank Gregory Clark and Christian Nielsen for making their replication data publicly available and for helpful correspondence on an earlier draft, David K Evans and Harry Patrinos for helpful comments, and Claude (Anthropic) for excellent research assistance. Claude assisted with the replication code, figures, tables, and drafting, under my direction; the arguments and conclusions are my own, all results are reproducible from the public replication repository, and I verified every estimate against the underlying data. Any errors are mine.

CENTER FOR GLOBAL DEVELOPMENT

2055 L Street, NW Fifth Floor
Washington, DC 20036
202.416.4000

1 Abbey Gardens
Great College Street
London
SW1P 3SE

www.cgdev.org

1 Introduction

Clark and Nielsen (2026) (henceforth CN) make a striking claim. Pooling 79 quasi-experimental estimates of the causal effect of an additional year of schooling on earnings—mostly from studies exploiting changes in compulsory schooling laws—they report an unweighted mean return of 8.2%, broadly consistent with surveys of the wider returns literature (Psacharopoulos and Patrinos, 2018). They argue, however, that this figure is severely inflated by publication bias, and that the true causal return is 0–3%. If correct, the conclusion would overturn a central tenet of human capital theory and undermine part of the empirical rationale for mass schooling.

CN's contribution is welcome in one respect: publication bias is a real and well-documented problem in economics (Ashenfelter et al., 1999), and meta-analyses that take it seriously are valuable. Their central conclusion, however, does not survive scrutiny. This comment applies the full battery of publication-bias corrections recommended by the recent practitioner's guide of Irsova et al. (2024) to CN's own replication data. None of the standard methods support a near-zero return. Bias-corrected estimates, including some reported by CN, cluster around 5–6%. These estimates are somewhat below the approximately 10% average found in broader reviews of IV-based returns (Patrinos and Psacharopoulos, 2026), but that gap reflects CN's exclusive focus on compulsory schooling law instruments, which identify a local average treatment effect for students at the margin of dropping out—a group that may have below-average returns. The relevant comparison is not this comment's roughly 5% against 10%: it is that figure against CN's claimed 0–3%.

The remainder of the comment proceeds as follows. Section 2 describes the standard publication-bias toolkit and my approach. Section 3 summarises CN's approach and where it departs from this standard. Section 4 applies the standard toolkit to CN's data. Section 5 examines the normal-distribution argument behind the near-zero end of their range. Section 6 concludes.

2 The Standard Toolkit and Approach of this Comment

Irsova et al. (2024), in the most recent practitioner's guide for meta-analysis in economics, recommend that researchers "always correct for publication bias" using "at least one technique from each of the following two model groups: 1) selection models (Andrews and Kasy, p-uniform*), and 2) funnel-based models (PET-PEESE, endogenous kink, WAAP, MAIVE)." The two model groups embody fundamentally different assumptions about

the mechanism generating selective publication, and a credible meta-analysis should triangulate across both.

This comment reports the uncorrected fixed-effect weighted mean as a baseline and applies the full menu to CN's data: five funnel-based methods—PET-PEESE (equivalent to the Egger slope test), trim-and-fill, the endogenous kink, the weighted average of the adequately powered (WAAP), and MAIVE, which instruments reported precision with sample size to guard against spurious precision—and four selection or significance-based models (p-uniform*, Copas, the Andrews–Kasy step-function model, and multiple-bias meta-analysis), together with right-truncated meta-analysis as a worst-case bound under *p*-hacking. Below, I provide more detail on each method before its implementation. Replication code is available at https://github.com/lcrawfurd/returns-to-education-comment.

3 Clark and Nielsen's Approach

CN assemble 79 independent estimates of the causal effect of an additional year of education on earnings, all drawn from studies exploiting changes in compulsory schooling laws as the identifying instrument. The unweighted mean across these 79 estimates is 8.2%. This is a specific and narrow corner of the quasi-experimental literature: a recent review of all IV-based estimates of returns to schooling—including twin studies, distance-to-school instruments, and other natural experiments—finds an average causal return of approximately 10% across 191 estimates from 145 studies (Patrinos and Psacharopoulos, 2026), if anything higher than CN's mean. Compulsory schooling law instruments identify a local average treatment effect for students on the margin of dropping out, who may have lower returns than average. CN's case for a near-zero causal return rests on two pieces of evidence.

First, they regress effect sizes on standard errors and observe that less precise studies report larger effects. Extrapolating the linear fit to a standard error of zero implies a return of approximately 3% (CN's Figure 7). This is a graphical version of the Precision-Effect Test (PET), but CN's regression is unweighted (OLS) and is not reported with standard errors. The standard inverse-variance-weighted PET applied to the same 65 estimates yields 6.5%, roughly double CN's figure. The gap is thus entirely attributable to the weighting choice—OLS treats a noisy estimate from a single-country study with the same weight as a precise estimate from a large dataset—not to any fundamental methodological difference. Section 4 shows that once standard weighting is applied, the PET estimate is stable across sample restrictions.

Second, and more centrally, they fit a normal distribution to the observed effect sizes by minimising squared errors across frequency bins (their Figure 9). The best-fitting normal is wide (standard deviation about 12%) with its centre close to zero. They read this as evidence that the observed distribution is the right half of a symmetric distribution centred near zero, with the left half "missing" because of selective publication.

Against this standard, CN report only one method from the toolkit—PET-PEESE. They do not report results from trim-and-fill, p-uniform*, the Andrews–Kasy estimator, the Copas selection model, or any other recommended approach.

