

Schooling and Labor Market Consequences of School Construction in Indonesia

A Comment on Duflo (*American Economic Review*, 2001)

David Roodman *

Journal of Comments and Replications in Economics, Volume 5, 2026-5, DOI: [10.18718/81781.56](https://doi.org/10.18718/81781.56)

JEL: I2, J31, O15

Keywords: Education, Wages, Reanalysis

Data Availability: The Stata code to reproduce the results of this replication can be downloaded at JCRE's data archive (DOI: [10.15456/j1.2025106.1743527847](https://doi.org/10.15456/j1.2025106.1743527847)).

Please Cite As: Roodman, D. (2026). Schooling and Labor Market Consequences of School Construction in Indonesia. A Comment on Duflo (*American Economic Review*, 2001). *Journal of Comments and Replications in Economics*, Vol. 5 (2026-5). DOI: [10.18718/81781.56](https://doi.org/10.18718/81781.56)

Abstract

Duflo (2001) exploits a 1970s schooling expansion in Indonesia to estimate individual-level schooling and labor impacts and the returns to schooling. I correct data errors, adjust for potential sources of bias, follow up later in life, and check sensitivity to two specification choices that are applied to some regressions and not others. Headline estimates from the original, 1995 follow-up are generally not robust to the revisions. Returns to schooling are weakly identified. But certain “fingerprints” from Duflo (2001) specification checks—impacts appearing mainly in the birth cohorts or at the grade levels where expected—recognizably persist. The best explanation, though weakened, is causation from the schooling expansion to individual schooling attainment to wages.

*Coefficient Giving. david.roodman@coefficientgiving.org

Thanks to Esther Duflo for data, code, and comments; to Daniel Feenberg at the National Bureau of Economic Research for running and trouble-shooting scripts; and to Mikey Jarrell, Nathan Lazarus, Jonathan Morduch, Otis Reid, Justin Sandefur, and anonymous referees for comments. Estimates and inferences are performed with *boottest* (Roodman et al. 2019) *cic* (Melly and Santangelo 2015), and *reghdfejl* (Roodman 2025). Tables and figures were made with *blindschemes* (Bischof 2017), *coefplot* (Jann 2014), *esttab* (Jann 2007), and *palettes* (Jann 2023).

Received April 16, 2025; accepted October 24, 2025; published April 07, 2026.

©Author(s) 2026. Licensed under the Creative Commons License - Attribution 4.0 International (CC BY 4.0).

Introduction

In 1973, the government of Indonesia dedicated part of an oil windfall to a project to erect thousands of schoolhouses across the nation's far-flung archipelagic domain. The Instruksi Presiden Sekolah Dasar program (Inpres SD) became one of the largest schooling expansions ever. In its first six years alone, it doubled the country's stock of primary schools (Duflo 2001). Following up on affected men in early adulthood, Duflo (2001, p. 812) concludes that Inpres SD "was effective in increasing both education and wages."

Duflo (2001) is influential for several reasons. Typifying its moment in intellectual history, it exploits strategies for causal inference such as difference-in-differences (DID) and natural experiment-based instrumentation (Angrist and Krueger 1999). It innovates, in part by applying such methods to a developing-world setting. The natural experiment it brings to light dwarfs most in the education literature. Threats to identification are confronted: a placebo check in the pre-treatment period returns a null result; the instruments pass a validity test; other potential sources of bias are named and addressed. Graphical analysis makes the case that the schooling shock left distinctive fingerprints in the data, in particular that it manifests with the expected timing and at the expected grade levels. Few studies so credibly claim to identify the impacts of schooling at scale.

Precisely because of its continuing importance, I revisit the Duflo (2001) quasi-experiment. The passage of time has brought improvements in econometric methods. It has brought more data too, as the subjects of the quasi-experiment have transited through their labor careers, leaving traces in regular natural surveys. And any fresh look may add insight.

The reanalysis proceeds in two steps. First, the data set is reconstructed from primary sources and the original analysis is reimplemented in code. Comparing the original and new versions surfaces bugs in both and builds confidence in the (debugged) new version.

Second, the Duflo (2001) specifications are modified in several ways to check robustness and improve inference. A rational reader will put some weight on the possibility that in assembling these modifications, I have "reverse *p*-hacked," trying many and selectively reporting, consciously or not, the most provocative results. Like Duflo (2001), this reanalysis was not pre-registered. To combat the threat and the perception of reverse *p*-hacking, the modifications are chosen to minimize consumption of "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011). The study population is not modified, nor the sample by, say, removal of outliers. No new controls are added. The original estimators are largely hewed to. Some modifications are borrowed from my earlier work rather than fabricated anew for this project.

The main changes resulting from both steps are:

1. Correcting data entry errors in baseline variables, which affect about 10% of observations.
2. Clustering standard errors, by the geographic unit of treatment.
3. Incorporating later data on the same subjects, as in the Easterly, Levine, and Roodman (2004) comment on Burnside and Dollar (2000). This increases the power to detect impacts

on schooling attainment and adds information on labor outcomes in the prime working years.

4. Applying now-standard methods to test for and accommodate weak identification in instrumental variables (IV) estimation, in particular the Kleibergen-Paap (2006) measure of identification strength and a bootstrapped Anderson-Rubin test to construct confidence intervals.
5. Using a two-piece linear spline model (“kink model”) to formally test an informal characterization in the original text about the presence of a trend break with timing attributable to Inpres SD.
6. Copying specification choices from some regressions to others within the original paper, as in the Roodman (2018) comment on Bleakley (2007)—choices that are in this case essentially undiscussed in the original. The choices are two: whether to incorporate survey weights and whether to include a particular vector of controls.

The first four items are relatively uncontroversial. Clustering by geographic treatment unit when performing DID on microdata has been standard since Bertrand, Duflo, and Mullainathan (2004). Duflo (2004a) also follows up later on the same population. Since 2001, the literature on weak identification in IV estimation has shifted norms among practitioners.

The last two items are more debatable, for each makes a trade-off of unknown degree between bias and power. Incorporating survey weights reduces the risk of bias from endogenous survey sampling but tends to reduce precision. Controlling for pre-treatment trends through the kink model should reduce bias from non-parallel trends but also consumes power since it asks not merely whether a post-treatment trend is positive, but whether it is more positive than the pre-trend. It is hard to know whether dropping a set of controls raises or lowers the exogenous fraction of identifying variation. Because these changes are more matters of judgment, they are performed and reported in all combinations. If they were introduced cumulatively in a particular order, the choice of that order would consume researcher degrees of freedom, justifiably raising a question in the reader’s mind about whether it was favored in order to create a certain impression.

Duflo (2001)’s headline ordinary least squares (OLS) findings—0.12–0.19 more years of schooling per new school per 1,000 children and 1.5–2.7% higher wages—both fade with data corrections, clustering, and reversal of the two specification choices. The point estimates fall, while, for schooling, the corresponding “placebo” impact rises. The impact on schooling attainment is ambiguous enough that instrumental variables estimates of the returns to schooling often suffer from weak identification, while weak-identification-robust inference sometimes produces infinite confidence intervals. The story does not fundamentally change when adding data from later follow-ups.

Yet, especially in those larger samples, the evidence from the Duflo (2001) fingerprint checks cuts against the null of zero impacts. In the kink-based regressions, the point estimates are almost always positive. True, many of the new estimates are not significant at conventional levels, especially when the questioned control set is dropped. But there is statistically strong evidence of an impact on primary school completion, with timing attributable to Inpres SD. A similar kink is found for wages, and with similar strength when the questioned controls are included.

For these reasons, the hypothesis that Inpres SD increased schooling and thereby wages remains

the best explanation for the evidence at hand. But the theory does not earn the same credence as in the original study. And the returns to Inpres-stimulated schooling are hard to measure with any precision.

Section 1 of this paper reviews the methods of Duflo (2001). Section 2 describes and motivates the changes to data and method. Section 3 replicates key Duflo (2001) results and documents the effects of changes in the original, 1995 follow-up. Section 4 moves to later follow-ups. Section 5 brings the same revisions to the analysis of schooling impact by grade level. Section 6 summarizes results from changes-in-changes (CIC), an estimator that is designed to accommodate non-parallel trends rather than model and control for them. Section 7 concludes.

1 The Duflo (2001) specifications

The authority to carry out the Inpres SD program flowed from annual presidential instructions, appendices of which set construction targets for every *kabupaten* (regency) and *kotamadya* (municipality); these are Indonesia's second-level administrative units, which I call "regencies" for short. Duflo (2001)'s treatment intensity indicator is the number of schools planned for construction in a regency between 1973/74 and 1978/79 per 1,000 children aged 5–14. Duflo (2001, note 1) cites a government finding that in this period, actual construction closely matched planned. Using data from the nationally representative Intercensal Population Survey (SUPAS) of 1995, the study examines how boys from less- and more-treated regencies fared in early adulthood: how many years they ultimately stayed in school; whether they held wage employment; and, if so, how much they earned per hour and month.

Duflo (2001) works in the difference-in-differences framework. Since the outcome data come from a single, 1995 cross-section, the study arrays observations into a two-dimensional structure by grouping on year of birth and place of birth, as in the Acemoglu and Angrist (2000) study set in the U.S. The oldest subjects are the pre-treatment cohorts because Inpres SD was launched too late to directly affect them. Analytical time therefore runs from older to younger people. The included cohorts range in age from 2 to 24 as of 1974.

To illustrate ideas, Duflo (2001) begins with classical 2×2 DID. It is performed by fitting

$$Y_{ijt} = D_j T_t \delta + \nu_j + \eta_t + \epsilon_{ijt} \quad (1)$$

where Y_{ijt} is an outcome for individual i in treatment group j in period t ; $j = 0, 1$ for the control and treatment group; $t = 0, 1$ for the pre- and post-treatment periods; D_j is a dummy for being from a high-treatment regency; T_t is a dummy for the younger, treated cohorts; δ is the impact parameter; ν_j and η_t are treatment group and cohort fixed effects; and ϵ_{ijt} is an exogenous, mean-zero error.

Next, Duflo (2001) elaborates beyond the 2×2 grid. Each yearly birth cohort gets its own time index and each of the approximately 290 regencies in Indonesia as of 1995 becomes its own treatment group. Treatment is taken as continuous, as planned schools per 1,000 children. These extensions lead to the two-way fixed effects (TWFE) model,

$$Y_{ijt} = \left(D_j \tilde{\mathbf{T}}_t \right)' \delta + \nu_j + \eta_t + \epsilon_{ijt} \quad (2)$$

where t and j index birth cohorts and regencies and D_j is now treatment intensity in regency j . $\tilde{\mathbf{T}}_t$ is a vector of variables that depend only on time, i.e., on age in 1974. Duflo (2001) defines $\tilde{\mathbf{T}}_t$ in several ways. In its least flexible form, $\tilde{\mathbf{T}}_t$ is the T_t defined earlier, a single dummy for the younger cohorts. In this case, men aged 2–6 in 1974 form the “after” cohorts, those aged 12–17 are the “before” cohorts, and the rest are dropped. (The slight asymmetry in these definitions, with one age range wider than the other, is not discussed in the original text.) In its most flexible form, $\tilde{\mathbf{T}}_t$ is a set of yearly cohort dummies, and all cohorts aged 2–24 constitute the sample. The estimate δ is then an event study, i.e., a vector of cohort-specific linear associations between treatment and outcome. An in-between option that Duflo (2001) sometimes favors makes $\tilde{\mathbf{T}}_t$ a set of yearly cohort dummies, except that the pre-treatment cohorts aged 12–24 in 1974 are given one dummy.

Alongside the treatment terms and fixed-effect dummies, most Duflo (2001) TWFE specifications include controls. All are products of baseline variables and cohort dummies, to allow the variables cohort-specific associations with the outcome. The augmented specifications read

$$Y_{ijt} = \left(D_j \tilde{\mathbf{T}}_t \right)' \delta + (\mathbf{C}_j \otimes \mathbf{T}_t)' \gamma + \nu_j + \eta_t + \epsilon_{ijt} \quad (3)$$

$\mathbf{T}_t$ is a full set of cohort dummies. In successive variants, $\mathbf{C}_j$ grows from one to three variables. It starts with the number of children in a regency in 1971 aged 5–14, which is the denominator of the treatment indicator. It gains the fraction of the population enrolled in primary school, to reduce bias from mean reversion. Then it gains regency-level spending on a separate water and sanitation program, which might confound the effects of school construction. I will call these variants the minimal, intermediate, and full control sets.

A final step in the Duflo (2001) empirics is to perform two-stage least squares (2SLS) to estimate the returns to schooling. The 2SLS specifications use (3) for the first stage, setting Y_{ijt} to years of schooling (S_{ijt}).¹ They adapt (3) for the second stage, replacing $D_j \tilde{\mathbf{T}}_t$ with the instrumented variable, S_{ijt} , and setting Y_{ijt} to either a dichotomous indicator of wage sector employment or to log monthly or hourly wages. Wage regressions are normally restricted to participants in the wage sector, although in a robustness test, wages of non-wage workers are imputed. $\tilde{\mathbf{T}}_t$ takes two of the forms described earlier: binary, and having one dummy for ages 12–24 and another for each yearly age from 11 down to 2. I call these the “instrument by young/old” and “instruments by birth year” variants.

One threat to identification in DID is the presence of non-parallel trends—that is, distinctive trends in outcomes within the various treatment arms that are separate from the causal story of interest. They may be visible before treatment and persist after. A standard specification check is therefore for relative trends before treatment, or “pre-trends.” Duflo (2001, Tables 3 & 4) checks

¹A more precise description of S_{ijt} is “highest year of school attended.” The two concepts can diverge. Suppose the construction of a new primary school causes a child to switch to it from a farther-away middle school. And suppose that the child exits schooling upon graduation from that primary school. Later in life that child could report a lower “highest year of school attended” without any reduction in years of schooling.

for pre-trends through a placebo or “control” experiment in which the pre- and post-treatment periods are shifted to lie fully within the true pre-treatment era. Those aged 18–24 in 1974 become the pre-treatment group and the actual pre-treatment group, aged 12–17, becomes the placebo post-treatment group.

Alongside the regression results, Duflo (2001) disaggregates the specifications along one dimension or another to check for fingerprints of the intervention. By surfacing patterns that would unlikely occur by chance, these increase confidence in the main results. One instance, already mentioned, is the event study, in which the treatment variable is interacted with a full set of birth cohort dummies, and the coefficients plotted. Here, it appears that the intermediate control set is used.² The other fingerprint check decomposes schooling attainment into a set of survival dummies: for whether a person attended school for at least one year, or at least two, etc. These dummies become dependent variables in the TWFE specification (2), the one without extra controls. The resulting plot of “difference in differences in CDF [cumulative distribution function]” shows whether the primary schooling expansion indeed manifested as greater continuation through primary school.³

2 Comments

I comment on seven aspects of the Duflo (2001) regressions: data transcription errors, clustering, the possibility of heterogeneous treatment effects, a vector of controls used in some regressions and not others, weak identification, observation weighting, and the handling of potentially non-parallel trends. The comment on effect heterogeneity leads to a reassuring diagnostic check. The rest lead to revisions in data and method.⁴

2.1 Data corrections

Duflo (2001) gathers early-1970s regency-level variables from government publications. For example, reports on the 1971 census provide total population by regency and population aged 5–14. Both are denominators for other variables.

I reconstruct all the 1970s variables but the one on water and sanitation spending, which is used in only some specifications and does not affect results much.⁵ I find some discrepancies between the Duflo (2001) figures and official sources (BPS 1972, 1974; Bappenas 1973–78). The transcription errors affect the treatment indicator for about 10% of regencies. For example, probably because regencies are listed in different orders in different publications, a few pairs or quadruplets have their numbers rearranged. The original and corrected versions of the treatment

²This is inferred by comparing to Duflo (1999, chapter 1, Table A1).

³Duflo (2001, p. 804) defines the dependent variables as the dummies $\mathbf{1}\{S_{ijt} \leq m\}$, i.e., whether individual i completed m years of schooling or less. In fact, the dummies appear to be $\mathbf{1}\{S_{ijt} > m\}$, which merely flips the signs on the estimates.

⁴Previous versions of this paper mention all of these issues. However, one, the vector of controls, was touched on only briefly. The first commit of the code to GitHub, in October 2022, [loops over control sets](#): no controls, and the minimal, intermediate, and full control sets. But I initially dropped most discussion of no-controls results because they did not seem important for kink-based inference, which I most emphasized.

⁵Page scans of primary sources and an annotated Excel tabulation are at github.com/droodman/Duflo-2001/tree/main/Printed%20sources.

intensity indicator D_j are correlated 0.81 across regencies.⁶ Duflo (2001, Table 2) confirms that Inpres SD targeted areas with many young children not enrolled in primary school, by reporting a regression of the log number of schools planned for each regency on the log number of children aged 5–14 and the log primary school non-enrolment rate; the corrections lift the R^2 from 0.78 to 0.83.

In response to an earlier version of this paper, Esther Duflo, Mikey Jarrell, and Nathan Lazarus endorsed the corrections.

2.2 Clustering

The variance estimates in Duflo (2001) are classical: they are not adjusted for the possibilities of heteroskedasticity and intra-regency and intertemporal dependence. In the other chapters of the Duflo (1999) thesis from which Duflo (2001) is derived, standard errors are clustered. And Bertrand, Duflo, and Mullainathan (2004) demonstrates the value of the clustering correction (Liang and Zeger 1986, Arellano 1987) when performing TWFE on microdata with treatment assigned by geography.

2.3 Heterogeneous treatment effects

Recent scholarship highlights the counterintuitive way that effect heterogeneity can bias TWFE estimates of mean effects (de Chaisemartin and D'Haultfœuille 2020; Goodman-Bacon 2021). This methodological critique does not apply with force to Duflo (2001). The scope for bias is greatest when treatment is staggered across units, unlike here. Moreover, Jakiela (2021)'s check for heterogeneity produces little evidence of such: see appendix A.

