A Report on "The Returns to Education: A Meta-Study" by Clark and Nielsen (2026)

Reviewer 2

February 13, 2026

v1

isitcredible.com

Disclaimer

This report was generated by large language models, overseen by a human editor. It represents the honest opinion of The Catalogue of Errors Ltd, but its accuracy should be verified by a qualified expert. Comments can be made here. Any errors in the report will be corrected in future revisions.

> I am wiser than this person; for it is likely that neither of us knows anything fine and good, but he thinks he knows something when he does not know it, whereas I, just as I do not know, do not think I know, either. I seem, then, to be wiser than him in this small way, at least: that what I do not know, I do not think I know, either.
>
> Plato, *The Apology of Socrates*, 21d

> To err is human. All human knowledge is fallible and therefore uncertain. It follows that we must distinguish sharply between truth and certainty. That to err is human means not only that we must constantly struggle against error, but also that, even when we have taken the greatest care, we cannot be completely certain that we have not made a mistake.
>
> Karl Popper, 'Knowledge and the Shaping of Reality'

Overview

Citation: Clark, G., and C. A. A. Nielsen. (2026). The Returns to Education: A Meta-Study. *Kyklos*. Vol. 0, No. 1, pp. 1–21.

Abstract Summary: This meta-study surveys 79 estimates of the causal effect of an additional year of education on earnings, finding an average return of 8.2%, but strong evidence of publication bias, which, once corrected, suggests the average causal returns are only in the $0\% - 3\%$ range.

Key Methodology: Meta-study of 79 causal estimates of returns to education, primarily derived from changes in minimum school-leaving age, utilizing publication bias tests (association between estimate size and standard error, p-hacking, and distribution assessment) and PET-PEESE procedure.

Research Question: Does more education lead to higher later-life earnings?

Summary

Is It Credible?

This article by Clark and Nielsen presents a provocative meta-analysis of the economic returns to education, challenging a deeply entrenched consensus in labor economics. By aggregating 79 causal estimates from 53 papers that utilize compulsory schooling laws (CSL) as instruments, the authors argue that the "true" return to an additional year of schooling is not the widely accepted 8–10%, but rather "in the range 0%–3%" (p. 1). Their central thesis is that the literature is distorted by severe publication bias: specifically, a "null-result penalty" that suppresses non-significant findings and a systematic omission of negative estimates. They support this by demonstrating a strong positive association between standard errors and effect sizes—implying that only large effects survive in small samples—and by identifying a suspicious clustering of estimates just above zero.

The credibility of the authors' diagnosis of publication bias is relatively high. The article describes a significant and substantially large positive association between estimate sizes and standard errors, a classic hallmark of publication selection where researchers or editors filter out imprecise, insignificant results (p. 2). Furthermore, the observation that 24% of studies cluster in the 0%–4% range, despite a mean of 8.2% and large sampling errors, provides compelling circumstantial evidence of what the authors term "0-hacking"—specifically manipulating negative estimates to appear positive—or the file-drawer problem (pp. 9–10). The authors reasonably argue that if the true distribution were normal around 8.2%, we should see far more negative estimates than the literature reports.

However, the credibility of the specific "0%–3%" replacement figure is lower than the credibility of their critique of the status quo. The authors derive the upper bound of this range (3%) from the intercept of a regression of effect sizes on standard errors,

while the lower bound (0%) is derived from correcting for the suspicious absence of negative estimates under the assumption of a normal distribution. While the authors do report p-values and ranges for the intercept in their controlled specifications, suggesting the estimate is statistically distinguishable from zero ($p = 0.000$), the precision of the specific *uncontrolled* intercept remains less transparent (p. 7). This is compounded by the authors' methodological choice to "average all causal estimates and their standard errors" when papers report multiple models without a primary specification (p. 6). While this only applies to a subset of the data, simply averaging standard errors is not a statistically valid method for aggregating uncertainty, as it ignores correlations between estimates and sample sizes.

Furthermore, there is a significant disconnect between the article's broad scope and the narrow population studied. The analysis is restricted to CSL reforms, which identify a Local Average Treatment Effect (LATE) for a specific sub-population: students at the margin of dropping out who are forced to stay in school. The authors anticipate this critique, citing arguments that these "low ability" students typically show *higher* returns than the average, implying that if the CSL return is zero, the general return must be even lower (p. 5). However, this relies heavily on the assumption that the "ceiling" logic holds; it remains plausible that returns are low for reluctant compliers but high for those who voluntarily pursue higher education, making the generalization to the entire field a strong claim.

Finally, the presentation of the results lacks transparency in key areas. The sample sizes fluctuate across different tests (from 79 to 49) without a clear reconciliation, and the results of robustness checks (such as including controls) are described in the text without accompanying regression tables (pp. 6–7). Additionally, the authors' treatment of formal bias tests is complex; they note that the standard PET-PEESE method shows "weak signs of publication bias" that are "very model dependent" (p. 7), but also highlight that it becomes significant when outliers are included (p. 8). While the article successfully casts doubt on the precision of the 8.2% consensus

estimate, the proposed correction to 0–3% appears to be an aggressive interpretation of data that is limited by both its specific population and the statistical opacity of some correction procedures.