4 Results

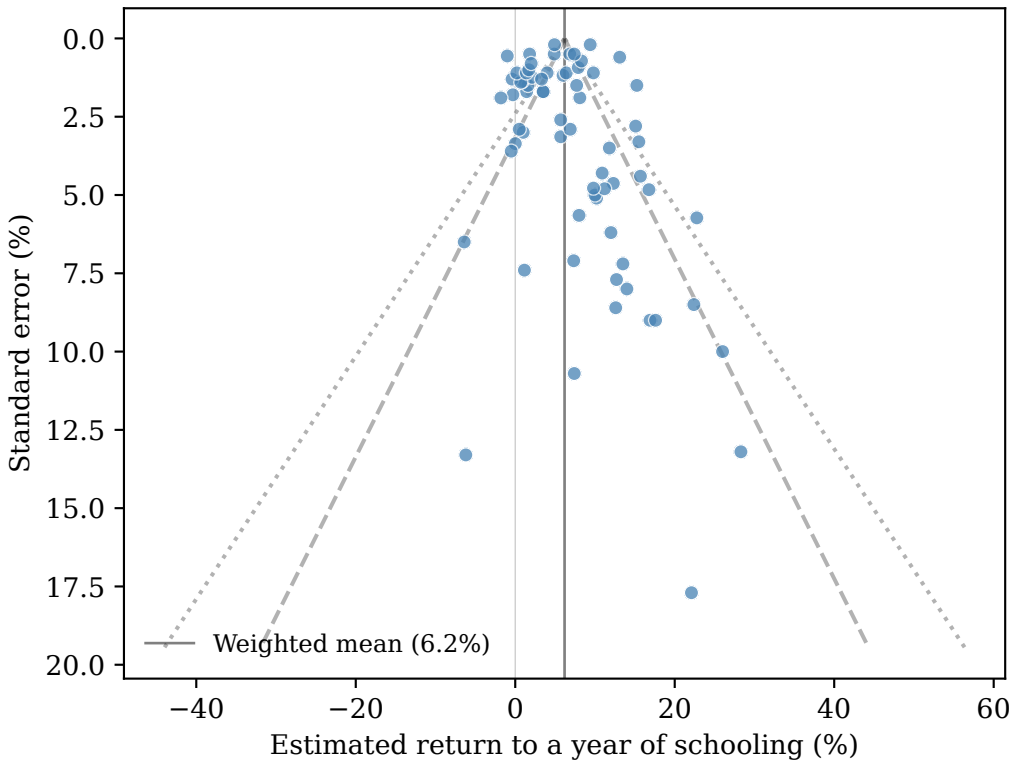
4.1 Funnel-based models

Of CN's 79 independent estimates, 71 report a standard error; the remaining 8 cannot enter any bias-correction method and are dropped throughout. These 71 are the main analysis sample, with a naive unweighted mean of 8.5%—marginally above the 8.2% across all 79, because the excluded estimates happen to report lower returns—and it is this baseline that the corrections reduce. Individual methods then use the subset their assumptions require: 65 estimates fall within CN's own precision restriction (SE $< 10.1\%$) for their PET-PEESE specification; 60 also report a sample size and so enter MAIVE; 33 are adequately powered (SE $\leq 2.21\%$) for WAAP; and the 71 collapse to 48 papers under inverse-variance pooling for the selection models and paper-level clustering. Table 1 states the sample for each method. Outliers are never dropped from estimation: the only exclusions are presentational—two estimates with standard errors above 20% are kept out of the funnel plot (Figure 1), and one return above 40% out of the histogram (Figure 5), so the remaining mass is legible.

Figure 1 shows the funnel plot of effect sizes against standard errors. The funnel is broadly symmetric around an inverse-variance-weighted mean of 6.2%. Estimates near zero or below appear among both the more and the less precise studies, contrary to what one would expect under severe selection against negative findings.

The Egger regression test (Egger et al., 1997) regresses t-statistics on precision (1/SE) and tests whether the intercept differs from zero. The intercept is -0.38 ($p = 0.595$): this implies no statistically significant funnel asymmetry. The slope of this regression is algebraically identical to the PET intercept (Stanley and Doucouliagos, 2014); both equal 6.4% ($p < 0.001$). Table 1 reports the Egger intercept alongside all other methods.

Figure 1: Funnel plot of returns-to-education estimates



Note: Each point is one of the 71 estimates from CN's dataset that report standard errors (8 of their 79 lack standard errors and are excluded from all analyses). Dashed lines indicate the region of statistical insignificance at the 5% level. Two outliers with standard errors exceeding 20% are excluded for presentational clarity. The vertical line marks the inverse-variance weighted mean of 6.2%.