2.4 Controls based on the number of children in the regency of birth

As described in section 1, most TWFE regressions in Duflo (2001) include in C_j the number of children aged 5–14 in the regency of birth as of 1971. The variable is interacted with the cohort dummies to control for cohort-specific associations between population and the outcomes. This specification choice warrants two comments. First, Duflo (2001) does not always include the controls. They enter the main TWFE regressions, but are dropped when the dependent variables are dummies for having completed at least a certain number of years of schooling (Duflo 2001, Figure 2). Second, there is little explanation for why they are included, except that they seem to matter.⁷

The concern that baseline population could confound, if omitted, is well founded. The cross-regency regression mentioned in section 2.1 shows that the number of new schools increased with subunitary elasticity with respect to population, meaning that more populous provinces received lower treatment per child. And since large regency population is correlated

⁶A particular complication affects baseline school attendance. Duflo (2001) appears to take the numerator, people aged 5 and older attending school, from BPS (1974a), and the denominator, total population, from a different publication, such as BPS (1972). The figures for Irian Jaya (Papua) are much lower across the board in BPS (1974a), perhaps because school attendance was only queried for a subpopulation. For consistency, I therefore use the denominator that accompanies the numerator in BPS (1974a), namely, total population aged 5 and above.

⁷The Duflo (2000) working paper states in an endnote, “The number of schools per child was negatively correlated with the number of children, and I want to abstract from any time-varying effect of regional population. It turns out to be important in the wage equation.”

with being on the main island of Java—there, average regency population was 1.08 million in 1995, against 0.41 million elsewhere—population is probably related to schooling and labor market outcomes for non-schooling reasons. In the 1970s and 1980s, Java became a locus of export-oriented industrialization, which affected individual schooling decisions, labor market experiences, and migration patterns (Heneveld 1979; Hill, Resosudarmo, and Vidyattama 2008). If these effects differed by birth cohort, the regency fixed effects would not fully remove them.

However, it is not obvious that controlling away a known correlate of treatment improves estimation, since the treatment variation that remains is not necessarily more exogenous than that expunged. One can construct stories of confounding for any nonrandom component of treatment, such as the time-varying effects of the baseline primary school attendance rate, which are also controlled for in the intermediate and full control sets. Duflo (2004b) argues that in Duflo (2001), “identification is made possible because the allocation rule for schools is known.” That point loses some force when one controls away the large, known components of allocation. For the reasons above, it is informative and minimally arbitrary to check the effect of dropping the controls based on population.

2.5 Weak identification

Econometricians have made progress since 2001 in understanding and managing the effects of weak identification on IV estimators (Stock and Yogo 2002, Moreira 2003, Kleibergen and Paap 2006, Montiel Olea and Pfluger 2013). As identification weakens, IV biases toward OLS. Various tests have been developed for strength of identification. But screening on such pre-tests can itself generate bias and size distortions. Better, argues Andrews, Stock, and Sun (2019, §4.3), to favor estimators that are robust to weak identification. For forming confidence intervals, Finlay and Magnusson (2019)’s Monte Carlo–based review favors the Anderson-Rubin (AR, 1949) test with a wild-bootstrapped distribution for the test statistic (Roodman et al. 2019).⁸ In comparison with the conditional likelihood ratio test of Moreira (2003), the AR test has the advantage of being defined for exactly- as well as over-identified regressions. The price of this generality is that the test compounds the hypothesis of primary interest—that coefficients on instrumented variables take given values—with a second—that the instruments are valid. Rejection on the test is therefore ambiguous. Of course, whichever hypothesis is the cause of a rejection, that rejection still speaks to causal interpretation of the results.

2.6 Endogenous sampling

The survey used for follow-up in Duflo (2001) employed a complex sampling design. It oversampled low-population regencies in order to produce comparably accurate statistics for all regencies.⁹ Within regencies, it sampled urban and rural areas non-proportionally. Survey weights included in the data set allow for adjustment for this non-proportional sampling. The code Duflo sent me incorporates weights in the 2×2 DID causal estimates but not in the TWFE causal estimates. This leads to the question of how the decisions to weight one and not the other affects results. The original text does not discuss weights.

⁸Simulations in Davidson and MacKinnon (2010) of the non-clustered but heteroskedasticity-robust analog also demonstrate the good size properties of the wild-bootstrapped AR test.

⁹Surbakti (1995, p. 19) states this motivation with respect to the contemporaneous socioeconomic survey (SUSENAS) rounds.

In general, there are two reasons to weight observations. One is to improve statistics as *descriptions*, i.e., to make them more representative for the original setting, in this case Indonesia of 1995. That rationale is not compelling here. Not incorporating the weights merely makes the results representative for a different, hypothetical population, one that may be no less relevant for testing theories or making decisions in other contexts. And if effects are fairly homogenous, as suggested above, weighted and unweighted results should not differ greatly. The other reason is to improve *causal estimates*. For example, if sampling is endogenous within a given estimation context, then incorporating weights can counteract selection on the basis of the outcome, a potential source of bias (Hausman and Wise 1981). However, incorporating survey weights typically increases the variance of estimators.

Appendix B reviews the conditions under which unweighted OLS consistently estimates the slope of the linear best fit in the population. Two sufficient conditions emerge. Let the variables x and y be the treatment and outcome and let the relationship of interest in the population be $y = \beta x + e$. And let p be the probability that a given member of the population is sampled. Approximately speaking, the *purely exogenous* sufficient condition allows p to be correlated with x but not e , while the *purely endogenous* condition allows p to be correlated with e but not x . Sampling can be related to the observed or unobserved determinant of the outcome, but not both (Solon, Haider, and Wooldridge 2015). In fact, as spelled out in the appendix, an additional condition is required in each case. For example, in the purely endogenous case, some patterns of heteroskedasticity can still render unweighted OLS inconsistent.

Unlike in the theory just thumbnailed, the regressions of interest here include controls, notably the dummies for year of birth and regency of birth. To adapt the sufficiency conditions, we apply them to outcome and treatment after partialling out the controls. If subjects did not *migrate* within Indonesia, this step would eliminate nearly all the association of the sampling probability with *both* the treatment and the error. Why? There is no sign that within the Duflo (2001) samples, the sampling probability is related to age—little sign that, for example, that men in their 30s were oversampled relative to men in their 20s from the same regency.¹⁰ And if the sampling probability is constant over time within regencies, then the variation in it would be a pure regional effect, which would effectively be absorbed by the regency fixed effects.

The catch in that argument lies in the distinction between regency of birth and regency of residence at follow-up. The Duflo (2001) regressions control for birth regency fixed effects. The sampling probability depends mainly on the place of residence at follow-up—urban or rural, in a low- or high-population regency. The two differ for migrants. Internal migration was an important feature of Indonesian development in this period (Hill, Resosudarmo, and Vidyattama 2008). The government promoted it to reduce overcrowding in Java, even as many people moved to cities for work. Hsiao (2024) finds that Inpres stimulated outmigration. If factors not controlled for affected both the outcomes and the decision to move, these could create a correlation between the sampling probability and the regression error terms. And if the sampling probability is related to both treatment and the error, unweighted OLS is inconsistent.

The literature takes two main approaches to testing whether a given unweighted OLS specification is consistent (Bollen et al. 2016). The *weight association* approach directly tests the sort of

¹⁰A weighted regression of age at follow-up on the sampling probability, restricted to the 1995 follow up and those aged 2–24 in 1974, and controlling for birthplace fixed effects, yields a t statistic of -0.58 .

sufficiency conditions just mentioned—though to date, it has only been used to check the purely exogenous condition. The *difference-in-coefficients* approach works in the tradition of Hausman (1982) testing, calculating the statistical significance of the gap between unweighted and weighted estimates (Pfeffermann 1993). Within the Hausman framework, unweighted OLS is the efficient estimator and weighted is the more-robust estimator.¹¹ Hausman testing avoids the error of checking only the purely exogenous sampling condition, when satisfaction of the purely endogenous condition could also avert inconsistency. I therefore opt for the Hausman approach below. Because errors are not assumed to be classical, I will use a bootstrap rather than analytical formulas to estimate the variance of the difference between the two estimates, which is the denominator of the test statistic (Cameron and Trivedi 2005, p. 378). In particular, I will run clustered wild bootstraps without the null imposed, with 9,999 replications (Cameron, Gelbach, and Miller 2008; Roodman et al. 2019).¹²

When a Hausman test casts doubt on the null that the unweighted and weighted estimates agree asymptotically, that can be explained in two ways. Treatment effects may be heterogeneous enough that weighting causes a large change in the estimated average effect. Cutting against that explanation would be the null result reported in appendix A from a check for effect heterogeneity. The alternative explanation is that the unweighted estimates are inconsistent.

One disadvantage of using inverse-sampling-probability weights is that their distribution can be skewed, making some observations inordinately influential. In the sample for the schooling regression just mentioned, the mean weight is 211, the 99th percentile is 772, and the maximum is 2,643. To limit the influence of extreme weights, I will winsorize them to the median weight plus 4 times the interquartile range. This is one simple approach evaluated through simulation in Potter and Zheng (2015).¹³ In this case, it yields a limit of 903.

2.7 Non-parallel trends

The geographic treatment variation in Inpres is not as good as random, for high- and low-treatment regencies presumptively differed at baseline. And it would be surprising if regencies that started out different would nevertheless have experienced the same *trends* in the absence of Inpres SD. In other words, the parallel trends assumption is almost surely violated (Kahn-Lang and Lang 2019). How much it is violated is unclear *a priori*. One common response to this sort of concern, exemplified by the Duflo (2001) placebo checks described in section 1, is to check for non-parallel trends in the pre-treatment era, and then rely on the experimental results if the pre-trend test result is not significant. But in simulations calibrated to a set of published DID studies, Roth (2022) concludes that half the time tests of this sort miss the presence of pre-trends that are large enough to cause substantial bias. The problem, as Kahn-Lang and Lang (2019) write, is that “not rejecting the null hypothesis is not equivalent to confirming it. Yet, in such

¹¹A striking fact makes the connection to Hausman testing: weighted OLS is equivalent to unweighted 2SLS regression in which all the predictors x_i are instrumented by their products with the inverse-probability weights, wx_i (Winship and Radbill 1994, note 13). So comparing weighted and unweighted OLS can be reframed as comparing OLS and 2SLS.

¹²Separate bootstraps are run for the unweighted and weighted regressions. The difference between the unweighted and weighted estimates is computed for each of the 9,999 bootstrap data sets. Then their variance is computed. The bootstraps for the two regressions are started with the same random number seed, to produce corresponding simulated data sets. For a modest boost in power, Davidson and MacKinnon (1999) favors imposing on the bootstrap data generating process the null that is being tested. That guidance is not applicable here because the null that the unweighted and weighted estimates agree cannot be imposed on either bootstrap alone.

¹³This is a late change prompted by a comment from a referee. It increases consistency with the Roodman (2023) reanalysis of the Khanna (2023), in which the weight distribution is even more skewed.

settings, economists may be tempted to, in Brad DeLong's terms, 'seize the high ground of the null hypothesis,' which can lead to incorrect estimation of treatment effects." Roth (2022) also shows that filtering estimates for passage of a pre-trend pre-test itself induces bias, because of regression to the mean in the post-treatment period. As when facing weak identification, it appears less biased to estimate effects in a way that is robust to the problem than to filter results with a pre-test for it.

Because Inpres treatment was roughly simultaneous in all regencies, if varied in intensity, it is relatively straightforward to model and control for pre-treatment trends.¹⁴ The approach taken here formalizes an informal characterization in Duflo (2001) of the event studies. The paper's Figure 1 shows how the cross-sectional association between treatment and schooling attainment evolves, birth cohort by birth cohort. Figure 3 adds a similar plot for log wages. Of the first, the text observes that the "coefficients fluctuate around 0 until age 12 [as of 1974] and start increasing after age 12" (Duflo 2001, p. 801). A sustained trend break at age 12 in 1974 is plausible because Inpres began adding primary schools in 1974 and continued doing so through at least 1980, making the 11-year-olds of 1974 approximately the first birth cohort young enough to attend the new schools, just for 6th grade. However, because there are uncertainty bands around the point estimates in Duflo (2001)'s Figure 1 (which will be widened by clustering), and because this observation of a change in trend is not formalized, it is not clear how easily the null of *no change* can be rejected.

To check for kinks, I will run regressions formulated by modifying $\tilde{T}_t$ in (3). $\tilde{T}_t$ will become a vector with two terms: a time trend t (age in 1974) and a spline term that emerges for younger cohorts after age 12:

$$\max(0, 12 - t) \quad (4)$$

In OLS/reduced-form regressions, both components of $D_j \tilde{T}_t$ will enter as regressors. In 2SLS, the first will enter as a control while the second will instrument the schooling variable. I will restrict the model to ages 2–22, to extend each spline segment 10 years. This choice is minimally arbitrary and nearly matches the 2–23 range of Duflo (2001)'s Figure 1. Making the lengths multiples of five also minimizes distortions from age heaping, which is the tendency of poorer people to have their survey-recorded ages rounded to the nearest multiple of 5.

3 *Revising the Duflo (2001) causal estimates*

3.1 *2×2 DID*

The reanalysis begins where the original analysis begins, with 2×2 DID estimates based on (1). A birth regency is "high treatment" if it garners a positive residual in a cross-regency regression of planned school construction on the number of children 5–14.

See Table 1 for the DID results. The first column contains the Duflo (2001) estimates. In the experiment of interest, which benchmarks outcomes for the 2–6-year-olds of 1974 against those for the 12–17-year-olds, the schooling impact is 0.12 years and the wage impact 0.026 log points. The ratio of those is the Wald DID estimate of the return to a year of schooling, 0.22, which

¹⁴Incorporating pre-trends is harder when treatment is staggered across cross-sectional units. See the debate over state-level no-fault divorce laws in Peters (1986), Wolfers (2006), and Lee and Solon (2011).

is also shown in the column. In the placebo check, which compares the pre-treatment groups aged 18–24 and 12–17 as of 1974, the estimated “impacts” are close to zero: 0.034 for years of schooling and 0.0070 for the log hourly wage.

The *third* column of Table 1 shows the results from the data and code provided by Duflo. They differ from the published results.¹⁵ The estimated impact on the log hourly wage falls from 0.026 to 0.016 and the return to schooling from 0.22 to 0.14. These estimates incorporate survey weights. For consistency with later tables, the second column shows the results when modifying the Duflo code to drop weights. This takes us further from the published results, which suggests that the latter indeed incorporate weights.

Table 1: 2×2 DID estimates of effects on schooling attainment and the log hourly wage, 1995

	Reported in Duflo (2001)	Duflo data & code		New	
<i>Effect on schooling</i>					
Experiment	0.12 (0.089)	0.10 (0.089)	0.11 (0.088)	−0.080 (0.15)	−0.029 (0.16)
Placebo	0.034 (0.098)	−0.037 (0.098)	0.0047 (0.098)	0.0030 (0.11)	0.072 (0.14)
<i>Effect on log hourly wage</i>					
Experiment	0.026 (0.015)	−0.0018 (0.015)	0.016 (0.015)	−0.010 (0.026)	0.0065 (0.026)
Placebo	0.0070 (0.016)	0.00083 (0.016)	0.0044 (0.016)	0.0047 (0.017)	0.014 (0.019)
<i>Return to schooling (Wald DID)</i>					
Experiment	0.22	−0.017 (0.15)	0.14 (0.13)	0.13 (0.24)	−0.22 (1.81)
Placebo	0.21	−0.023 (0.48)	0.93 (17.5)	1.55 (55.2)	0.19 (0.30)
Data corrections				✓	✓
Clustered				✓	✓
Weighted	✓?		✓		✓

Notes: Sample is restricted to wage earners aged 12–17 and 18–24 in the placebo experiment and 2–6 and 12–17 in the experiment of interest, with the younger range in each case being the post-treatment group. Wald DID estimates are ratios of corresponding results in first two panels. The corrected high-treatment indicator is constructed by returning to primary sources for planned school construction and student populations, then defining high treatment as having a positive residual in a linear regression of the former on the latter (Duflo 2001, note 2). “Clustered” means clustered by regency of birth. The Duflo code does not run unweighted 2×2 DID regressions, but is here modified to do so. In the last regression, weights are winsorized to the median weight plus 4 times the interquartile range.

The last two columns document that the 2×2 results are not robust to correcting data errors. With or without weighting, impacts are now higher in the placebo experiment than in the experiment of interest. The schooling impact is wrong-signed. That said, these contrasts are statistically weak, especially in light of the clustered standard errors now reported. The weighted Wald estimate of the return to schooling is −0.22, but is weakly identified and indistinguishable from zero.

The 2×2 DID estimates turn out not to be replicable or robust. This finding does not speak directly to the main results in Duflo (2001), which come from TWFE specifications that extract more information.

¹⁵Despite engagement with the original author, I have not managed to reconcile this discrepancy.

3.2 OLS/reduced-form results

3.2.1 Years of schooling

We move next to OLS TWFE estimates. Samples are restricted to the same cohorts as in the 2×2 DID. Fixed-effect dummies are included for birth year and birth regency. In the original, weights are now dropped. Here, the original estimates are exactly reproduced.

I focus first on the educational outcome, years of schooling. Table 2 displays the original results along with the consequences of introducing changes to address issues raised in section 2. The table is arranged so that the original results appear in the upper left and the ones incorporating all changes are in the bottom right.

In panel A, the minimal Duflo (2001) control set is included; recall that it consists of interactions between cohort dummies and the number of children aged 5–14 in the regency of birth. In the original, adding controls beyond the minimal set does not greatly affect results, so for parsimony those specifications are omitted. As shown in the upper left of the panel, Duflo (2001) estimates that Inpres increased schooling attainment by 0.12 years (standard error 0.025) for each planned new school per 1,000 children. The corresponding placebo estimate is also reported. However, unlike in Duflo (2001), it is rescaled for comparability with the experimental effect. The reasoning runs as follows. The average lag between the two placebo group age ranges—between ages 18–24 and 12–17 as of 1974—is the distance between their midpoints, $21 - 14.5 = 6.5$ years. The placebo estimates are therefore of cumulative trends over 6.5 years. The experimental comparison—between the 12–17 and 2–6 groups—spans an average gap of $14.5 - 4 = 10.5$ years. Comparability calls for rescaling the placebo results by $10.5/6.5$. Despite this magnification, in the original specification, the placebo effect is reassuringly estimated to be just 0.015 years per unit of treatment, barely a tenth of the experimental effect (panel A, column 1).