The Bottom Line

Clark and Nielsen provide strong evidence that publication bias inflates the reported returns to compulsory schooling, successfully challenging the certainty of the consensus 8–10% estimate. However, their counter-claim that the true return is merely 0–3% is less credible, marred by the difficulty of generalizing from compulsory schooling compliers to the broader population and some methodological choices in aggregation. The article is a valuable corrective that likely overcorrects.

Potential Issues

Generalization from a narrow sample: The article's main conclusion that "the implied average causal returns to an extra year of schooling will be only in the range 0%–3%" suggests a broad, general finding (p. 1). However, the analysis is based exclusively on studies of compulsory schooling law changes. These studies estimate a Local Average Treatment Effect (LATE) that applies only to a specific, non-representative sub-population: students who are compelled to stay in school by the law change. The authors acknowledge this limitation but argue that since this group is often thought to have *higher* returns than the average student, a near-zero finding here is damning for the general consensus (p. 5). While this is a logical counter-argument within the literature, it relies on a specific theoretical assumption about the relationship between LATE and ATE. If that relationship does not hold—if voluntary students actually benefit far more than reluctant ones—the generalization collapses.

Transparency of the main estimate's uncertainty: The article's central quantitative claim is that the "true" return to education is "around 3%" based on the regression intercept (p. 7). While the authors report p-values and ranges for the intercept in *controlled* specifications, they do not provide a standard error or confidence interval for the headline *uncontrolled* intercept in the text (p. 7). This omission makes it slightly more difficult for the reader to immediately assess the precision of the primary visual evidence presented in their figures. For an article whose primary goal is to provide a more accurate point estimate, the absence of a clear confidence interval around the main uncontrolled estimate is a shortcoming in reporting.

Clarity of publication bias analysis: The article's analysis of publication bias relies on multiple statistical tests, but the reporting is difficult to follow and the interpretation is not presented systematically. The authors report that the standard PET-PEESE procedure shows "weak signs of publication bias" in the main specification,

but then note that in other specifications (e.g., with outliers included), the PEESE test becomes significant (pp. 7–8). The results of these different model specifications are scattered throughout the text without a summary table, making it difficult for a reader to track which assumptions produce which results. This lack of a clear, systematic presentation of the formal test results may reduce confidence in the authors' ultimate conclusion.

Use of non-standard data aggregation methods: The article's procedure for handling studies that report multiple estimates may not be statistically optimal. The authors state that when a study did not specify a "primary model," they "average all causal estimates and their standard errors" (p. 6). Simply averaging standard errors is not a statistically valid method for calculating the uncertainty of a pooled estimate, as it does not account for the likely correlation between estimates from the same study or their respective sample sizes. While this pragmatic choice applies only to a subset of the data, a more rigorous approach would involve calculating a pooled variance. This methodological choice may have produced an inaccurate measure of uncertainty for the data points constructed through this averaging procedure.

Presentation and transparency issues: Several aspects of the article's presentation could be improved to enhance transparency and reproducibility. First, the sample size varies across different analyses (e.g., N=79, N=65, N=59, N=49) due to missing data or specific exclusion criteria, and while the text explains these drops, the article does not provide a clear reconciliation table to help the reader track which estimates are included in each test (pp. 6–9). Second, the article claims its main finding is robust to the inclusion of numerous controls, but these results are summarized only in the text without a standard regression table (pp. 6–7). The absence of a table prevents readers from assessing the stability of the main coefficient and the effect of the controls.

Future Research

Expansion of instrumental variables: Future work should apply the same rigorous publication bias detection methods to other identification strategies beyond compulsory schooling laws. Analyzing studies utilizing twin fixed effects, lottery admissions, or distance-to-college instruments would determine if the "null-result penalty" is unique to the low-ability population affected by CSLs or if it is a systemic issue across all estimates of educational returns.

Preregistered meta-analysis: To resolve the ambiguity regarding model specification and sample selection, researchers could conduct a preregistered meta-analysis. This would involve defining inclusion criteria and statistical models—specifically the method for correcting publication bias—prior to data collection. Such a study should explicitly report confidence intervals for bias-corrected intercepts to provide a statistically valid range for the "true" return.

Forensic analysis of "0-hacking": Future research could investigate the mechanism behind the clustering of estimates near zero. This could involve a granular forensic analysis of the distribution of t-statistics in the 0–4% range, or qualitative surveys of labor economists regarding their practices when encountering negative or null results. Understanding whether this clustering results from p-hacking (manipulating significance) or selective submission (file-drawer effect) is crucial for designing better correction methods.

isitcredible.com