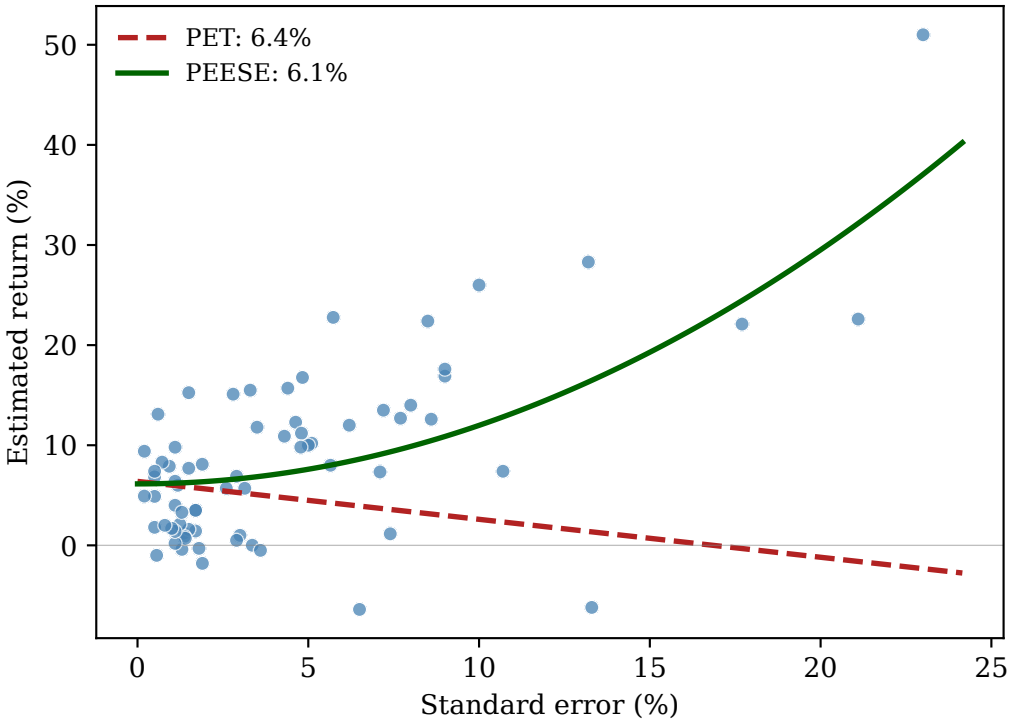
PET-PEESE

PET-PEESE (Stanley and Doucouliagos, 2014)—Precision Effect Test (PET) and Precision Effect Estimate with Standard Error (PEESE)—regresses effect sizes on standard errors (PET) or on squared standard errors (PEESE), with inverse-variance weights. The intercept estimates the effect size at zero standard error—that is, the bias-corrected true effect. The standard decision rule prescribes PEESE when the PET intercept is significant.

Applied to the full sample of 71 independent estimates with non-missing standard errors, PET yields an intercept of 6.4% ($p < 0.001$) and PEESE yields 6.1% ($p < 0.001$) (Figure 2). Both are precisely estimated and far from zero. Running CN's own PET-PEESE specification (their replication do-file, restricted to SE $< 10.1\%$) reproduces this almost exactly: PET $= 6.5\%$, PEESE $= 6.1\%$, the estimates CN themselves report as a bias-corrected true effect of 6.1–6.4%. Because each of the 48 papers in the sample may contribute multiple estimates, the inverse-variance specification may overweight studies with many estimates. Applying robust variance estimation (RVE; Hedges et al., 2010), which clusters standard errors at the paper level and applies a small-sample CR2 correction ($n = 48$ clusters), yields PET $= 3.9\%$ ($p < 0.001$) and PEESE $= 5.4\%$ ($p < 0.001$)—lower than the FE estimates but still far from zero. Table 1 reports all specifications across weighting schemes.

CN report these results but discount them, noting that PET-PEESE has been criticised for imprecise, model-dependent corrections and for false negatives when studies are few or heterogeneity is high (Pustejovsky, 2017; Doucouliagos et al., 2018; Bartoš et al., 2022; Kvarven et al., 2020). Two of these conditions should be separated. The number of estimates here is not small—71 estimates from 48 papers—so that part of the concern does not apply; heterogeneity, by contrast, genuinely is high ($I^2 = 95.6\%$), a setting in which PET-PEESE can perform poorly. But the documented failure, and the one CN themselves name, is a bias toward *false negatives*: under these conditions PET-PEESE tends to miss true effects and understate them (Stanley, 2017). A method that errs toward finding nothing cannot conjure a 6% return from a true zero—if it is biased here, it is biased downward, which makes CN's own 6% a floor, not a ceiling. Nothing hinges on PET-PEESE in any case: the selection models in the next section start from entirely different assumptions and all return positive, significant estimates (roughly 5–6% at the selection intensity the data imply), and the random-effects and cluster-robust variants appropriate under high heterogeneity (Section 4.4) leave the reported (PEESE) bias-corrected return at 5.0–5.4%. What CN's own regressions show to be model-dependent is only whether the publication-selection slope reaches significance; their bias-corrected intercept stays near 6% throughout.

Figure 2: PET-PEESE publication-bias correction



Note: PET (dashed red) regresses effect sizes on standard errors, weighted by inverse variance; PEESE (solid green) uses squared standard errors. Both intercepts—the bias-corrected estimate of the true return at SE $= 0$—are approximately 6% and highly significant ($p < 0.001$). The PET slope is negative, indicating that noisier studies report *smaller*, not larger, effects.

Trim-and-fill

Trim-and-fill (Duval and Tweedie, 2000) iteratively imputes "missing" studies on the under-represented side of the funnel and recomputes the adjusted mean. Applied to CN's data, the algorithm finds no studies to impute and returns an adjusted mean of 6.2%.

Endogenous kink

The endogenous-kink estimator (Bom and Rachinger, 2019) refines PET-PEESE by letting the effect–standard-error relationship be flat among the most precise estimates and linear only beyond a threshold. Selection cannot inflate estimates that are significant on their own—those precise enough that the standard error falls below $|\hat{\mu}|/1.96$—so the bias-generating slope applies only above that kink, whose location is estimated jointly with the corrected intercept. On the 71 independent estimates (inverse-variance weighted) the kink falls at a standard error of 3.1%, leaving 33 estimates in the biased region above it; the bias-corrected intercept is 6.2% ($p < 0.001$).