Moving to column 2, we see that unlike in 2×2 DID, correcting the data errors barely perturbs the point estimates. However, clustering the standard errors increases them by about 50%.

Weighting is introduced in column 3 of panel A. It too increases standard errors, though more modestly. It also makes the placebo impacts bigger than the impacts of the experiment of interest: the main estimate drops from 0.12 (standard error 0.042) to 0.077 (0.054) while the rescaled placebo effect rises from 0.013 (0.055) to 0.083 (0.068). The p values from the Hausman tests comparing weighted and unweighted results, reported in the last column of panel A, cast these shifts as unlikely if treatment effects are homogeneous and unweighted OLS is consistent. The p value is 0.23 for the experimental effect and 0.080 for the placebo effect. And for specification tests, a low significance standard such as $p < 0.05$ is the opposite of conservative. The weighted findings challenge the assumption that pre-treatment schooling trends are close enough to parallel that, if continued into the post-treatment period, they could explain only a negligible fraction of the measured effect.

The bottom row of panel A reports the results from the preferred method for controlling for regency-specific linear trends, using the kink model. These estimates are also depicted in the plots of Figure 1, which are event studies like Duflo (2001, Figure 1) with the kink fits superimposed. For example, in the specification closest to that in Duflo (2001), but subject to data corrections

and clustering, the trend in years of schooling steepens by an estimated 0.0072 per year at the allowed kink point at age 12; see the upper left plot of the figure. Following that steeper trend 8 years to the right from the kink brings us to the center of the age 2–6 range, which is the post-treatment cohort. The kink implies that the 2–6 cohort averaged $8 \times 0.0072 = 0.058$ more years in school for each unit of treatment—more, that is, than would have been the case with no deviation from the pre-trend. This kink-based estimate is displayed in the figure and in panel A along with its clustered standard error, 0.071. The corresponding weighted estimate appears to the right in both displays; it is statistically similar.¹⁶

One specification change has been held back to this point: dropping the Duflo (2001) controls based on child population. Section 2.4 laid out reasons to be somewhat skeptical of their inclusion. In addition, when estimating impacts from distinctive kinks in long-term trends, as preferred here, extra controls are in principle not needed for consistency, even as they consume statistical power. Aside from its connection through Inpres SD, we have little reason to expect initial population to create a confounding kink—little reason, that is, to expect the time-varying impact of initial population to break distinctively at the cohort aged 12 in 1974.

For these reasons, the bottom halves of Table 2 and Figure 1 show the results from rerunning the preceding analysis without the extra controls. (The results in the “original” column of panel B do not appear in Duflo (2001) and are obtained by modifying the corresponding regressions in panel A.) Precision slightly increases. The experimental effect barely falls and the placebo effect rises further. Going by the Hausman p values, weighted and unweighted estimates now better agree. The kink in the long-term schooling trend is even closer to zero, indeed slightly negative.

Under this testing regime, it is not clear that Inpres SD increased time spent in school. As one reverses choices that are applied to these specifications and not some others, the treatment effect falls, the placebo effect rises, and the trend break fades. Where Duflo (2001) estimates that Inpres SD increased schooling by 0.12–0.19 years per unit of treatment, the specification preferred here—kink-based, with weights and without the extra controls—produces a slightly negative and statistically insignificant estimate, -0.027 (0.082). That estimate rejects the Duflo (2001) range at $p < 0.08$.

¹⁶Switching from the minimal to the intermediate control set as in Duflo (2001, Figure 1)—adding interactions between the baseline enrollment rate and birth year dummies—yields a kink estimate of 0.054 (0.075) with clustering and 0.054 (0.063) without. The latter result corresponds most directly to the original figure, but incorporates data corrections. This test provides little evidence of a trend break.

Table 2: Reduced-form TWFE impact estimates for schooling, 1995

	Original	New		
	(1)	(2)	(3)	Hausman p : weighted vs. not
<i>A. With controls based on number of children aged 5–14</i>				
Experiment	0.12 (0.025)	0.12 (0.042)	0.077 (0.054)	0.23
Placebo	0.015 (0.042)	0.013 (0.055)	0.083 (0.068)	0.08
Kink		0.058 (0.071)	0.040 (0.087)	0.73
<i>B. Without extra controls</i>				
Experiment	0.10 (0.023)	0.100 (0.041)	0.075 (0.050)	0.45
Placebo	0.058 (0.038)	0.063 (0.052)	0.11 (0.068)	0.25
Kink		−0.016 (0.064)	−0.027 (0.082)	0.83
Data corrections		✓	✓	
Clustered		✓	✓	
Weighted			✓	
Experiment N	78,470	78,470	78,470	
Placebo N	78,488	78,488	78,488	
Kink N		143,143	143,143	

Notes: Coefficient estimates are for $D_j T_t$ where D_j is planned Inpres schools per 1,000 children and T_t is a dummy for the younger cohorts in each experiment. Hausman p is for the Hausman test of the unweighted estimate against the weighted, using a 9,999-replication wild cluster bootstrapped estimate of the covariance matrix for the difference between the two. In the last regression, weights are winsorized to the median weight plus 4 times the interquartile range. Experiment samples are restricted to men aged 2–6 or 12–17 in 1974 and placebo samples to ages 12–17 and 18–24. “Original” results without extra controls are obtained by modifying regressions that exactly replicate the original. All regressions control for survey year, year-of-birth, and regency-of-birth dummies. Standard errors are in parentheses—classical in the original estimates and clustered by birth regency in the new. Since the average age difference between treatment and control is 10.5 years in the experiment and 6.5 in the placebo test, for comparability, the placebo results are scaled by 10.5/6.5. Kink estimates are coefficients on the kink term $D_j \times \max(0, 12 - t)$, multiplied by 8, which is the average lag from the allowed moment of kink, age 12 in 1974, to the treatment group aged 2–6 in 1974.

3.2.2 Reduced-form labor impacts

We turn next to reduced-form impacts on the labor outcomes—namely, employment (wage sector participation) and the logarithm of the hourly wage. Although the original does not report reduced-form employment estimates, they are covered here. The analysis proceeds exactly as before: see Table 3 and Figure 2. With one striking exception, the story is similar. The original reduced-form wage results are exactly reproduced. Clustering and weighting reduce precision. Dropping the extra controls increases precision modestly, even as it reduces the kink-based estimates.

For employment, most impact estimates are positive and reasonably significant. The estimate from the specification closest to the original, 2.1 (0.44) points of employment per unit of treatment, becomes 1.7 (0.63) when weighting and 1.2 (0.60) when also dropping the extra controls. The kink estimates are similar without weights, but shrink with weighting. The Hausman test tends to reject the null that the unweighted estimators are consistent. (See columns 1 and 2 of Table 3 and Figure 2).

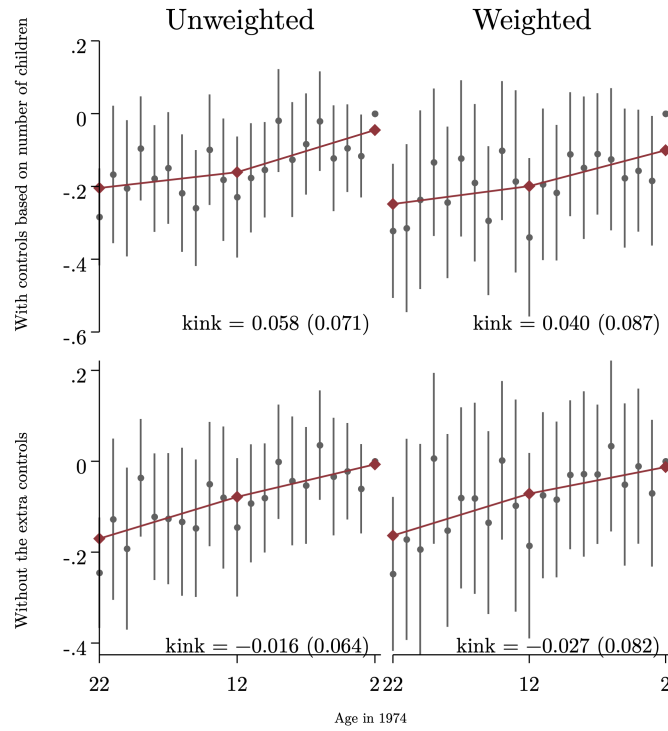


Figure 1: Coefficients on interactions between age dummies and Inpres intensity in schooling regressions, 1995

Notes: Each plot depicts, in grey, the coefficient estimates and 95% confidence intervals for interactions between birth year cohort dummies and the treatment intensity indicator, in regressions reported in Table 2, panel A. Age 2 in 1974 is the omitted reference cohort. Controls based on child population are products of the elementary school-age population in regency of birth and age cohort dummies. Kink model fits are superimposed.

The “striking exception” is that the preferred specification generates a large impact estimate for wages. On the one hand, dropping the child population–based controls from the wage specification substantially alters the original (non-kink) impact estimates, from 0.015 (0.0072) log points per unit of treatment to –0.014 (0.0063) (Table 3, column 3). This confirms a statement in the Duflo (2000, note 21) working paper that the controls are important for the wage equation. On the other hand, the pre-trend is negative in all four variants, as seen in the right half of Figure 2. Against that benchmark, even a relatively flat post-treatment trend in the event study constitutes a large, positive trend break. And the kink is about twice as large with weighting, so that the Hausman test casts doubt on the unweighted estimators. With weights and without the extra controls (as preferred), the kink-based impact estimate is 0.057 (0.024) log points per unit of treatment. The point estimate is 2–4 times the 1.5–2.7% reported in Duflo (2001), and the standard error is small enough to easily reject the lower end of the Duflo (2001) range.

The finding of a strong trend break timed to Inpres, in wages but not schooling, poses a puzzle. It increases the credence of the proposition that Inpres *did* raise wages via schooling. Perhaps the impact on total schooling was too small to be clearly detected in the kink-based regressions preferred above. Or perhaps Inpres SD’s effects will show more clearly in rates of completion of primary school than in total years of schooling—a possibility that we will investigate in section 5.

Table 3: Reduced-form TWFE impact estimates for labor outcomes, 1995

	Formal sector employment			Log hourly wage			
	New		Hausman <i>p</i> : weighted vs. not	Original	New		Hausman <i>p</i> : weighted vs. not
	(1)	(2)		(3)	(4)	(5)	
<i>A. With controls based on number of children aged 5–14</i>							
Experiment	0.021 (0.0044)	0.017 (0.0063)	0.34	0.015 (0.0072)	0.019 (0.011)	0.018 (0.016)	0.95
Placebo	0.0011 (0.0060)	0.0079 (0.0078)	0.19	0.0051 (0.013)	−0.0046 (0.013)	−0.0065 (0.018)	0.88
Kink	0.022 (0.0079)	0.011 (0.011)	0.15		0.035 (0.018)	0.065 (0.026)	0.06
<i>B. Without extra controls</i>							
Experiment	0.013 (0.0044)	0.012 (0.0060)	0.74	−0.014 (0.0063)	−0.014 (0.0096)	−0.011 (0.014)	0.68
Placebo	0.0014 (0.0057)	0.0053 (0.0074)	0.41	−0.0070 (0.011)	−0.017 (0.012)	−0.023 (0.017)	0.63
Kink	0.015 (0.0082)	0.0059 (0.010)	0.14		0.032 (0.016)	0.057 (0.024)	0.08
Data corrections	✓	✓			✓	✓	
Clustered	✓	✓			✓	✓	
Weighted		✓				✓	
Experiment <i>N</i>	78,470	78,470		31,061	31,061	31,061	
Placebo <i>N</i>	78,488	78,488		30,458	30,458	30,458	
Kink <i>N</i>	143,143	143,143			57,234	57,234	

Notes: See notes to Table 2.

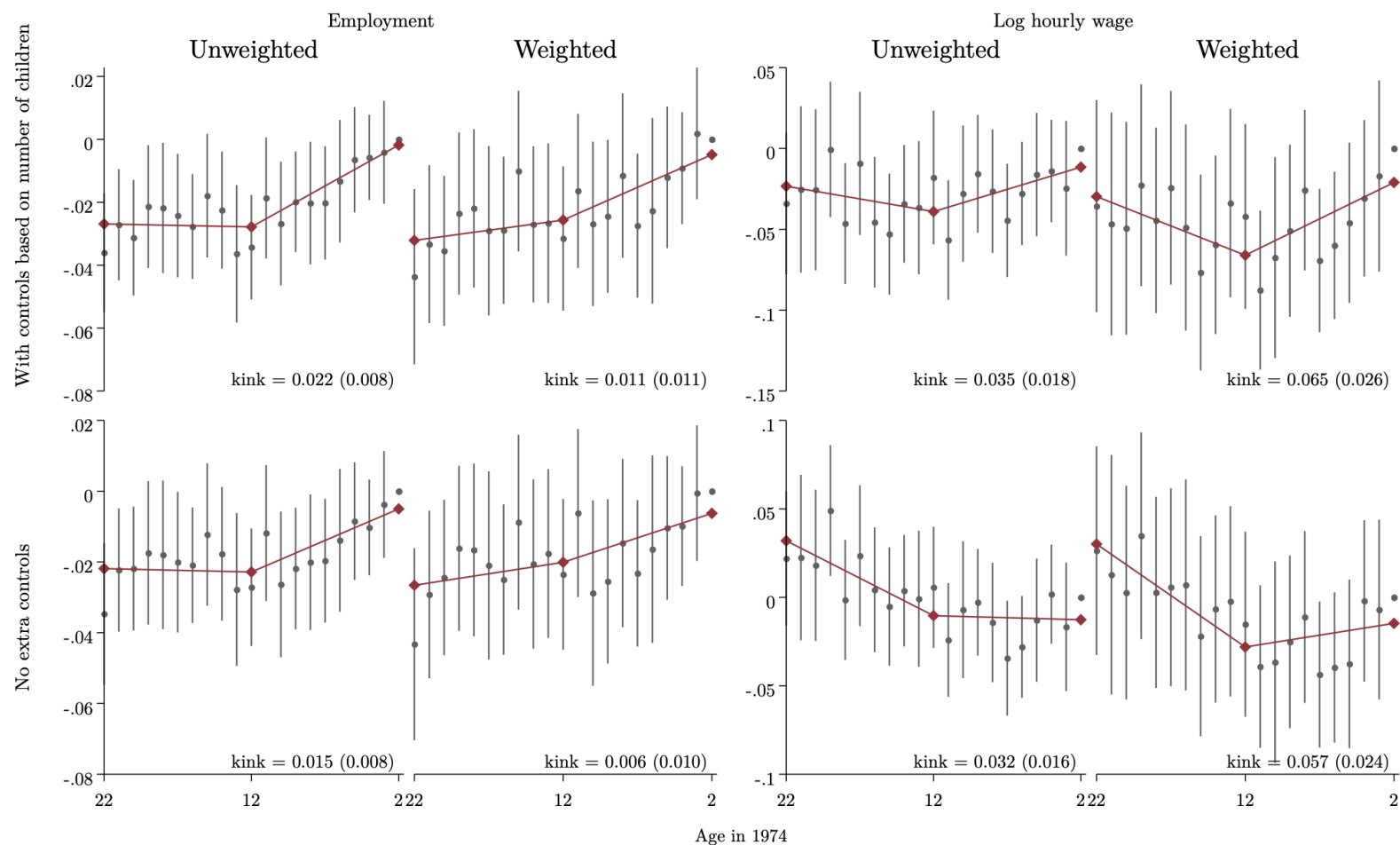


Figure 2: Coefficients on interactions between age dummies and Inpres intensity in labor outcome regressions, 1995

Notes: See notes for Figure 1.

3.3 2SLS estimates of the returns to schooling

I finish the reanalysis of the 1995 follow-up with 2SLS estimates of the returns to schooling. As noted, the preceding results are compatible with the hypothesis that Inpres SD boosted schooling, at least modestly, and thereby wages. However, the weakness of the schooling impacts may make it difficult to measure the return to schooling.

As explained in sections 1 and 2.7, three instrument sets of the form $D_j \tilde{T}_t$ are tried, where D_j is treatment intensity in regency j . The first two sets are from Duflo (2001): the exactly identified “by young/old” specification, in which $\tilde{T}_t$ consists solely of a dummy for the group aged 2–6 in 1974; and the overidentified “by birth year” specification, with one dummy for those aged 12–24 and separate dummies for each yearly age in 2–11. The second appears preferred in Duflo (2001) because results from the first are reported only once. The third variant is new. It incorporates the kink model. $D_j t$ is added as a control and $D_j \times \max(0, 12 - t)$ becomes the instrument capturing slope change at the allowed kink point.

Once again, the original results are exactly reproduced. And once again, I correct baseline variables, cluster standard errors by birth regency, and, in various combinations, incorporate survey weights and drop the controls based on child population. I also report the Kleibergen-Paap (KP, 2006) measure of identification strength, which was not available when Duflo (2001) was written.¹⁷ Results for the “by young/old” and kink instrument set appear in Table 4. The estimates with by-birth-year instruments are in appendix Table C–1.

Instrument weakness proves prevalent. Only the specifications closest to the original (unweighted, with child population controls, using the original’s less-preferred “by young/old” instrument), earn KP F statistics above 7 (Table 4, panel A, columns 1 and 6), where 10 is a rule-of-thumb minimum (Andrews, Stock, and Sun 2019). The KP statistics for the by-birth-year instrument set never clear 2, which is why, for parsimony, they are moved to appendix C.¹⁸ For the kink specifications (columns of 3, 4, 8, and 9 of Table 4), identification is similarly weak, presumably because measuring a trend change is more demanding than measuring a trend.

As mentioned in section 2.5, recent scholarship argues that filtering IV estimates on the basis of a pre-test introduces bias. Instead of putting all inferential weight on the results with, say, KP $F > 5$, I engage in weak identification–robust inference for all of the estimates. To this end, Table 4 reports 95% confidence sets constructed using the wild-bootstrapped Anderson-Rubin test. Most of these confidence sets are disjoint or unbounded. Still, a confidence interval of $(-\infty, \infty)$ does not mean that the regression extracts no information. To provide insight, Figure 3 displays the associated confidence curves. These are functions of trial values for the return to schooling. For a given trial value, such as a return to schooling of 10%, the test statistic’s distribution is

¹⁷The test of Montiel Olea and Pflueger (2013) is in principle superior here because it is robust to heteroskedasticity and within-cluster dependence. However, it coincides with the KP F in the exactly identified regressions that emerge as preferred here, produces comparably low values for the overidentified regressions, and is less practical because the available software is less well developed.