WAAP

The Weighted Average of Adequately Powered studies (WAAP; Stanley et al. 2017) restricts to estimates from studies with sufficient precision to detect the pooled effect at 80% power. Of the 71 estimates, 33 meet the adequacy threshold (SE $\leq$ 2.21%); their inverse-variance-weighted mean is 6.1% ($p < 0.001$).

MAIVE: spurious precision

The funnel-based methods above, and the inverse-variance weights behind them, take the reported standard error as an honest measure of precision. Under extreme heterogeneity that is worth probing: the weights load on a handful of very precise estimates, and if their precision were spurious—a product of specification or clustering choices rather than genuine information—the correction could be distorted. MAIVE (Irsova et al., 2025) guards against this by instrumenting the reported standard error with sample size, on the logic that a genuine standard error scales with $1/\sqrt{N}$.

Sixty of the 71 estimates report a sample size; on these (41 papers) the first-stage F is 28.4, a strong instrument. MAIVE's recommended default—unweighted PET-PEESE with the instrument—returns 2.6%. That default is unweighted, however, and unweighting is exactly the choice that separates CN's $\approx$ 3% from the inverse-variance-weighted $\approx$

6%. Holding the weighting fixed and switching the instrument on isolates the spurious-precision correction itself: the inverse-variance-weighted PET intercept then moves from 6.3% to 6.9%—up, not down. The fall to 2.6% comes from dropping the weights, not from instrumenting precision. The very precise large-N studies have small standard errors because their samples are genuinely large, and MAIVE confirms it.

4.2 Selection models

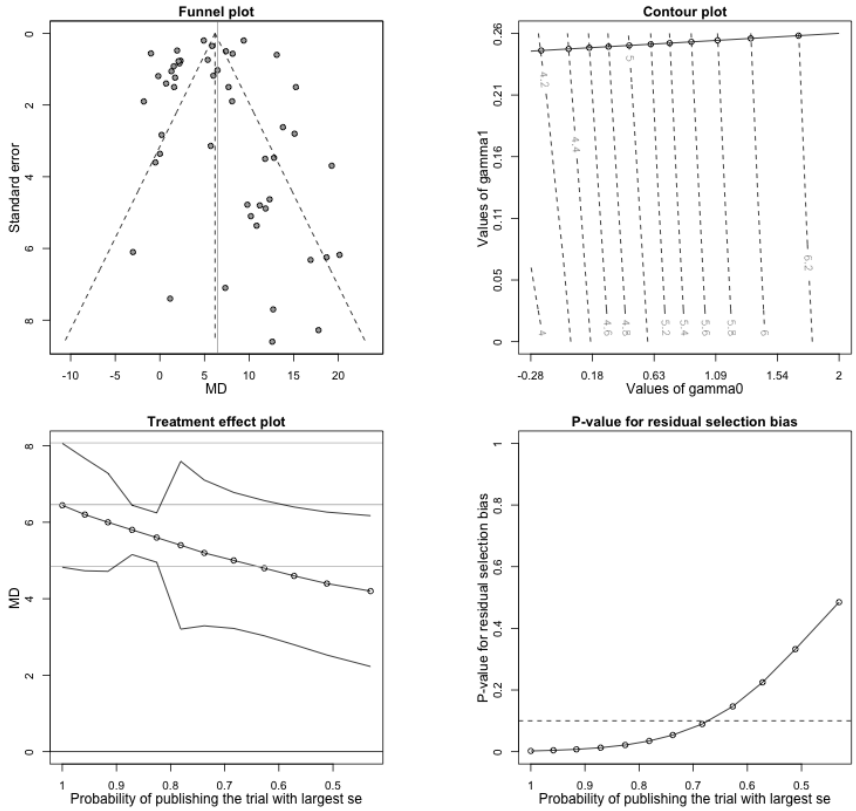
The funnel-based methods above all assume that selective publication operates on study precision. Selection models instead assume selection on statistical significance—that studies with insignificant results are less likely to be published regardless of their precision. Four implementations were applied to study-collapsed estimates ($n = 48$): p-uniform* (van Aert and van Assen, 2026), the Copas selection model (Copas, 2001), the Andrews–Kasy / Vevea–Hedges step-function model (Andrews and Kasy, 2019), and—to bound the joint effect of publication selection and within-study p-hacking—multiple-bias meta-analysis (Mathur, 2024b).

p-uniform* yields a bias-corrected estimate of 6.2% ($p < 0.001$; confidence interval unavailable due to non-convergence of test inversion with extreme heterogeneity). The Copas model yields 4.8% (95% CI: 3.0% to 6.6%, $p < 0.001$). Figure 3 shows the Copas sensitivity plot. The Andrews–Kasy step-function model, which estimates the probability that a non-significant result is published relative to a significant one, yields a bias-corrected estimate of 5.3% (95% CI: 2.9% to 7.6%, $p < 0.001$); the estimated relative publication probability for non-significant results is $\hat{\omega} = 0.50$, indicating that non-significant findings are about half as likely to be published as significant ones. Multiple-bias meta-analysis (Mathur, 2024b), which models publication selection and within-study bias jointly, gives 5.2% (95% CI: 3.5% to 6.8%) at the selection intensity Andrews–Kasy implies (affirmative results about twice as likely to be published). The estimate declines as the assumed selection strengthens but stays positive and significant throughout: 3.9% if affirmative results are four times as likely to appear, and 1.4% even under the extreme assumption of a two-hundred-fold advantage, always with a confidence interval excluding zero. These selection models, like the funnel-based methods, deliver estimates well above zero.