¹⁸It is counterintuitive that adding instruments weakens their collective strength. The cause of the paradox is that identification strength is, broadly, the ratio of two competing factors: the ability of the instruments to explain the instrumented variables, and their expected ability to explain the first-stage error, i.e., the endogenous component of the explanatory variables. The first is desirable; the second is not, since it generates bias toward OLS. But, mechanically, the second also rises with the instrument count. Evidently, moving to the larger instrument set is increasing the undesirable factor more.

simulated using the Wild Restricted Efficient bootstrap of Davidson and MacKinnon (2010).¹⁹ The two-sided p value extracted from the bootstrap is then plotted against the trial coefficient. The curve's intersections with $p = 0.05$ are pinpointed in order to demarcate the confidence sets. The plots are arrayed to parallel the presentation in Table 4.

Though the 95% confidence intervals are often unbounded, the original Duflo (2001) estimates sometimes remain recognizable within them. When the controls based on child population are included, the test assigns lower p values to negative rates of return in both employment and wages (upper half of the figure).

Once more, reanalysis of the 1995 data produces ambiguity. The 2SLS results imply that returns to schooling are positive, but are not much more precise than that. In particular, they weaken the confidence we can have in the Duflo (2001) estimates of the returns to schooling, 6.8–10.6%. The cluster-robust 95% confidence interval for the coefficient in the specification closest to the original, at the top of column 6 of Table 4, is $[-0.03, 0.32]$. While that rejects the null of no return to schooling at a low p value, it does so on the basis of some identifying assumptions that are plausibly violated, including that eschewing survey weights does not cause bias, and that ambient trends are parallel. Both assumptions are challenged by earlier results: the Hausman tests for the first stage (Table 2, final column) lean against the equivalency of the unweighted and weighted estimators, while the kink-based wage analysis has shown that pre-trends are in general not negligible. Confidence intervals from the regressions that relax those assumptions (far right of the table and graph) are unbounded.

¹⁹99,999 replications are performed for each test. Auxiliary weights are Rademacher-distributed and drawn by regency. See Roodman et al. (2019).

Table 4: 2SLS estimates of the effect of years of schooling on labor outcomes, 1995

	Employment				Log hourly wage				
	Instrument by young/old		Kink instrument		Instrument by young/old			Kink instrument	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
<i>A. With controls based on number of children aged 5–14</i>									
Coefficient	0.17 (0.065)	0.22 (0.16)	0.35 (0.38)	0.36 (0.71)	0.0752 (0.0338)	0.12 (0.058)	0.12 (0.094)	0.17 (0.11)	0.19 (0.098)
	[0.09, 0.58]	$(-\infty, -0.43] \cup [0.05, \infty)$	$(-\infty, -0.27] \cup [0.07, \infty)$	$(-\infty, \infty)$		[-0.03, 0.32]	[-0.38, 1.04]	$(-\infty, \infty)$	[0.00, $\infty)$
KP <i>F</i>	7.89	2.08	0.92	0.28		7.44	4.29	2.22	4.23
<i>B. Without extra controls</i>									
Coefficient	0.13 (0.063)	0.16 (0.11)	-1.73 (9.97)	-0.38 (0.93)		-0.50 (1.15)	-0.41 (1.38)	1.48 (13.3)	0.43 (0.54)
	[0.04, 0.69]	$(-\infty, -0.24] \cup [-0.00, \infty)$	$(-\infty, -0.07] \cup [0.08, \infty)$	$(-\infty, \infty)$		$(-\infty, 0.07] \cup [0.20, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
KP <i>F</i>	5.96	2.21	0.030	0.21		0.28	0.16	0.011	0.45
Data corrections	✓	✓	✓	✓		✓	✓	✓	✓
Clustered	✓	✓	✓	✓		✓	✓	✓	✓
Weighted		✓		✓			✓		✓
Pre-trend control			✓	✓				✓	✓
Observations	78,470	78,470	152,989	152,989	31,061	31,061	31,061	60,663	60,663

Notes: Each column of each panel shows results for a different dependent variable. Coefficients are for years of schooling. “Young/old” regressions are restricted to ages 2–6, 12–17. The kink specification adds $D_j t$ as a control and instruments with $D_j \times \max(0, 12 - t)$ where t is age in 1974. All regressions control for year-of-birth and regency-of-birth dummies. Standard errors are in parentheses and 95% confidence sets are in brackets. The latter are from the Anderson-Rubin test with a wild-bootstrapped distribution with 99,999 replications. “KP *F*” is the Kleibergen-Paap measure of instrument strength. Column 5 is copied from Duflo (2001), Table 7.

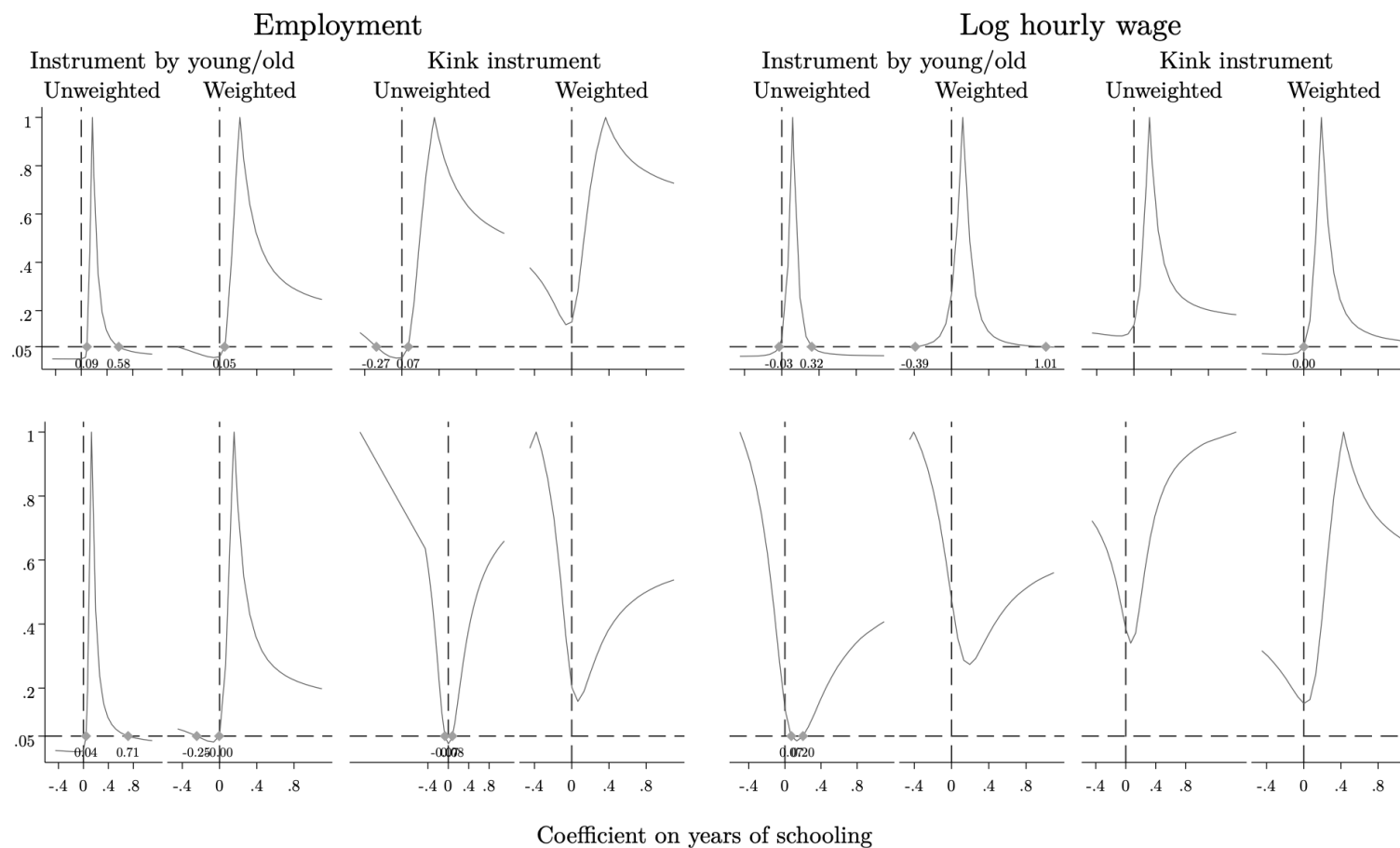


Figure 3: Confidence curves from wild-bootstrapped tests of the coefficient on years of schooling in 2SLS regressions in Table 4

Notes: Each plot is of the p value for the null that the instruments in a particular regression are valid and the coefficient on schooling takes a given trial value, as a function of that trial value. Each test statistic is from an Anderson-Rubin test, whose distribution is simulated using a wild bootstrap. Horizontal dashed lines mark $p = 0.05$. Their intersections with the curves define 95% confidence intervals. The plots are arranged in parallel with the results in Table 4, to which they correspond.

4 Following up later

4.1 The added surveys

The Indonesian statistical agency has long conducted several national surveys, with themes including labor and household welfare. As a result, the passage of time has generated more data about the subjects of the quasi-experiment framed in Duflo (2001). Because schooling attainment hardly changes during adulthood, the added data should help us to better measure the schooling impacts already studied in the 1995 data. In addition, information on labor market outcomes in prime work years is of interest in its own right. In the 1995 follow-up, the main treated cohorts are 2–6 years old in 1974, thus 23–27 at survey, early in their careers.

The choice of follow-ups is constrained by the need to construct key variables in Duflo (2001): regency and year of birth, years of schooling, employment, and wages. The rarest of those is regency of birth. Feasible follow-ups include:

- *The 2005 intercensal survey (SUPAS)*. The successor to the 1995 survey, this one provides all the needed variables except wages.²⁰ It is hosted by the IPUMS project (MPC 2020).
- *Socioeconomic survey (SUSENAS)*. The rounds of 2011–14 and 2017–19 cover schooling. The 2011–14 SUSENAS waves also include standard labor survey questions.²¹

One complication arises in moving to these later surveys: the possibility of selective attrition. This could result from mortality if, for example, subjects from more-treated areas are poorer and die sooner. In addition, historically, the standard retirement age for civil servants in Indonesia was 56. This may have caused those with formal-sector employment to exit the work force earlier. (A common formal profession in the data set is schoolteacher.) Duflo (2001)’s pre-treatment cohort, men aged 12–17 in 1974, started to turn 56 in 2013.

The attrition biases might not be large. In Indonesia in this period only tertiary schooling—not the target of Inpres—is strongly associated with mortality (Sudharsanan et al. 2020; Xu et al. 2024). And the progression of the pre-treatment cohorts into retirement may have been more complex and gradual than the existence of an official retirement age would imply. Government employees could work beyond age 56 under some circumstances (Muliati and Wiener 2014). And civil servants with at least 20 years of service could retire as early as 50. To the extent that attrition from retirement is gradual, it becomes just one more ambient trend that the statistical models are more or less designed to adjust for, e.g., with the linear time trend in the kink model.

This matter can be probed empirically. If the bias is severe, then as pre-treatment cohorts (the 12–24-year-olds of 1974) progress through the surveys of 2011–19, the association *within* this demographic between treatment intensity in place of birth and outcomes later in life will shift as

²⁰Previous versions of this paper imputed wages in 2005 with a model calibrated to the 1995 SUPAS data. This model was inspired by the income scores IPUMS has constructed for U.S. census data to proxy for earnings. But the validity of estimates based on imputed wages could be questioned. In fact, *all* wage indicators contain measurement problems (see Surbakti 1995, p. 44, on issues in wage values in Indonesia’s surveys). Still, I dropped the imputation because I had developed it before I learned about the option of using later SUSENAS data. I probably would never have imputed the 2005 wages if I had known about the later data. Skipping the imputation reduces my exercise of researcher discretion. For much the same reasons, I also dropped the use of 2010 labor survey (SAKERNAS) data, despite the precedent in Duflo (2004a)’s use of the same survey to estimate medium-run impacts of Inpres. The data have the disadvantage of including regency of residence but not regency of birth.

²¹Hsiao (2024) first used the 2011–14 rounds to revisit Duflo (2001).

they age. For example, if more-treated regencies were poorer before Inpres and produced fewer well-paid, formal sector workers, then retirement of such workers could flatten the cross-sectional dependence of wages on treatment intensity. To check, I regress the outcomes of interest on interactions between treatment intensity and survey year dummies. Much as in the main TWFE specifications studied here, birth year, birth regency, and survey year fixed effects are controlled for. Observation weights are again incorporated, while controls based on child population are not.²² Samples are restricted to the pre-treatment 12–24-year-olds. The coefficient estimates are plotted in Figure 4 along with clustered 95% confidence intervals. The last available year for a given outcome is always the omitted reference year, so it is represented with an exact zero. To the extent that the tests are powered to detect selective attrition, they do not find it. Except for employment, the confidence intervals easily embrace zero. And the positive relative trend in employment starts in 1995–2005 and thus is not strongly linked to the end of (work) life.²³

Since attrition bias does not appear to be large, especially in the key schooling and wage variables, I will retain all the feasible follow-ups in the 2010s.

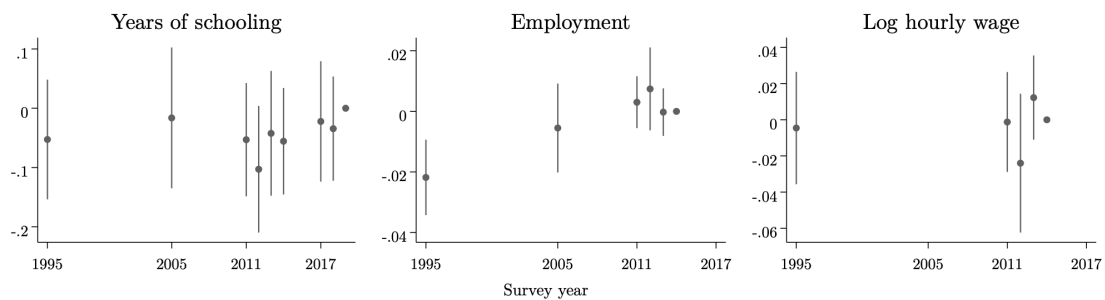


Figure 4: Coefficients on interactions between survey year dummies and Inpres intensity, among 12–17-year-olds of 1974

Notes: Each plot shows coefficients from a single regression, on the products of treatment intensity in regency of birth with survey year dummies. Each dependent variable is available in a distinct subset of survey rounds. Observation weights are incorporated. Samples are restricted to the pre-treatment cohorts aged 12–24 in 1974. All regressions control for birth year, birth regency, and survey year fixed effects. The last available year is always the omitted reference year, so it is represented with an exact 0.

4.2 Results from later follow-ups

To provide an out-of-sample test, the TWFE specifications are transferred from the 1995 follow-up to the later survey waves, pooled. Samples continue to be restricted to men who were 2–24 in 1974. The only specification change is to add survey year fixed effects. As was seen in Figure 4, the wage regressions can now take data from 2011–14. When formal employment is the dependent variable, 2005 is added. And when years of schooling is the dependent variable, the samples extend to the 2017–19 rounds as well.

The new schooling impact estimates are more precise than the old, yet do not differ statistically: compare the post-1995 results, which are in Table 5 and Figure 5, with the 1995 results in

²²Reversing these two choices make no difference for the conclusion.

²³A previous version of this paper stopped follow-up at 2012 because the 12–17-year-olds of 1974, the Duflo (2001) pre-treatment cohorts, began reaching age 56 in 2013. Through the analysis described here, I changed my mind.

Table 2 and Figure 1. This consistency is expected since schooling attainment hardly changes during adulthood. The larger samples provide more information about the same variable in the same population, and attrition does not appear to be large. That said, the kink-based schooling estimates all shift in the same direction: upward. All are now positive. Interestingly, the kink plot closest in specification to Duflo (2001, Figure 1), validates Duflo (2001)’s informal reading of a trend break (upper left of Figure 5 below). According to the new, formal estimate, among those aged 2–6 in 1974, Inpres increased schooling by an average 0.086 (0.040) years per unit of treatment. In the specification preferred here—with weights, without extra controls—the kink is smaller, at 0.038 (0.038). Overall, the higher-powered schooling estimates provide slightly more evidence that Inpres SD boosted schooling attainment.

For the employment outcome, the shift to prime working years turns out to introduce a downward pre-trend, especially when weighting. And low p values on the Hausman test tend to justify weighting. The preferred kink fit implies an impact of 0.93 (0.62) percentage points per unit of treatment. See Table 6 and Figure 6. That is larger and more precise than the corresponding estimate for 1995, 0.59 (1.0) points.

In contrast, the estimated impact on wages, conditional on employment, is smaller in the later follow-ups, though it is still always positive. In the kink specification closest to the original, the estimate falls only slightly, from 0.035 (0.018) log points in 1995 to 0.034 (0.015) in 2011–14. In the preferred specification, the estimate declines from 0.057 (0.024) in 1995 to 0.0088 (0.020) in 2011–14 (comparing the right halves of Table 3 and Table 6, Figure 2 and Figure 6).

Moving to the larger samples does not systematically strengthen identification in the 2SLS regressions. In the wage regressions, the KP F statistic is lower than in 1995 when the population controls are entered and higher than in 1995 when they are not. For example, the confidence interval in the wage specification closest to the original shifts from $[-0.03, 0.32]$ in 1995 (column 6 of Table 4) to $[0.11, \infty)$ in 2011–14 (Table 7). That for the preferred specification narrows from $(-\infty, \infty)$ to $[-0.23, 0.14]$ (bottom right of Table 7 and Figure 6). Regardless, the returns to schooling remain imprecisely measured.

Appendix D repeats the analysis of this section with the 1995 data included as well. Since the samples are dominated by the new, larger follow-ups, the results are similar.