As a deliberately pessimistic check, right-truncated meta-analysis (Mathur, 2024a) estimates the corrected mean under *worst-case* p-hacking—the assumption that every statistically significant, correctly-signed estimate owes its significance to manipulation. Even then the corrected mean is positive, around 2–4%, with a credible interval that excludes zero. This worst case assumes exactly the within-study p-hacking that the caliper test in

the next section finds no trace of; absent that manipulation, the relevant estimates are the higher ones above.

Figure 3: Copas selection-model sensitivity plot



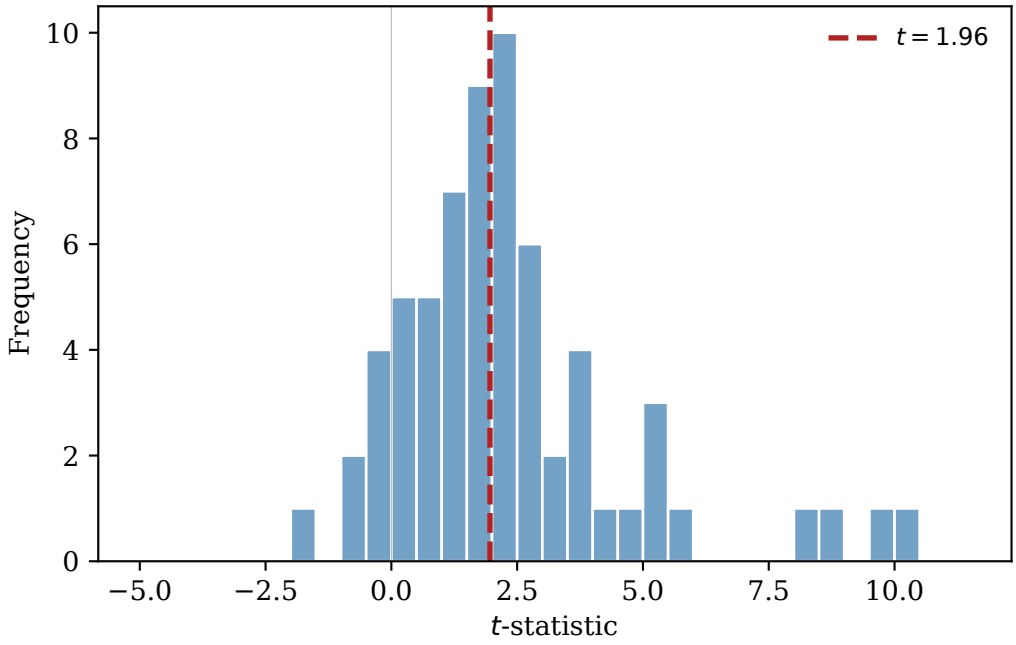
Note: The plot shows how the adjusted effect-size estimate varies as the assumed probability of selection changes. Even under strong selection assumptions, the Copas-adjusted estimate remains well above zero, settling at 4.8% (95% CI: 3.0–6.6%).

4.3 No bunching of *t*-statistics

A caliper test examines whether *t*-statistics bunch just above 1.96, a signature of within-study *p*-hacking—where researchers manipulate specifications to push a result across the significance threshold. This is distinct from the file-drawer mechanism CN allege, in which entire studies with insignificant results go unpublished. The two selection mechanisms leave different distributional traces: *p*-hacking produces bunching at the threshold, while file-drawer suppression produces a deficit of small *t*-statistics throughout the sub-significant region. The caliper test therefore diagnoses a different form of bias from the one under discussion; the Andrews–Kasy model in the previous section directly addresses the file-drawer mechanism.

Figure 4 shows the t-statistic distribution. With a bandwidth of 0.5, there are 10 estimates on each side of 1.96 (binomial $p = 1.0$); with a bandwidth of 1.0, 16 on each side ($p = 1.0$). There is no evidence of p-hacking.

Figure 4: Distribution of t-statistics



Note: Histogram of t-statistics from the 71 of CN's 79 estimates that report standard errors. The dashed vertical line marks the 5% significance threshold ($t = 1.96$). There is no visible bunching of estimates just above the threshold, and a caliper test confirms equal counts on both sides ($p = 1.0$ for bandwidths of 0.5 and 1.0).

4.4 Heterogeneity

Between-study heterogeneity in this literature is extreme. A random-effects meta-analysis using Restricted Maximum Likelihood (REML) estimation of the 71 independent estimates yields $I^2 = 95.6\%$ with $\tau^2 = 20.0$ and Cochran's $Q(70) = 1072.8$ ($p < 0.001$), all well above the 75% I^2 threshold conventionally regarded as substantial. Here I^2 is the share of total variation attributable to between-study differences rather than sampling error, τ^2 is the estimated between-study variance, and Q is the standard homogeneity statistic. This heterogeneity is not a nuisance to be explained away: returns to schooling genuinely vary across countries, cohorts, and the specific reforms being studied, and LATE estimates from different compliers should not be expected to coincide.

Under severe heterogeneity, the behaviour of PET-PEESE has been debated. The esti-

mator assumes that the bias-generating selection mechanism operates identically across studies, and inverse-variance weights place large weight on a handful of very precise estimates that may not be representative of the broader evidence base. Two adjustments address this concern. First, random-effects (REML) weighting gives more balanced weight to less-precise estimates and yields a PEESE intercept of 5.0%. Second, cluster-robust variance estimation with paper-level clustering (RVE; Hedges et al. 2010; Pustejovsky and Tipton 2022), which does not assume independent observations within the same paper, yields a PEESE intercept of 5.4%. These RE and RVE specifications are the appropriate ones under extreme heterogeneity. Under the standard PET–PEESE decision rule (Stanley and Doucouliagos, 2014), the reported bias-corrected estimate is the PEESE intercept whenever the PET test rejects a zero effect—as it does in every specification here ($p < 0.001$). The reported funnel-based estimate is therefore 5.0–6.1% across FE, RE, and RVE weighting; the PET intercepts (6.4%, 3.3%, and 3.9%, reported in Table 1) are the precision-effect significance test, not separate magnitude estimates. Even the lowest PEESE estimate, 5.0%, is well above the near-zero return CN's normal-fitting argument points to.

4.5 Summary

Table 1 summarises results across all methods. Every correction—funnel-based, selection-based, and robust to spurious precision or *p*-hacking—yields a bias-corrected return that is positive, statistically significant, and economically meaningful. Under standard inverse-variance weighting the funnel-based estimates cluster near 6% (PEESE 6.1%, WAAP 6.1%, trim-and-fill 6.2%, endogenous kink 6.2%, with the PET significance test at 6.4%); the selection models, at the selection intensity the data imply, give 4.8–6.2% (Copas 4.8%, Andrews–Kasy 5.3%, multiple-bias 5.2%, p-uniform* 6.2%). Under the specifications built for extreme heterogeneity the reported PEESE estimate is 5.0–5.4% (RE 5.0%, RVE 5.4%); the corresponding PET intercepts (3.3% and 3.9%) are the significance pretest, not the reported magnitude. Figures reach 1–4% only under worst-case *p*-hacking assumptions that the caliper test rejects. Every one of these figures is well above the near-zero return CN infer from their normal-fitting argument, and—outside the worst-case *p*-hacking bound—each is statistically distinguishable from zero. MAIVE confirms that the $\approx 6\%$ under standard weighting is not an artefact of spurious precision.

Table 1: Publication-bias corrections applied to Clark and Nielsen's (2026) replication data

Method	Sample	Estimate [95% CI]	p-value
Panel A: Funnel-based methods (fixed-effect weighting)			
Fixed-effect weighted mean	All ind. est. ($n = 71$)	6.2%	—
WAAP	Adequately powered ($n = 33$)	6.1%	< 0.001
Egger intercept (bias)†	All ind. est. ($n = 71$)	-0.38	0.595
PET (= Egger slope)‡	All ind. est. ($n = 71$)	6.4%	< 0.001
PEESE	All ind. est. ($n = 71$)	6.1%	< 0.001
Trim-and-fill	All ind. est. ($n = 71$)	6.2%	—
Endogenous kink	All ind. est. ($n = 71$)	6.2%	< 0.001
Panel A′: Funnel-based methods, heterogeneity-robust weighting			
Random-effects weighted mean	All ind. est. ($n = 71$)	6.0%	< 0.001
RE PEESE intercept (REML)	All ind. est. ($n = 71$)	5.0%	< 0.001
RVE PEESE intercept	All ind. est. ($n = 71$, 48 clusters)	5.4%	< 0.001
Heterogeneity: $I^2 = 95.6\%$, $\tau^2 = 20.0$, $Q(70) = 1072.8$ ($p < 0.001$)			
Panel A″: Spurious-precision-robust (MAIVE; N instruments precision)			
MAIVE PET-PEESE (default / IV-weighted)	Est. with N ($n = 60$)	2.6% / 6.9%	—
Panel B: Selection / significance-based models			
p-uniform*	Collapsed by study ($n = 48$)	6.2%	< 0.001
Copas	Collapsed by study ($n = 48$)	4.8% [3.0%, 6.6%]	< 0.001
Andrews–Kasy (Vevea–Hedges)	Collapsed by study ($n = 48$)	5.3% [2.9%, 7.6%]	< 0.001
Multiple-bias ($\eta \approx 2$)	Collapsed by study ($n = 48$)	5.2% [3.5%, 6.8%]	< 0.001
Panel C: Clark and Nielsen's own estimates			
PET (SE $< 10.1\%$)	$n = 65$	6.5%	< 0.001
PEESE (SE $< 10.1\%$)	$n = 65$	6.1%	< 0.001

Notes: Bias-corrected returns to one additional year of schooling (percentage points), applied to Clark and Nielsen's (2026) data. Each method and its citation is described in Section 4; column 2 gives the sample for each. No estimate is dropped as an outlier from any row; the only outlier exclusions are presentational (Figures 1 and 5). For MAIVE the default is unweighted and the IV-weighted variant holds the weighting fixed, isolating the spurious-precision correction. † Egger intercept is a dimensionless t-statistic, not a return. ‡ PET is the precision-effect significance test; significant throughout ($p < 0.001$), so under the PET–PEESE decision rule the reported estimate is the PEESE intercept. '—' denotes no analytic p-value.

5 The Normal-Distribution Argument

The push toward the bottom of CN's 0–3% range—the claim that the true return lies near zero rather than near the $\approx 3\%$ their slope regression implies—comes neither from PET-PEESE nor from any selection model. It rests on the argument behind their Figure 9: a normal distribution is fitted to frequency bins of the positive effect-size estimates, on the premise that the true distribution of effects is normal and that the observed estimates are only the right half of a symmetric distribution centred near zero, the left half being "missing" due to selective non-publication of negative results. It is this assumption, not any particular fitted value, that pushes the estimate toward zero.