The overall picture from the new follow-ups resembles that from the 1995 data, but is a bit sharper. In all variants, the schooling, employment, and wage trends now kink upward at the allowed moment. For schooling and wages, kinks are statistically significant when the controls based on child population are included. This means that one’s reading of the OLS evidence depends on how much weight one puts on the inclusion of controls whose presence in the original is mostly not explained. Meanwhile, the continuing weakness of the first stage leaves the returns to schooling poorly identified. It remains the case that causation from Inpres to schooling to wages is compatible with the evidence. And it remains the case that the evidence does not emphatically ratify the theory.

Table 5: Reduced-form TWFE impact estimates for years of schooling, pooled data for 2005, 2011–2014, 2017–19

	(1)	(2)	Hausman <i>p</i> : weighted vs. not
<i>A. With controls based on number of children aged 5–14</i>			
Experiment	0.10 (0.032)	0.099 (0.036)	0.86
Placebo	0.050 (0.041)	0.10 (0.046)	0.05
Kink	0.086 (0.040)	0.069 (0.041)	0.55
<i>B. Without extra controls</i>			
Experiment	0.076 (0.031)	0.079 (0.035)	0.88
Placebo	0.054 (0.038)	0.094 (0.043)	0.12
Kink	0.043 (0.038)	0.038 (0.038)	0.85
Data corrections	✓	✓	
Clustered	✓	✓	
Weighted		✓	
Experiment <i>N</i>	589,671	589,671	
Placebo <i>N</i>	486,230	486,230	
Kink <i>N</i>	1,023,726	1,023,726	

Notes: Notes to Table 2 apply.

Table 6: Reduced-form TWFE impact estimates for labor outcomes, 2005, 2011–14

	Formal sector employment			Log hourly wage		
	(1)	(2)	Hausman <i>p</i> : weighted vs. not	(3)	(4)	Hausman <i>p</i> : weighted vs. not
<i>A. With controls based on number of children aged 5–14</i>						
Experiment	0.0020 (0.0026)	–0.0043 (0.0032)	0.003	0.022 (0.0088)	0.0091 (0.011)	0.05
Placebo	–0.0056 (0.0050)	–0.025 (0.0068)	0.000005	–0.034 (0.014)	–0.024 (0.022)	0.49
Kink	0.0081 (0.0055)	0.016 (0.0067)	0.08	0.034 (0.015)	0.034 (0.022)	0.99
<i>B. Without extra controls</i>						
Experiment	–0.0026 (0.0025)	–0.0077 (0.0029)	0.01	–0.0061 (0.0087)	–0.018 (0.0099)	0.08
Placebo	–0.0062 (0.0047)	–0.021 (0.0065)	0.0001	–0.025 (0.013)	–0.015 (0.018)	0.47
Kink	0.0047 (0.0051)	0.0093 (0.0062)	0.25	0.013 (0.014)	0.0088 (0.020)	0.70
Data corrections	✓	✓		✓	✓	
Clustered	✓	✓		✓	✓	
Weighted		✓			✓	
Experiment <i>N</i>	382,383	382,383		113,738	113,738	
Placebo <i>N</i>	325,963	325,963		68,938	68,938	
Kink <i>N</i>	667,593	667,593		184,723	184,723	

Notes: See notes to Table 3. The 2005 data covers employment but not wages.

Table 7: 2SLS estimates of the effect of years of schooling on labor market outcomes, pooled data for 2011–14

	Employment				Log hourly wage			
	Instrument by young/old		Kink instrument		Instrument by young/old		Kink instrument	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>A. With controls based on number of children aged 5–14</i>								
Coefficient	0.023 (0.028) [−0.06, 0.11]	−0.055 (0.052) (−∞, 0.03]	0.13 (0.13) (−∞, ∞)	3.12 (34.2) (−∞, 0.00] ∪ [0.02, ∞)	0.31 (0.15) [0.11, ∞)	0.32 (0.55) (−∞, ∞)	−0.44 (0.71) (−∞, −0.01] ∪ [0.30, ∞)	−0.11 (0.13) (−∞, 0.06]
KP <i>F</i>	7.15	4.33	1.63	0.0082	2.56	0.22	0.56	3.72
<i>B. Without extra controls</i>								
Coefficient	−0.043 (0.051) (−∞, 0.03] ∪ [0.93, ∞)	−0.11 (0.077) (−∞, −0.02]	0.33 (1.50) (−∞, ∞)	−0.27 (0.69) (−∞, ∞)	0.25 (0.35) (−∞, ∞)	0.27 (0.18) (−∞, ∞)	−0.024 (0.085) (−∞, 0.11]	0.022 (0.061) [−0.23, 0.14]
KP <i>F</i>	3.70	3.75	0.051	0.23	0.30	1.45	4.01	8.37
Data corrections	✓	✓	✓	✓	✓	✓	✓	✓
Clustered	✓	✓	✓	✓	✓	✓	✓	✓
Weighted		✓		✓		✓		✓
Pre-trend control			✓	✓			✓	✓
Observations	382,383	382,383	705,366	705,366	113,297	113,297	188,949	188,949

Notes: See notes to Table 4 and Table 5.

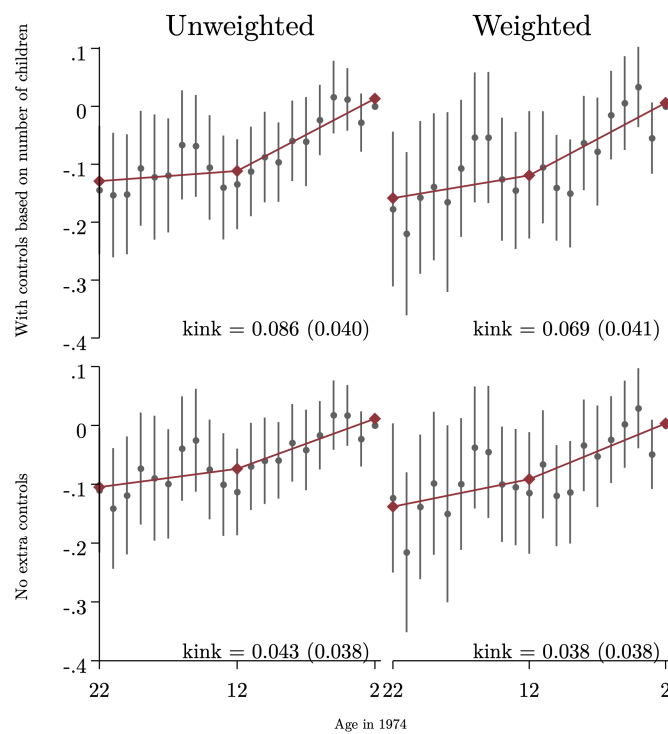


Figure 5: Coefficients on interactions between age dummies and Inpres intensity in schooling regressions, 2005, 2011–14, 2017–19

Notes: See notes to Figure 1.

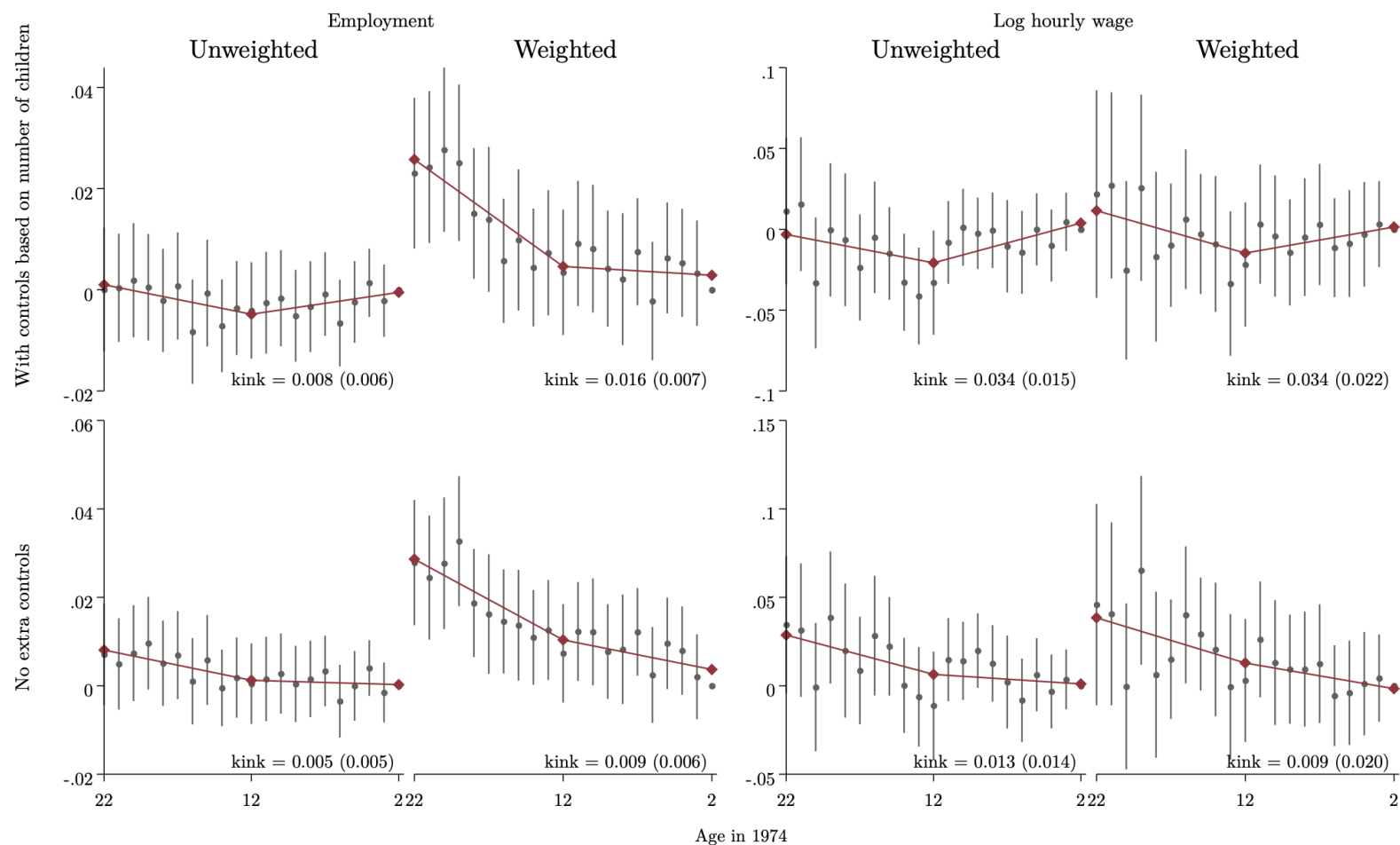


Figure 6: Coefficients on interactions between age dummies and Inpres intensity in labor outcome regressions, 2005, 2011–14

Notes: See notes to Figure 1. The 2005 data cover employment but not wages.

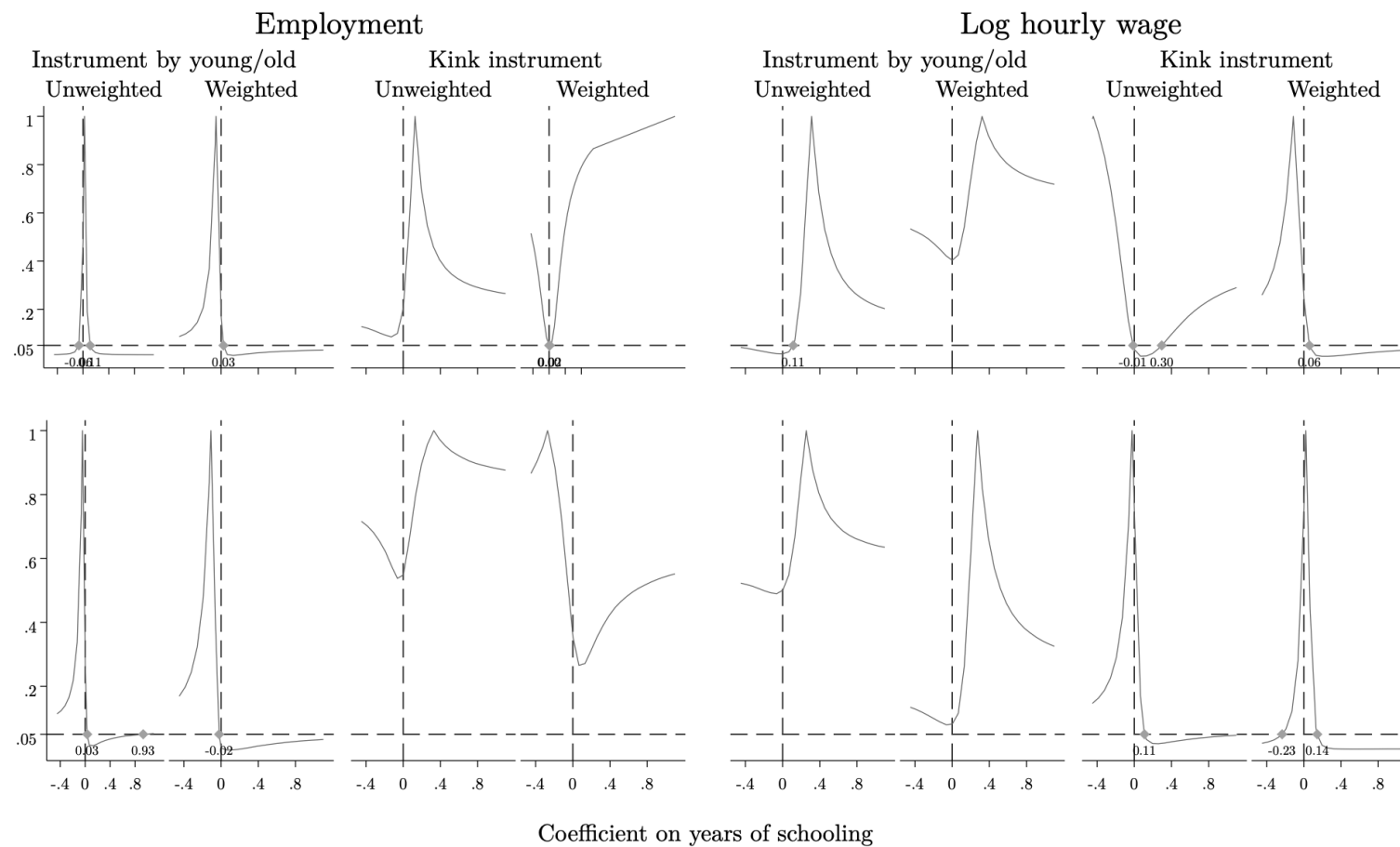


Figure 7: Confidence curves from wild-bootstrapped tests of the coefficient on schooling in 2SLS regressions in Table 7, 2005, 2011–14
 Notes: See notes to Figure 3 and Table 7.

5 Impacts by grade level

Recall from section 1 that Duflo (2001) disaggregates the main specifications along two dimensions to check for fingerprints of Inpres SD. The analysis so far has heavily emphasized the kink model, which breaks out results by birth cohort. This section turns to the other fingerprint test, for the grade-by-grade impacts of the school construction impulse, which Duflo (2001) calls DID-in-CDF analysis. The examination will corroborate the view that Inpres affected primary schooling patterns, but in a surprisingly circumscribed way. Inpres seems to have increased primary school completion, but not so much total *years* spent in primary school. A potential confounder will also be exposed that further contributes to the seeming muted impact on total schooling.

This fingerprint check consists in running a set of regressions like (2)—meaning without any controls based on child population—but with the treatment variable dichotomized as in 2×2 DID. The dependent variables are dummies for whether a person has at least 1 year of schooling, or 2, or 3, and so on. The impact estimates for these indicators are plotted along with 95% confidence intervals.

Duflo (2001) Figure 2 is almost exactly replicated in the upper left of Figure 8, below.²⁴ The impact of Inpres on schooling attainment builds through the six years of primary school. The effect mostly evaporates in the three years of lower secondary school (grades 7–9). These results look like a fingerprint of Inpres. Then, in upper secondary school (grades 10–12) as well as the early tertiary years, the “impact” goes negative. “The negative difference in differences at the senior high school level,” Duflo (2001) reasonably surmises, “may indicate that some variable predicting the probability of attending senior high school is omitted from this regression (and changed in low program regions more than in high program regions).”

I vary this replication in the same ways as before. In the second row, I make the treatment variable continuous (planned schools per 1,000 children), incorporate data corrections, cluster standard errors, weight observations, and incorporate all feasible surveys (1995, 2005, 2011–14, 2017–19). Other than a modest shift toward zero in the lower secondary years, the pattern of results hardly changes.

Then, in the right column, I transfer these specifications to the placebo experiment. As before, I scale these estimates by 10.5/6.5 for comparability. Notably, and especially in the larger sample represented in the second row, the experiment and placebo patterns are similar. This suggests that most of the grade-wise fingerprint detected in Duflo (2001) is characteristic of long-term convergence, not Inpres specifically. However, some of the placebo effects in primary school are smaller, leaving open the possibility—to which we will return—that Inpres *accelerated* convergence. Meanwhile, the negative “impact” on secondary schooling is largely absent in the placebo experiment. The omitted variable detected by Duflo (2001) apparently only affects the younger cohorts.

What might the omitted variable be? A good candidate is Indonesia’s contemporaneous expansion in *secondary* schooling. Under President Suharto, Indonesia managed development policy through

²⁴The code Duflo sent me does not produce this figure.

five-year plans (Repelitas). One reason the government so quickly translated the oil windfall of late 1973 into the Inpres SD program is that Repelita II was nearly finalized. Much more so than Repelita I, it prioritized education, and at all levels. Heneveld (1979, p. 147) describes the scale of state investment in secondary education as “staggering”: nearly 1,000 new schools built, another 2,000 rehabilitated, more than 25,000 teachers hired. The growth continued under Repelita III. The annual number of middle school graduates doubled between 1974 and 1979, from 293,000 to 595,000, then doubled again by 1984, to 1.25 million. High school graduating classes mushroomed too, from 78,000 to 154,000 to 456,000 (IEES 1986, pp. 6, 8). In contrast with primary school students, the majority of new secondary students went to *private* schools, which indicates that the supply expansion, however great, was outstripped by demand. And the demand was not evenly distributed. “The most rapid increases in demand have occurred precisely in those areas which are profiting most from Indonesia’s new economic growth, the cities, and in East Java” (Heneveld 1978, p. 74). Since Javanese regencies had higher populations, they tended to receive fewer Inpres schools per child. The intensity of the primary and secondary “big pushes” may have been anti-correlated across regencies.