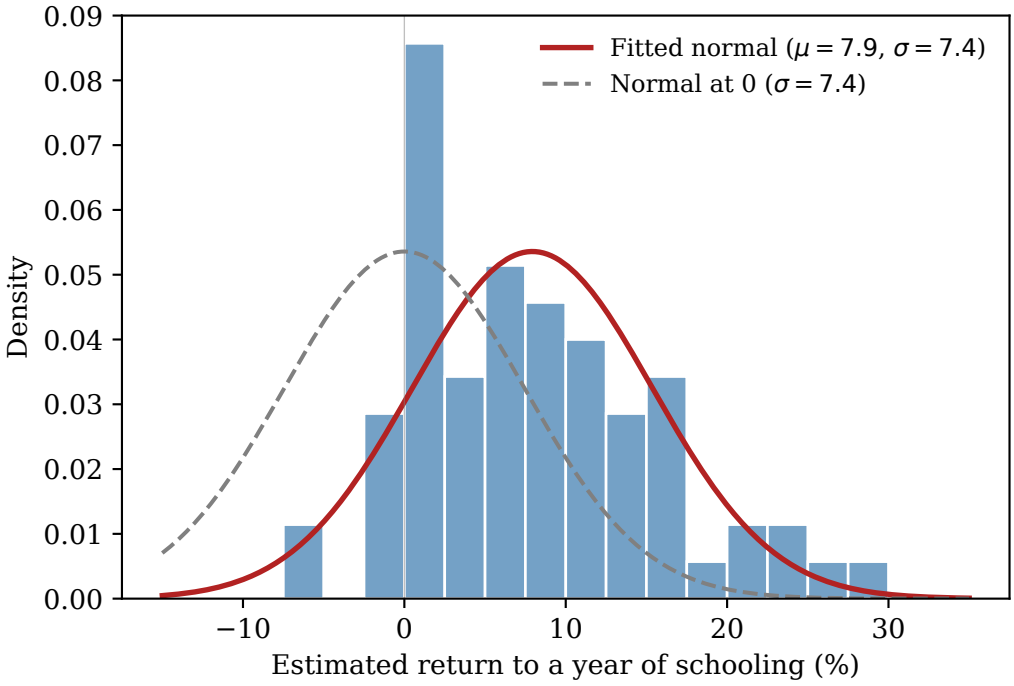
This procedure has two decisive problems.

The first problem is that the truncation assumption, not the data, produces the near-zero mean. CN fit the positive estimates with a normal constrained to be zero for all negative values, premising that every negative effect was suppressed by publication. That constraint does the work. Figure 5 makes the point: the unconstrained maximum-likelihood normal fit to the same data—which does not assume negative observations are "missing"—has its mean at approximately 8%, not at zero.

The left panel of Figure 6 shows the same contrast directly: CN's truncated-normal fit (red), which assumes the negatives are censored, centres near zero, while the unconstrained fit (green) recovers $\hat{\mu} = 8.0\%$. The near-zero result is an artefact of the censoring assumption, not a finding derived from the data.

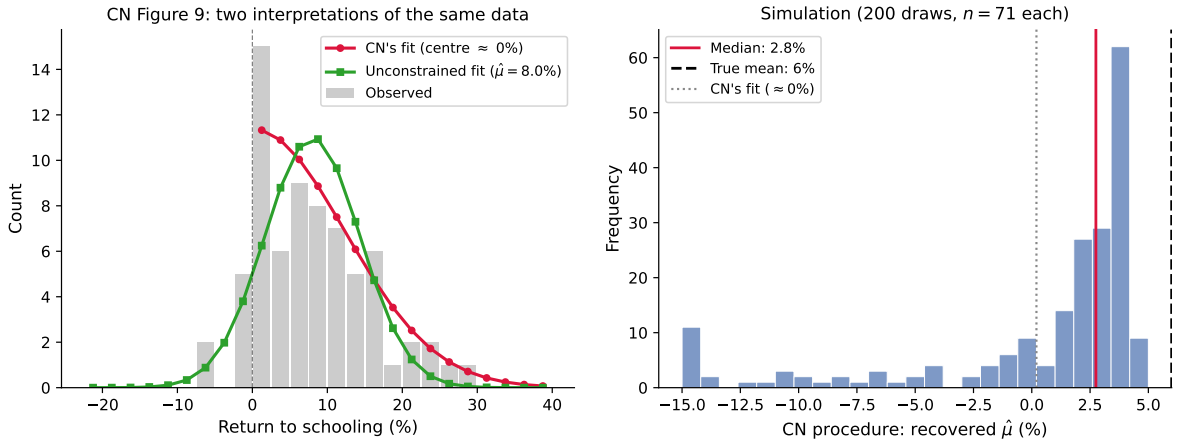
The second problem is that the procedure understates even a known, positive mean. The check is simple: feed it data whose true answer is already known and see whether it recovers it. The right panel of Figure 6 applies CN's exact grid-search procedure to 200 simulated datasets of $n = 71$ draws from a log-normal distribution with a known true mean of 6%, a standard deviation of 6%, and no publication bias at all—nothing dropped, nothing inflated. The median recovered value is 2.75%; half the draws fall below 3%, and none reaches the true 6%. The exercise tests CN's procedure rather than estimating the return, which the standard methods above already put at around 6%; and the procedure fails the test, returning a number near zero from data with a real, positive effect and nothing to correct. A near-zero output is therefore a property of the procedure, not evidence that the true return is near zero.