Likely, then, the effects of the Inpres SD primary school expansion on *total* schooling are partially confounded by a secondary schooling expansion. This might help explain why an impact on total schooling trends has not emerged consistently in the data.

To sharpen the picture of the schooling impacts, the third row of Figure 8 introduces another methodological change: each regression for attending a grade is restricted to those who attended the previous grade, making the dependent variables dummies for *continuation*. Since the samples are small for the highest schooling levels, their confidence intervals are wide, so they are dropped to avoid compressing the vertical scale. In the primary years, we see positive impacts at each grade. The largest is in grade 1—that is, in making it practical for children to attend elementary school at all. However, the same effect is seen in the placebo group. Meanwhile, at the secondary level, the negative results are now concentrated in grades 7 and 10, i.e., on whether students continue to the next schooling level.

To systematize the experiment-placebo comparisons, the bottom row of the figure brings the kink model to these finely sliced schooling variables, once more allowing a trend break at age 12 as of 1974. The plot of kink estimates in the bottom left shows no clear trend break for continuation into grade 7 but does show large negative kinks at grades 10 and 13. These imply that the relative slow-down in continuation into lower secondary school in high-Inpres areas did not break with the historical trends, while the slow-downs in continuation into upper secondary and university did.

Somewhat surprisingly, within the primary school years, by far the largest deviation from historical trend is in continuation into 6th grade, the final year of primary school. One cause might be an inappropriate choice of kink point for the lower grades. The reason to expect a kink in grade 6 continuation at age 12 as of 1974 is the same as for total schooling: the 11-year-olds of 1974 were the first cohort to be able to attend the new schools—which they would only have been able to do for one year, in grade 6. Likewise we might expect a kink in continuation into grade 5 to appear one cohort later, and so on down to a kink in grade 1 continuation at the cohort aged 7 in 1974. Figure 9 investigates—and mostly rejects—this theory. The kink plots for continuation at each grade level are constructed in the usual way except that the spline is shifted

in the way just described. The kink in grade 6 continuation remains far larger than the others, which are themselves diverse in sign.

In sum, the trend break in schooling attributable to Inpres SD is largely confined to continuation into the final year of primary school. This effect may be real, or it may reflect measurement noise in the lower grades. Perhaps survey respondents accurately recall whether they completed primary school but, conditional on not completing it, do not accurately recall when they left school.

The apparent concentration of continuation effects in the final primary grade and the contrary developments at the secondary and even tertiary levels may both mask the impact of Inpres on schooling careers in the main Duflo (2001) regressions. If so, using a narrowed schooling variable might improve identification. There are two natural candidates: total *primary* schooling, capped at 6 years; and a dummy for primary completion, more precisely for whether a person attended grade 6 or higher. Figure 10 studies the impact of Inpres on these variables using kink plots. Total primary schooling shows remarkably little trend break upon the arrival of Inpres, but primary completion does, making the latter a more promising candidate for kink-based 2SLS estimation of the returns to schooling. Appendix E reports 2SLS estimates of the impact of primary completion on employment outcomes in 2011–14. Once more, the estimates are imprecise.

Bringing the specification changes developed in previous sections to the Duflo (2001) DID-in-CDF analysis reveals a sort of two-dimensional fingerprint of Inpres, along the dimensions of schooling grade and birth cohort. The results support the view that Inpres SD accelerated convergence in schooling between low- and high-treatment areas, at least in primary completion. Also exposed is the likely fingerprint of simultaneous post-primary schooling expansion. Unfortunately, removing this confounding variation does not sharpen the estimates of returns to schooling.

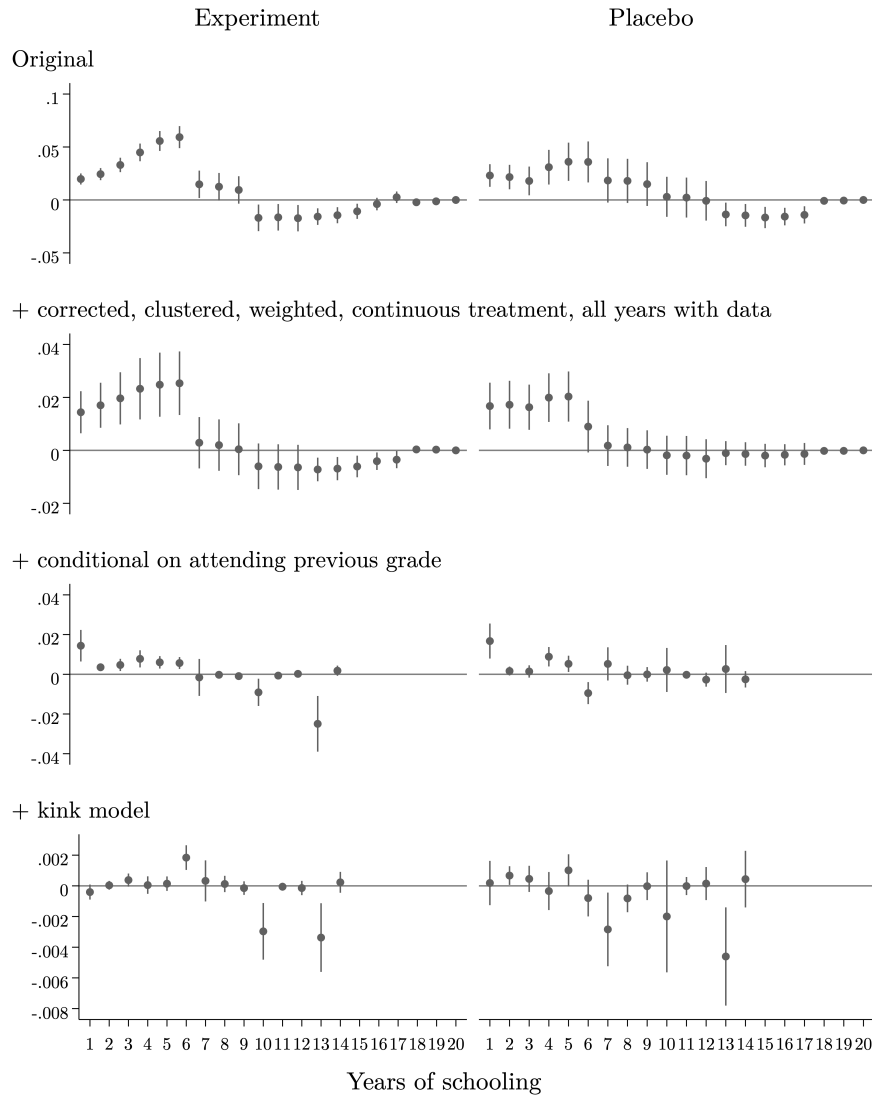


Figure 8: TWFE estimates for schooling attainment and continuation

Notes: Though not marked, the vertical axis ranges are the same within each row. Each point estimate and 95% confidence interval is from a separate TWFE regression, incorporating fixed effects for year and regency of birth. No other controls are added, except the one included by definition in the kink model. The dependent variables are dummies for whether a person attended school for $\geq m$ years, $m = 1, \dots, 20$. The upper left plot is a close replication of Duflo (2001), Figure 2.

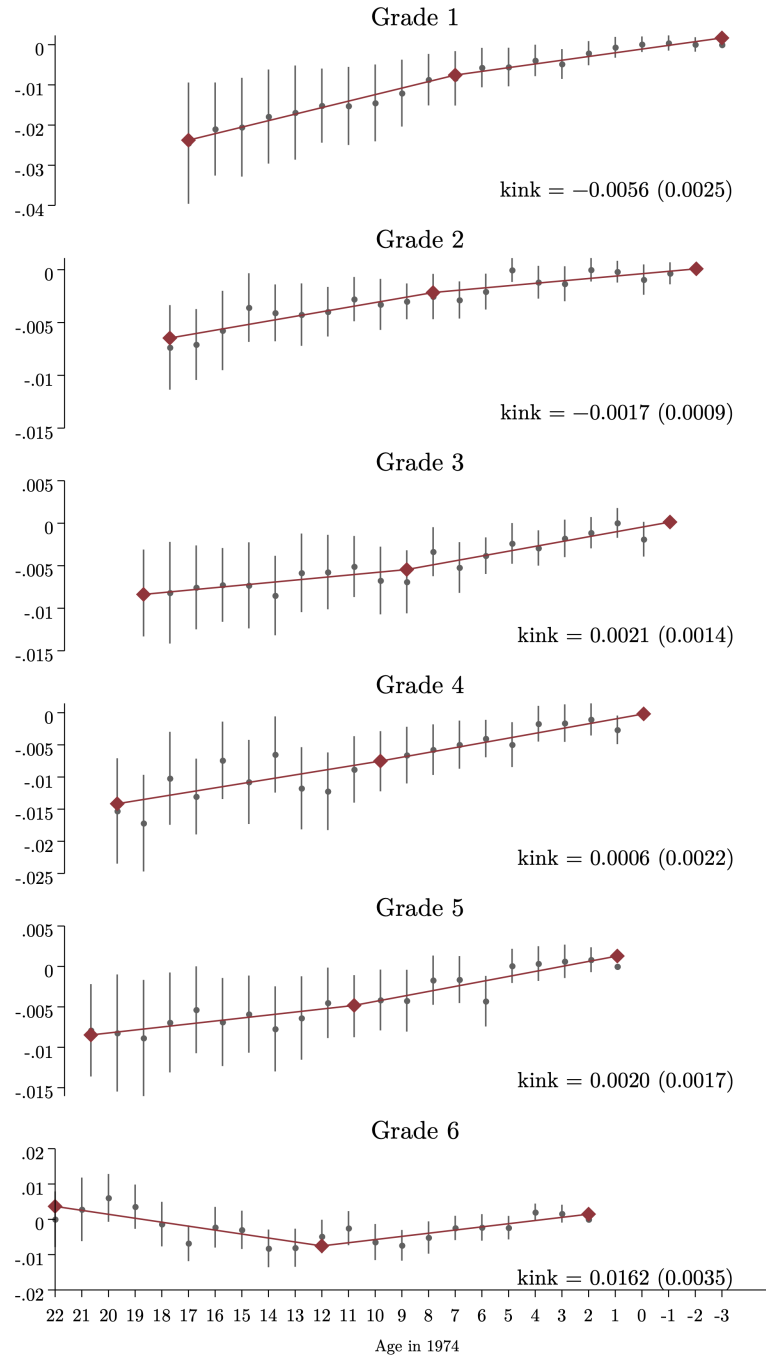


Figure 9: Kink estimates for schooling continuation rates, all available survey rounds

Notes: The plots are constructed in the same way as in earlier figures except that the spline is shifted as shown. The dependent variable is attendance of a given grade conditional on having attended the previous grade. No controls based on child population are included.

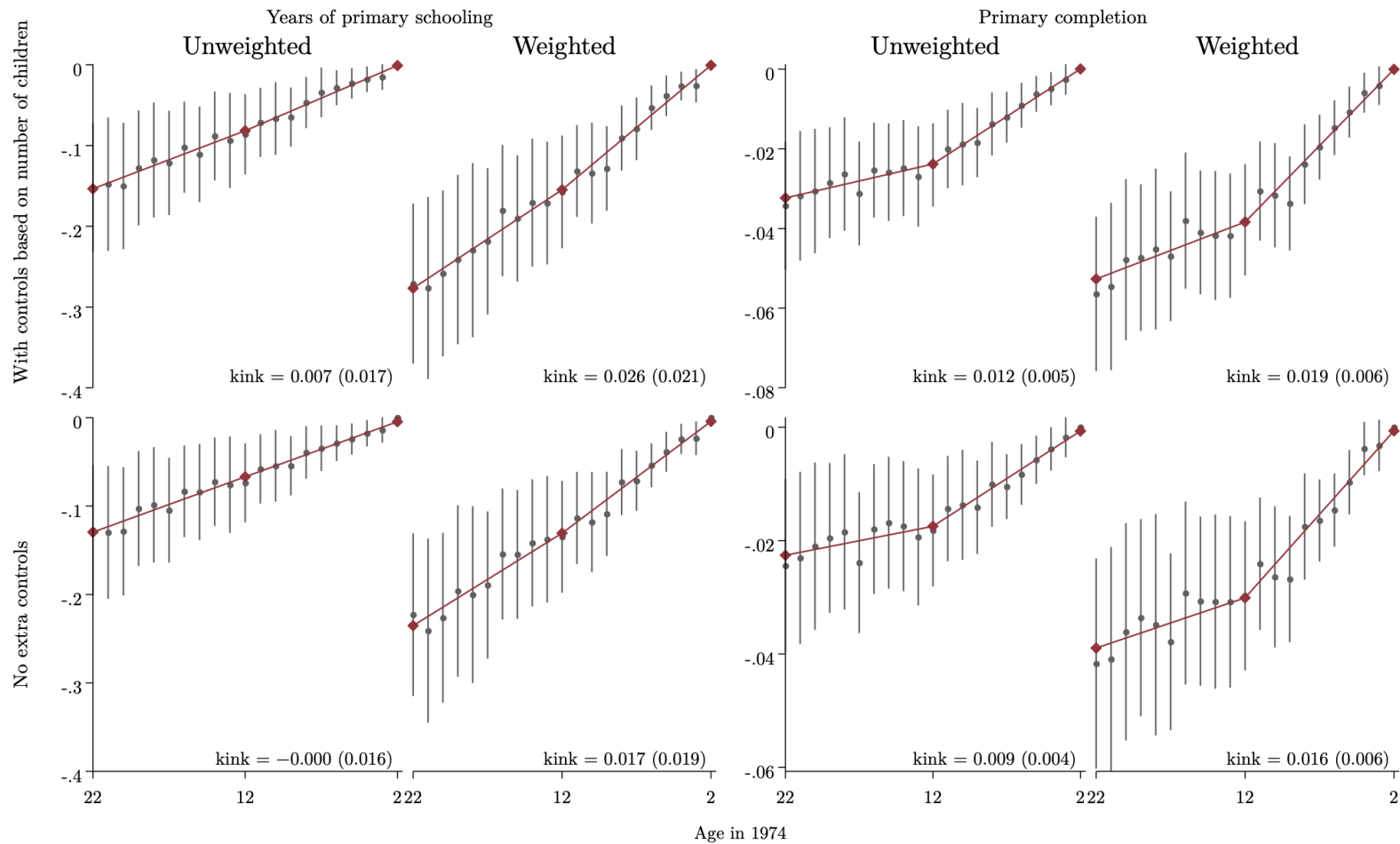


Figure 10: Kink estimates for primary schooling and primary completion, all available survey rounds

Notes: See notes to Figure 1. Data are for 1995, 2005, 2011–14, 2017–19. No controls based on child population are included.

6 *Changes in changes*

An alternative to controlling for nonparallel trends, as with the kink model, is to apply an estimator that is intrinsically robust to their presence. A leading example is the quantile-based changes-in-changes (CIC) estimator of Athey and Imbens (2006). Appendix F reports results from applying CIC to reduced-form wage impacts. Despite the looser identifying assumptions, the standard errors are similar to those from the kink-based estimates in Table 3 and Table 6. The strongest finding is of positive wages impacts at low deciles during the prime earnings years (2011–14). This progressive impact (in the distributional sense) is consistent with causation by Inpres, which would primarily have lengthened the schooling careers of poorer people, for whom the marginal year was more often in primary school. There is a hint of negative impacts at high deciles, which might reflect the confounding from post-primary schooling expansions. Appendix G explains why the “Wald CIC” estimator of de Chaisemartin and D’Haultfœuille (2017) is not deployed to estimate the returns to schooling.

7 *Conclusion*

The scrutiny here brought to Duflo (2001) is a testament to its enduring importance. Few studies so credibly claim to identify the impacts of schooling at scale. Most of the comments assess the robustness of the results to relaxing identifying assumptions. And many stem from the passage of time, which has brought more data and econometric insight. Four changes are relatively uncontroversial: correcting baseline data, clustering standard errors by geographic treatment unit, following up later, and employing weak identification–robust inference. Others are best seen as matters of judgment, making tradeoffs that are typically clear in principle and hard to measure in practice. In various combinations, the latter produce an ambiguous set of results, which frustrates the desire for a clean and definitive narrative.

Overall, Inpres appears to have accelerated progress on primary completion, at a rate of about 1.6 percentage points per school per 1,000 children. The impacts on total years in school, and even just years in primary school, are remarkably muted, which might owe to measurement error in primary schooling and confounding from secondary schooling. More surely, though still not definitively, the onset of Inpres is associated with wage gains among natives of higher-treatment areas. The ratio of these two effects—the returns to schooling—proves impossible to measure with precision.

On the strength of this reanalysis, I can form conclusions on three subjects: the interpretive reasoning in Duflo (2001); the strength of the evidence, as now expanded; and the credibility of a particular theory to explain the evidence. As for the first, the more demanding assessment performed here does not support the confident conclusion in the original that Inpres SD “was effective in increasing both education and wages in Indonesia.” As for the second, if the definition of strong evidence is the ability to overcome a judiciously skeptical prior, then the evidence here is weak. If a reader *does* come to this reanalysis with a prior that Inpres most likely boosted schooling and thereby wages, the evidence reported here will mildly reinforce that prior. A more skeptical reader will hardly be moved.

Finally, as for theory, I do not endorse a prior that a massive schooling expansion had next to no impact on schooling. But I do not have a strong prior about the return to primary schooling in

the Duflo (2001) quasi-experiment. Spending five years in primary school instead of four may not increase pay in a future career of manual labor. I take the upward, Inpres-timed kinks in trends in primary completion and wages, with some variants of the estimate more statistically significant than others, as mildly supporting causation from the first to the second. The effect is unfortunately hard to quantify.

I hope that researchers will continue to take inspiration from Duflo (2001) as a creative, rigorous effort to extract evidence from a natural experiment that is relevant to major theoretical questions and global issues. That said, this reanalysis shows that the hunt for causal truth outside of actual experiments is at least as hard as valuable.

8 References

- Acemoglu, Daron, and Joshua Angrist. 2000. "How Large Are Human-Capital Externalities? Evidence from Compulsory Schooling Laws." *NBER Macroeconomics Annual* 15 (January): 9–59. <https://doi.org/10.1086/654403>.
- Anderson, T. W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20 (1): 46–63. <https://jstor.org/stable/2236803>.
- Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11 (1): 727–53. <https://doi.org/10.1146/annurev-economics-080218-025643>.
- Angrist, Joshua D., and Alan B. Krueger. 1999. "Chapter 23 - Empirical Strategies in Labor Economics." In *Handbook of Labor Economics*, edited by Orley C. Ashenfelter and David Card, 3:1277–1366. Elsevier. [https://doi.org/10.1016/S1573-4463\(99\)03004-7](https://doi.org/10.1016/S1573-4463(99)03004-7).
- Athey, Susan, and Guido W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica: Journal of the Econometric Society* 74 (2): 431–97. <https://doi.org/10.1111/j.1468-0262.2006.00668.x>.
- Bappenas. 1973–78. *Pedoman Pelaksanaan Bantuan Pembangunan Sekolah Dasar Menurut Instruksi Presiden Republik Indonesia*. Jakarta. Selected pages at <https://github.com/droodman/Duflo-2001/tree/main/Printed%20sources>.
- Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *Quarterly Journal of Economics* 119 (1): 249–75. <https://doi.org/10.1162/003355304772839588>.
- Biro Pusat Statistik (BPS). 1972. *Sensus Penduduk 1971: Diperintji Menurut Propinsi dan Kabupaten/Kotamadya*. Selected pages at <https://github.com/droodman/Duflo-2001/tree/main/Printed%20sources>.
- BPS. 1974. *Sensus Penduduk 1971: Series E*. Selected pages at <https://github.com/droodman/Duflo-2001/tree/main/Printed%20sources>.
- Bischof, Daniel. 2017. "New Graphic Schemes for Stata: plotplain and plottig." *Stata Journal* 17 (3): 748–59. <https://doi.org/10.1177/1536867X1701700313>.

- Bleakley, Hoyt. 2007. "Disease and Development: Evidence from Hookworm Eradication in the American South." *Quarterly Journal of Economics* 122 (1): 73–117. <https://academic.oup.com/qje/article-abstract/122/1/73/1924773>.
- Bollen, Kenneth A., Paul P Biemer, Alan F. Karr, Stephen Tueller, and Marcus E. Berzofsky. 2016. "Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis." *Annual Review of Statistics and Its Application* 3 (1): 375–92. <https://doi.org/10.1146/annurev-statistics-011516-012958>.
- Brewer, K. R. W., and R. W. Mellor. 1973. "The Effect of Sample Structure on Analytical Surveys." *Australian Journal of Statistics* 15 (3): 145–52. <https://doi.org/10.1111/j.1467-842X.1973.tb00195.x>.
- Burnside, Craig, and David Dollar. 2000. "Aid, Policies, and Growth." *American Economic Review* 90 (4): 847–68. <https://doi.org/10.1257/aer.90.4.847>.
- Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90 (3): 414–27. <https://doi.org/10.1162/rest.90.3.414>.
- Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press. <https://ipcid.org/evaluation/apoio/Microeconometrics%20-%20Methods%20and%20Applications.pdf>.
- Davidson, Russell, and James G. MacKinnon. 1999. "The Size Distortion of Bootstrap Tests." *Econometric Theory* 15 (3): 361–76. <http://www.jstor.org/stable/3533339>.
- Davidson, Russell, and James G. MacKinnon. 2010. "Wild Bootstrap Tests for IV Regression." *Journal of Business & Economic Statistics: A Publication of the American Statistical Association* 28 (1): 128–44. <https://doi.org/10.1198/jbes.2009.07221>.
- de Chaisemartin, Clément, and Xavier D'Haultfoeuille. 2017. "Fuzzy Differences-in-Differences." *The Review of Economic Studies* 85 (2): 999–1028. <https://doi.org/10.1093/restud/rdx049>.
- de Chaisemartin, Clément de, and Xavier D'Haultfoeuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96. <https://doi.org/10.1257/aer.20181169>.
- Duflo, Esther. 1999. *Essays in Empirical Development Economics*. Dissertation. Massachusetts Institute of Technology. <https://dspace.mit.edu/handle/1721.1/9516>.
- Duflo, Esther. 2000. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." Working Paper 7860. NBER. <https://doi.org/10.3386/w7860>.
- Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4): 795–813. <https://doi.org/10.1257/aer.91.4.795>.
- Duflo, Esther. 2004a. "The Medium Run Effects of Educational Expansion: Evidence from a Large School Construction Program in Indonesia." *Journal of Development Economics* 74 (1): 163–97. <https://doi.org/10.1016/j.jdeveco.2003.12.008>.
- Duflo, Esther. 2004b. "Scaling up and Evaluation." In Francois Bourguignon and Boris Pleskovic, eds. *Annual World Bank Conference on Development Economics: Accelerating Development*: 341–69. <http://hdl.handle.net/10986/14889>.

- DuMouchel, William H., and Greg J. Duncan. 1983. "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples." *Journal of the American Statistical Association* 78 (383): 535–43. <https://doi.org/10.2307/2288115>.
- Easterly, William, Ross Levine, and David Roodman. 2004. "Aid, Policies, and Growth: Comment." *American Economic Review* 94 (3): 774–80. <https://doi.org/10.1257/0002828041464560>.
- Finlay, Keith, and Leandro M. Magnusson. 2019. "Two Applications of Wild Bootstrap Methods to Improve Inference in cluster-IV Models." *Journal of Applied Econometrics* 34 (6): 911–33. <https://doi.org/10.1002/jae.2710>.
- Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225 (2): 254–77. <https://doi.org/10.1016/j.jeconom.2021.03.014>.
- Hausman, J. A. 1978. "Specification Tests in Econometrics." *Econometrica* 46 (6): 1251–71. <https://doi.org/10.2307/1913827>.
- Hausman, Jerry A., and David A. Wise. 1981. "Stratification on an Endogenous Variable and Estimation: The Gary Income Maintenance Experiment." In Charles F. Manski and Daniel L. McFadden, eds. *Structural Analysis of Discrete Data with Econometric Applications*. MIT Press. <https://eml.berkeley.edu/~mcfadden/discrete/chapter10.pdf>.
- Heneveld, Ward. 1978. "The Distribution of Development Funds: New School Building in East Java." *Bulletin of Indonesian Economic Studies* 14 (1): 63–79. <https://doi.org/10.1080/00074917812331333251>.
- Heneveld, Ward. 1979. "Indonesian Education in The Seventies: Problems of Rapid Growth." *Southeast Asian Affairs*: 142–54. <http://www.jstor.org/stable/27908373>.
- Hill, Hal, Budy P. Resosudarmo, and Yogi Vidyattama. 2008. "Indonesia's Changing Economic Geography." *Bulletin of Indonesian Economic Studies* 44 (3): 407–35. <https://doi.org/10.1080/00074910802395344>.
- Hsiao, Allan. 2024. "Educational Investment in Spatial Equilibrium: Evidence from Indonesia." October 3. https://allanhsiao.com/files/Hsiao_schools.pdf.
- Improving the Efficiency of Educational Systems (IEES). 1986. "Indonesia: Education and Human Resources Sector Review." Chapter 6. https://pdf.usaid.gov/pdfs_docs/PNAAX852.pdf.
- Jakiela, Pamela. 2021. "Simple Diagnostics for Two-Way Fixed Effects." *arXiv [econ.GN]*. <http://arxiv.org/abs/2103.13229>.
- Jann, Ben. 2007. "Making Regression Tables Simplified." *Stata Journal* 7 (2): 227–44. <https://doi.org/10.1177/1536867X0700700207>.
- Jann, Ben. 2014. "Plotting Regression Coefficients and Other Estimates." *Stata Journal* 14 (4): 708–37. <https://doi.org/10.1177/1536867X1401400402>.
- Jann, Ben. 2023. "Color Palettes for Stata Graphics: An Update." *Stata Journal* 23 (2): 336–85. <https://doi.org/10.1177/1536867X231175264>.
- Kahn-Lang, Ariella, and Kevin Lang. 2019. "The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications." *Journal of Business & Economic Statistics* 38 (3): 613–20. <https://doi.org/10.1080/07350015.2018.1546591>.
- Khanna, Gaurav. 2023. "Large-Scale Education Reform in General Equilibrium: Regression Discontinuity Evidence from India." *Journal of Political Economy*, 131 (2): 541–91. <https://doi.org/10.1086/7350015>.

[//doi.org/10.1086/721619](https://doi.org/10.1086/721619).

Kleibergen, Frank, and Richard Paap. 2006. "Generalized Reduced Rank Tests Using the Singular Value Decomposition." *Journal of Econometrics* 133 (1): 97–126. <https://doi.org/10.1016/j.jeconom.2005.02.011>.

Lee, Jin Young, and Gary Solon. 2011. "The Fragility of Estimated Effects of Unilateral Divorce Laws on Divorce Rates." *B.E. Journal of Economic Analysis & Policy* 11 (1). <https://doi.org/10.2202/1935-1682.2994>.

Liang, Kung-Yee, and Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73 (1): 13–22. <https://doi.org/10.1093/biomet/73.1.13>.

Melly, Blaise, and Giulia Santangelo. 2015. "The Changes-in-changes Model with Covariates." <http://www.hec.unil.ch/documents/seminars/deep/1591.pdf>.

Minnesota Population Center (MPC). 2020. Integrated Public Use Microdata Series, International: Version 7.3. <https://international.ipums.org>.

Moreira, Marcelo J. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica* 71 (4): 1027–48. <https://doi.org/10.1111/1468-0262.00438>.

Muliati, Iene, and Mitchell Wiener. 2014. "Pension Reform Experience in Indonesia." In Benedict Clements, Frank Eich, and Sanjeev Gupta, eds. *Equitable and Sustainable Pensions: Challenges and Experience*. International Monetary Fund. <https://doi.org/10.5089/9781616359508.071>.

Peters, H. Elizabeth. 1986. "Marriage and Divorce: Informational Constraints and Private Contracting." *American Economic Review* 76 (3): 437–54. <http://www.jstor.org/stable/1813361>.

Pfeffermann, Danny. 1993. "The Role of Sampling Weights When Modeling Survey Data." *International Statistical Review/Revue Internationale de Statistique* 61 (2): 317–337. <https://doi.org/10.2307/1403631>.

Potter, Frank, and Yuhong Zheng. 2015. "Methods and Issues in Trimming Extreme Weights in Sample Surveys." Proceedings of the Joint Statistical Meetings Survey Research Methods Section. American Statistical Association. <http://www.asasrms.org/Proceedings/y2015/files/234115.pdf>.

Roodman, David. 2018. "The Impacts of Hookworm Eradication in the American South. A Replication Study of Bleakley (The Quarterly Journal of Economics, 2007)." *International Journal for Re-Views in Empirical Economics*. <https://doi.org/10.18718/81781.7>.

Roodman, David. 2023. "Large-Scale Education Reform in General Equilibrium: Regression Discontinuity Evidence from India: Comment." arXiv [Econ.GN]. <http://arxiv.org/abs/2303.11956>.

Roodman, David. 2025. "Julia as a Universal Platform for Statistical Software Development." *Stata Journal* 25 (2). <https://doi.org/10.1177/1536867X251341105>.

Roodman, David, James G. MacKinnon, Morten Ørregaard Nielsen, and Matthew D. Webb. 2019. "Fast and Wild: Bootstrap Inference in Stata Using boottest." *Stata Journal* 19 (1): 4–60. <https://doi.org/10.1177/1536867X19830877>.

Roth, Jonathan. 2022. "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends." *American Economic Review: Insights* 4 (3): 305–22. <https://doi.org/10.1257/aeri.20210236>.

- Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. <https://doi.org/10.1177/0956797611417632>.
- Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50 (2): 301–16. <https://doi.org/10.3368/jhr.50.2.301>.
- Stock, James H., and Motohiro Yogo. 2002. "Testing for Weak Instruments in Linear IV Regression." Technical Working Paper Series. National Bureau of Economic Research. <https://nber.org/papers/t0284>.
- Sudharsanan, Nikkil, Yuan Zhang, Collin F. Payne, William Dow, and Eileen Crimmins. 2020. "Education and Adult Mortality in Middle-Income Countries: Surprising Gradients in Six Nationally-Representative Longitudinal Surveys." *SSM - Population Health* 12 (December): 100649. <https://doi.org/10.1016/j.ssmph.2020.100649>.
- Surbakti, Pajung. 1995. *Indonesia's National Socio-Economic Survey: A Continual Data Source for Analysis on Welfare Development*. Central Bureau of Statistics. https://rand.org/content/dam/rand/www/external/labor/bps/manualpdf/susenas/surbakti_1995_review.pdf.
- Wang, Feng, Haiying Wang, and Jun Yan. 2022. "Diagnostic Tests for the Necessity of Weight in Regression with Survey Data." *International Statistical Review/Revue Internationale de Statistique* 91 (1): 55–71. <https://doi.org/10.1111/insr.12509>.
- Winship, Christopher, and Larry Radbill. 1994. "Sampling Weights and Regression Analysis." *Sociological Methods & Research* 23 (2): 230–57. <https://doi.org/10.1177/0049124194023002004>.
- Wolfers, Justin. 2006. "Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results." *American Economic Review* 96 (5): 1802–20. <https://doi.org/10.1257/aer.96.5.1802>.
- Xu, Kim Qinzi, Jessica Yi Han Aw, and Collin F. Payne. 2024. "Inequalities in Mortality in the Asia-Pacific: A Cross-National Comparison of Socioeconomic Gradients." *Journals of Gerontology. Series B, Psychological Sciences and Social Sciences* 79 (4): 1–11. <https://doi.org/10.1093/geronb/gbad193>.

Appendices

A *Testing for effect heterogeneity*

Recent scholarship (de Chaisemartin and D’Haultfœuille 2020; Goodman-Bacon 2021) has highlighted how effect heterogeneity can bias TWFE estimates of mean effects. A key insight is that controlling for the two sets of fixed effects rearranges as well as restricts the identifying variation in the treatment. For example, if treatment is binary, then treated observations may effectively be assigned negative treatment after partialling out the fixed effects, while untreated observations may effectively be assigned positive treatment. When combined with certain patterns of effect heterogeneity, these shifts can reverse the sign of the effect estimate.

In this context, the Jakiela (2021) test for homogeneity is straightforward. One partials the fixed-effect dummies out of the treatment and outcome variables. Then one regresses the “residualized” outcome on the residualized treatment—separately for the treated and untreated subsamples. If the treatment effect is homogeneous, the two slope estimates will not differ (though the converse is not true).

I run the test on a variant of a Duflo (2001) regression. In the penultimate columns of Table 2 and Table 3 appear results from reduced-form regressions of years of schooling and the log hourly wage in 1995 on the treatment intensity indicator $D_j T_i$. The samples are restricted to men aged 2–6 and 12–17 in 1974. The minimal Duflo (2001) control set is included. Data corrections are incorporated. Standard errors are clustered by birth regency. Observations are weighted.

I modify the Jakiela test in two minor ways. Since treatment is not binary, I split the sample according to the (corrected) Duflo (2001) indicator of high-versus-low treatment. And since the representative regression includes other controls, I partial them out too. Figure A–1 depicts the data so processed. The blue and black circles are for treated and untreated observations respectively. For clarity, values are grouped and averaged for each age cohort–birth regency pair. Superimposed on the two scatter plots are linear and local polynomial smoothed fits. For schooling and the log hourly wage, the slopes of the linear fits are nearly identical. The p values for the nulls that they match are 0.61 and 0.91.

This check for treatment heterogeneity therefore finds essentially none.

B *Weighting, consistency, and efficiency*

Section 2.6 points out that if in the context of a particular estimation problem, sampling is endogenous, then weighting observations in inverse proportion to their sampling probability may be necessary for consistency. Typically, though, it increases the variance of estimators. This appendix investigates more precisely how the sampling program affects the consistency of OLS and its efficiency relative to weighted OLS (WOLS).

Consider the homogeneous treatment effect model

$$y = \alpha + x\beta + e \tag{5}$$

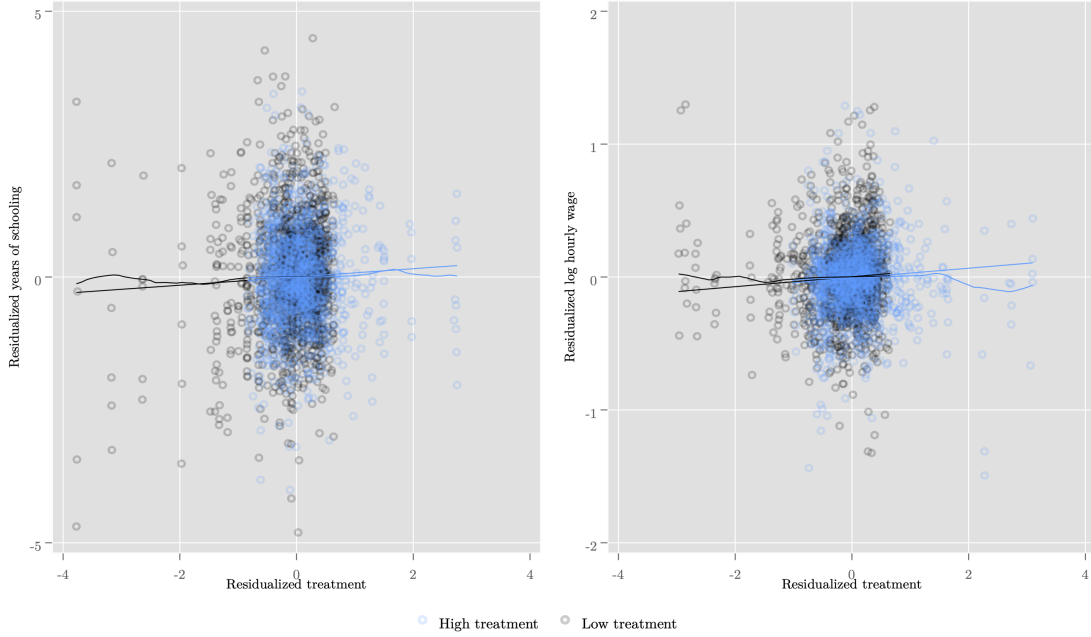


Figure A-1: Residualized outcome vs. residualized treatment in revised Duflo (2001) reduced-form wage regression, age and birth regency

Here, x , e , and y are variables in the population, with realizations (x_i, e_i, y_i) that are sampled with independent probabilities p_i . For the moment, the standard exogeneity assumption $E[e|x] = 0$ is not imposed to identify α and β . Instead, α and β are *defined* as the results from (unweighted) OLS regression of y on x in the population.²⁵ This implies only that $E[e] = 0$.