Figure 5: Distribution of estimated returns with fitted normal densities



Note: Histogram of the 71 of CN's 79 estimates that report standard errors (excluding, for visual clarity only, one estimate with a return above 40%; it is retained in all estimation). The solid red curve is the unconstrained maximum-likelihood normal fit ($\hat{\mu} = 8.0$, $\hat{\sigma} = 6.3$); the dashed grey curve centres a normal at zero with the same standard deviation, as CN's procedure assumes. The unconstrained fit centres near 8%, not zero: it is the truncation-at-zero assumption, not the data, that produces CN's near-zero mean.

Figure 6: CN's procedure is biased toward zero by construction



Note: Left panel: CN's Figure 9 exercise on the actual data. The red curve shows the truncated-normal fit that assumes all negative effects are censored; its centre sits near zero. The green curve shows the unconstrained MLE normal fit ($\hat{\mu} = 8.0\%$, $\hat{\sigma} = 6.3\%$). The censoring assumption, not the data, produces the near-zero result. *Right panel:* Distribution of CN's recovered $\hat{\mu}$ across 200 simulated datasets of $n = 71$ draws from a log-normal with true mean = 6%, SD = 6%, and no publication bias. The median recovered value is 2.75% (dashed red), well below the true mean of 6% (dashed black). Around 50% of draws return a value below 3%; no draw recovers the true mean.

6 Conclusion

The standard toolkit for publication-bias correction in meta-analysis was developed precisely to answer questions of the kind CN pose. Applied to their own data, every recommended method—funnel-based, selection-based, and robust to spurious precision or *p*-hacking—returns a bias-corrected estimate that is positive, significant, and economically meaningful: around 5–6% under standard weighting and the selection intensity the data imply, and near 5% even under the heterogeneity-robust (PEESE) specifications. CN's own PET-PEESE regressions, reported in their paper, give the same answer as the standard-weighted methods.

A bias-corrected return of roughly 5% for CN's sample is consistent with the broader causal evidence once instrument differences are accounted for. Reviews drawing on a wider range of quasi-experimental designs—twin studies, distance-to-school instruments, and other natural experiments as well as CSL reforms—find average IV-based returns of approximately 10% (Patrinos and Psacharopoulos, 2026); the gap reflects the fact that compulsory schooling law instruments identify returns for students at the margin of dropping out, who may have below-average gains from an additional year of schooling. Even at 5%, the return is economically meaningful: it implies substantial lifetime earnings gains from policies that extend schooling, and is well above any plausible threshold

at which public investment in education would fail a cost–benefit test.

The naive mean of around 8% is plausibly inflated, and a figure around 5% may be more reasonable as a summary of the causal evidence. But the leap from "some publication bias exists" to a true return near zero is unsupported by any standard method, and rests on a bespoke distribution-fitting procedure whose near-zero result follows from its own assumption that all negative estimates were suppressed, not from the data. A simulation applying CN's exact procedure to 200 right-skewed datasets with a known true mean of 6% and no publication bias recovers a median of 2.75%, with 50% of draws below 3% and none reaching the true mean. The unconstrained MLE fit to CN's data, which drops the truncation-at-zero assumption, recovers 8.0%—far above the near-zero figure their truncated fit produces. Publication bias is a real concern; the appropriate response is to apply the methods designed to correct for it, not to fit ad hoc parametric models that deliver predetermined conclusions.

References

Andrews, I. and Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–2794.

Ashenfelter, O., Harmon, C., and Oosterbeek, H. (1999). A review of estimates of the schooling/earnings relationship, with tests for publication bias. *Labour Economics*, 6(4):453–470.

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):1–19.

Bom, P. R. D. and Rachinger, H. (2019). A kinked meta-regression model for publication bias correction. *Research Synthesis Methods*, 10(4):497–515.

Clark, G. and Nielsen, C. A. A. (2026). The returns to education: A meta-study. *Kyklos*.

Copas, J. (2001). A sensitivity analysis for publication bias in systematic reviews. *Biostatistics*, 2(4):463–471.

Doucouliagos, H., Paldam, M., and Stanley, T. D. (2018). Skating on thin evidence: Implications for public policy. *European Journal of Political Economy*, 54:16–25.

Duval, S. and Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2):455–463.

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

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.

Irsova, Z., Bom, P. R. D., Havranek, T., and Rachinger, H. (2025). Spurious precision in meta-analysis of observational research. *Nature Communications*, 16:7647.

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.

Kvarven, A., Strømland, E., and Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, 4(4):423–434.

Mathur, M. B. (2024a). P-hacking in meta-analyses: A formalization and new meta-analytic methods. *Research Synthesis Methods*, 15(6):1048–1063.

Mathur, M. B. (2024b). Sensitivity analysis for the interactive effects of internal bias and publication bias in meta-analyses. *Research Synthesis Methods*, 15(2):334–348.

Patrinos, H. A. and Psacharopoulos, G. (2026). Causal returns to education. *International Journal of Educational Development*, 122:103565.

Psacharopoulos, G. and Patrinos, H. A. (2018). Returns to investment in education: A decennial review of the global literature. *Education Economics*, 26(5):445–458.

Pustejovsky, J. E. (2017). You wanna PEESE of d's? https://jepusto.com/posts/PET-PEESE-performance/. Working paper, University of Wisconsin–Madison.

Pustejovsky, J. E. and Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, 23:425–438.

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, 8(6):581–591.

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

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

van Aert, R. C. M. and van Assen, M. A. L. M. (2026). Correcting for publication bias in a meta-analysis with the p-uniform* method. *Psychonomic Bulletin & Review*, 33(3).