Suppose we perform OLS of y on x within the sample. What regression in the *population* will produce the same result in the large-population, large-sample limit? The answer: regression of y on x while weighting by p . While this population regression would weight each observation by its probability of observation, the unweighted sample regression would weight each observation in the population by its *realized* chance of observation, 0 or 1. Thus, asymptotically, the unweighted OLS estimate of β in the sample is

$$\hat{\beta}_u = \frac{Cov_p[x, y]}{Var_p[x]}$$

where the variances are computed over the population and subscripts indicate weighting by p . Substituting with (5), the asymptotic bias of $\hat{\beta}_u$ is

$$\hat{\beta}_u - \beta = \frac{Cov_p[x, \alpha + x\beta + e]}{Var_p[x]} - \beta = \frac{Cov_p[x, e]}{Var_p[x]} \quad (6)$$

$\hat{\beta}_u$ is unbiased for β if and only if the numerator on the right is 0.

There are at least two circumstances under which this condition holds:

²⁵This estimand is one considered in Brewer and Mellor (1973) and DuMouchel and Duncan (1983).

1. *The sampling probability is exogenous.* This condition permits p to be correlated with x , but not with e : $Cov[p, e] = 0$. In fact, it also requires $Cov[px, e] = 0$. I will call these two requirements second- and third-order conditions since they involve a second and third moment of the variables.²⁶
2. *The sampling probability is purely endogenous.* The second-order condition here is the mirror image of the previous one. p may be correlated with e but not x : $Cov[p, x] = 0$. Here too, a less-intuitive third-order condition is also required: $Cov[pe, x] = 0$.²⁷

To bring these mathematical statements to life, I simulate several data generating and sampling processes, apply OLS and WOLS in each case, and summarize the results in Figure B–1. In all cases, a population is generated with 5,000 independent standard normal draws for both x and a primary error term ε . The outcome is then computed as $y = x + e$ where e is set to ε (homoskedastic error; first row) or $\varepsilon(x+3)/5$ (heteroskedastic; second row). For each observation, the sampling probability p is 1 or 0.05 depending on whether $x > 0$ (purely exogenous; first column), or whether x and e have the same sign (second), or solely whether $e > 0$ (purely endogenous; third column), or $y > 0$ (fourth). Each pane shows the asymptotic best-fit line in the population, $y = x$, along with scatter plots of the samples, and OLS and WOLS fits thereto. Confidence intervals for the fits are also shown; for legibility they are some 11 standard errors wide, corresponding to the 99.99995% confidence level.

The first column of Figure B–1 confirms that under exogenous sampling, OLS is consistent—even in the presence of heteroskedasticity. In both panes of the column the OLS and WOLS best-fit lines all but coincide. Here, the sampling probability depends deterministically on x alone, which suffices to satisfy both conditions for the consistency of OLS under exogenous sampling—that p and px are uncorrelated with e in the population. Under homoskedasticity (top of column 1), OLS is also more efficient than WOLS, with its narrower confidence band rendered in purple. Under heteroskedasticity (next pane down), weighting by the inverse sampling probability happens in this case to counteract the inefficiency created by unequal error variances, by up-weighting the lower-variance observations on the left. (If p were higher for $x < 0$ instead of lower, then the weighting would compound rather than counteract the inefficiency caused by the heteroskedasticity.) Being thinner, the green band for the weighted estimator is fully obscured by the purple one for the unweighted estimator. The backgrounds of both panels in the left column are slightly shaded to indicate that here, unweighted OLS is consistent. In sum, under exogenous sampling, OLS is consistent and typically more efficient than WOLS, though heteroskedasticity can augment or counteract this efficiency advantage.

The second column shows why the third-order condition for exogenous sampling is necessary. Here, sampling is dense when x and e have the same sign. We still have the naïve exogenous sampling condition, $Cov[p, e] = 0$, because knowing the value of either variable tells one nothing about the expectation of the other. For example, the average error is the same—zero—in the densely and lightly sampled regions. But $Cov[px, e] \neq 0$: where p and x are both at their largest in magnitude, in the lower left and upper right, the average of e is too. So while WOLS remains consistent, OLS does not.

The third and fourth columns move to endogenous sampling. The third illustrates purely

²⁶The condition is equivalent to $E[pe] = E[p]E[e]$ and $E[pxe] = E[px]E[e]$, both of which are 0 since $E[e] = 0$. Define $\bar{p}$ as the normalized sampling probability $\frac{p}{E[p]}$. Then the numerator on the right side of (6) is $Cov_p[x, e] = Cov_{\bar{p}}[x, e] = E[\bar{p}xe] - E[\bar{p}x]E[\bar{p}e] = 0 - E[\bar{p}x] \cdot 0 = 0$.

²⁷The conditions now imply $E[px] = E[p]E[x]$ and $E[pe] = E[p]E[e]$. The numerator in (6) can now be developed as $Cov_p[x, e] = E[\bar{p}xe] - E[\bar{p}x]E[\bar{p}e] = E[\bar{p}e]E[x] - E[\bar{p}]E[x]E[\bar{p}e] = E[\bar{p}e]E[x](1 - E[\bar{p}])$, in which $E[\bar{p}] = 1$.

endogenous sampling as defined above: p depends only on e . Under homoskedasticity (top row), WOLS is once more consistent, at least for the slope of the best fit. This time the counterexample demonstrating the need for the third-order condition appears in the same column, in the second row. It is possible for heteroskedasticity to cause $Cov[pe, x] \neq 0$. Here, when p and e are at their largest, so is x . This makes OLS inconsistent even for the slope.

The last column of Figure B–1 shows that if p is related directly to the outcome y , as distinct from its unexplained component e , then unweighted OLS is inconsistent even under homoskedasticity. An exception here would be if the true best-fit line were horizontal—if there were no effect—for then direct dependence on y would be equivalent to direct dependence on e , bringing us back to the third column.

WOLS is consistent throughout. Usually it has higher variance than OLS. But in these examples, when only WOLS is consistent, its mean squared error—variance plus the squared bias—is lower.

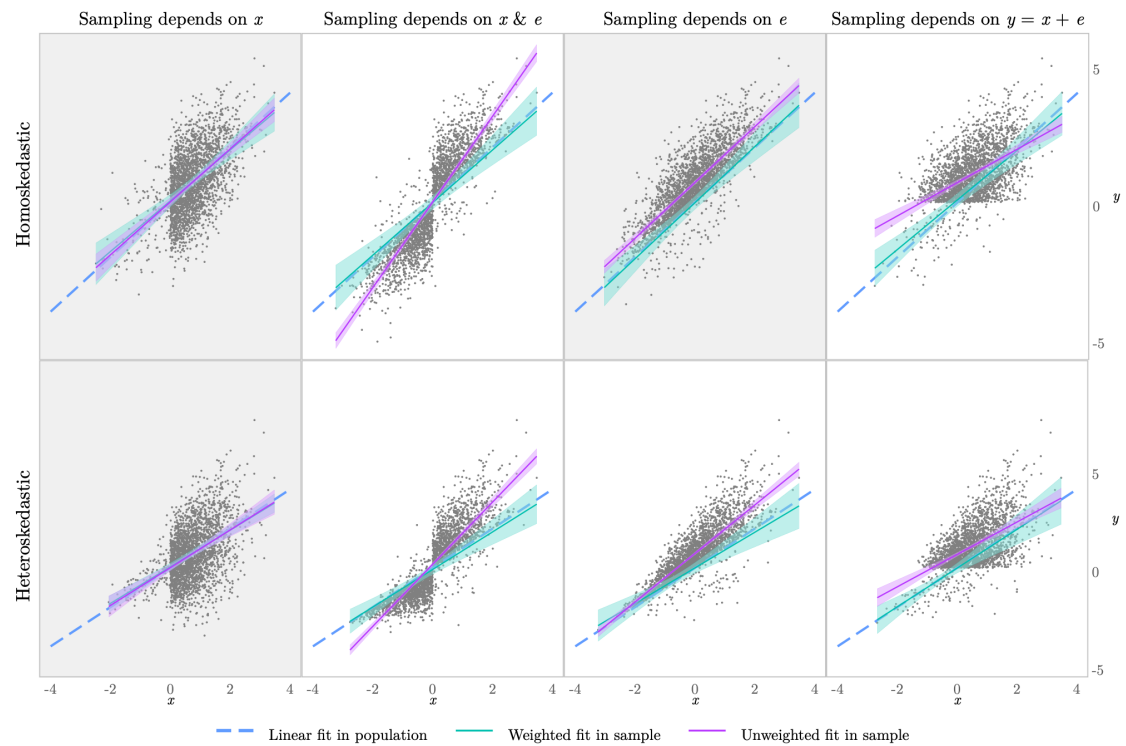


Figure B-1: Illustration of Unweighted and Weighted Regression in the presence of disproportionate sampling

Notes: For all panels, a single population of 5,000 independent standard normal draws for x and a primary error term ε are drawn. The outcome is constructed as $y = x + e$ where e is ε (first row) or $\varepsilon(x + 3)/5$ (second row). The sampling probability is 1 or 0.05 depending on whether or not $x > 0$ (first column), or x and e have the same sign (second), or $e > 0$ (third), or $y > 0$ (fourth). The OLS fit in the population is plotted along with fits on the sample with and without weighting by $1/p$. For legibility the confidence level on the confidence intervals is 99.999995%. Confidence intervals are robust to heteroskedasticity except in the unweighted fits in the first row. Where the unweighted fit is consistent for the slope, the background is slightly shaded.

C 2SLS estimates of the effect of years of schooling on labor outcomes, 1995, with instruments “by birth year”

Table C-1: 2SLS estimates of the effect of years of schooling on labor outcomes, 1995, with instruments “by birth year”

	Employment			Log hourly wage		
	(1)	(2)	(3)	(4)	(5)	(6)
<i>A. With controls based on number of children aged 5–14</i>						
Coefficient	0.101 (0.0210)	0.10 (0.031)	0.090 (0.036)	0.0675 (0.0280)	0.11 (0.048)	0.14 (0.052)
		$(-\infty, \infty)$	$(-\infty, \infty)$		$(-\infty, -0.24]$	$(-\infty, -0.23] \cup [0.34, \infty)$
KP <i>F</i>		1.86	1.04		1.08	1.44
Observations		152,989	152,989		60,663	60,663
<i>B. Without extra controls</i>						
Coefficient		0.072 (0.030)	0.068 (0.039)		0.046 (0.071)	0.12 (0.065)
		$(-\infty, \infty)$	$(-\infty, \infty)$		$(-\infty, -0.25]$	$(-\infty, -0.14]$
KP <i>F</i>		1.07	0.66		0.34	0.90
Data corrections		✓	✓		✓	✓
Clustered		✓	✓		✓	✓
Weighted			✓			✓
Observations	152,989	152,989	152,989	60,663	60,663	60,663

Notes: See notes to Table 4. The regressions reported here differ only in using the “by birth year” instrument set, which consists of interactions between D_j and dummies for ages 2–24 as of 1974, except that age ≥ 12 defines one category. Columns 1 and 4 are copied from Dufo (2001), Table 7.

D Pooled results, all available years

Table D-1: Reduced-form TWFE impact estimates for years of schooling, pooled data for 1995, 2005, 2011–2014, 2017–19

	(1)	(2)	Hausman <i>p</i> : weighted vs. not
<i>A. With controls based on number of children aged 5–14</i>			
Experiment	0.11 (0.032)	0.098 (0.036)	0.68
Placebo	0.043 (0.037)	0.099 (0.042)	0.04
Kink	0.081 (0.037)	0.070 (0.040)	0.64
<i>B. Without extra controls</i>			
Experiment	0.081 (0.030)	0.079 (0.035)	0.93
Placebo	0.055 (0.035)	0.094 (0.040)	0.12
Kink	0.035 (0.035)	0.035 (0.038)	0.99
Data corrections	✓	✓	
Clustered	✓	✓	
Weighted		✓	
Experiment <i>N</i>	668,141	668,141	
Placebo <i>N</i>	564,718	564,718	
Kink <i>N</i>	1,166,869	1,166,869	

Notes: Notes to Table 2 apply. As dictated by availability of needed variables.

Table D-2: Reduced-form TWFE impact estimates for labor outcomes, 1995, 2005, 2011–14

	Formal sector employment			Log hourly wage		
	(1)	(2)	Hausman <i>p</i> : weighted vs. not	(3)	(4)	Hausman <i>p</i> : weighted vs. not
<i>A. With controls based on number of children aged 5–14</i>						
Experiment	0.0057 (0.0024)	–0.00045 (0.0031)	0.002	0.021 (0.0073)	0.0093 (0.0094)	0.04
Placebo	–0.0035 (0.0039)	–0.018 (0.0057)	0.0001	–0.015 (0.011)	–0.015 (0.016)	0.97
Kink	0.010 (0.0048)	0.014 (0.0065)	0.27	0.025 (0.012)	0.035 (0.018)	0.35
<i>B. Without extra controls</i>						
Experiment	0.0011 (0.0022)	–0.0039 (0.0028)	0.008	–0.0048 (0.0070)	–0.015 (0.0086)	0.07
Placebo	–0.0035 (0.0036)	–0.015 (0.0055)	0.001	–0.0092 (0.0099)	–0.012 (0.014)	0.75
Kink	0.0064 (0.0045)	0.0080 (0.0061)	0.64	0.010 (0.011)	0.015 (0.017)	0.59
Data corrections	✓	✓		✓	✓	
Clustered	✓	✓		✓	✓	
Weighted		✓			✓	
Experiment <i>N</i>	460,853	460,853		144,799	144,799	
Placebo <i>N</i>	404,451	404,451		99,396	99,396	
Kink <i>N</i>	810,736	810,736		241,957	241,957	

Notes: See notes to Table 3. The 2005 data covers employment but not wages.

Table D-3: 2SLS estimates of the effect of years of schooling on labor market outcomes, pooled data for 1995, 2005, 2011–14

	Employment				Log hourly wage			
	Instrument by young/old		Kink instrument		Instrument by young/old		Kink instrument	
	(1)	(2)	(3)	(4)	(6)	(7)	(8)	(9)
<i>A. With controls based on number of children aged 5–14</i>								
Coefficient	0.059 (0.026)	–0.0056 (0.039)	0.17 (0.14) $(-\infty, -0.24] \cup$ $[-0.01, \infty)$	0.92 (3.05) $(-\infty, \infty)$	0.24 (0.085)	0.17 (0.16) $(-\infty, \infty)$	–0.37 (0.74) $(-\infty, 0.07] \cup$ $[0.19, \infty)$	–0.17 (0.24) $(-\infty, 0.08] \cup$ $[0.40, \infty)$
KP <i>F</i>	9.02	4.59	1.86	0.087	5.16	0.91	0.42	1.55
<i>B. Without extra controls</i>								
Coefficient	0.016 (0.030) $[-0.15, 0.12]$	–0.055 (0.053) $(-\infty, 0.02]$	1.35 (12.4) $(-\infty, \infty)$	–0.29 (0.69) $(-\infty, \infty)$	0.59 (2.25) $(-\infty, \infty)$	0.32 (0.28) $(-\infty, \infty)$	0.013 (0.074) $(-\infty, 0.15]$	0.010 (0.076) $(-\infty, 0.16]$
KP <i>F</i>	5.29	4.03	0.012	0.25	0.048	0.82	4.60	5.08
Data corrections	✓	✓	✓	✓	✓	✓	✓	✓
Clustered	✓	✓	✓	✓	✓	✓	✓	✓
Weighted		✓		✓		✓		✓
Pre-trend control			✓	✓			✓	✓
Observations	460,853	460,853	858,355	858,355	144,358	144,358	249,612	249,612

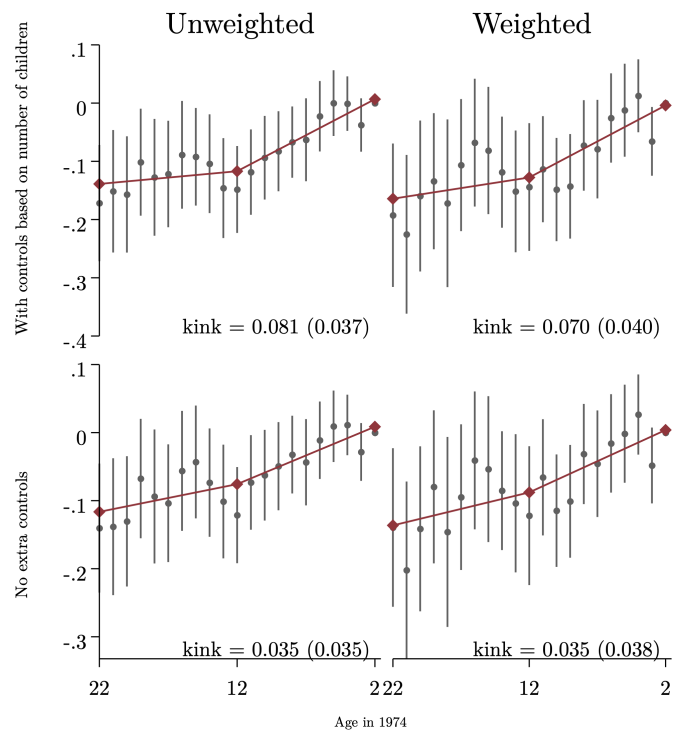


Figure D-1: Coefficients on interactions between age dummies and Inpres intensity in schooling regressions, 1995, 2005, 2011-14, 2017-19

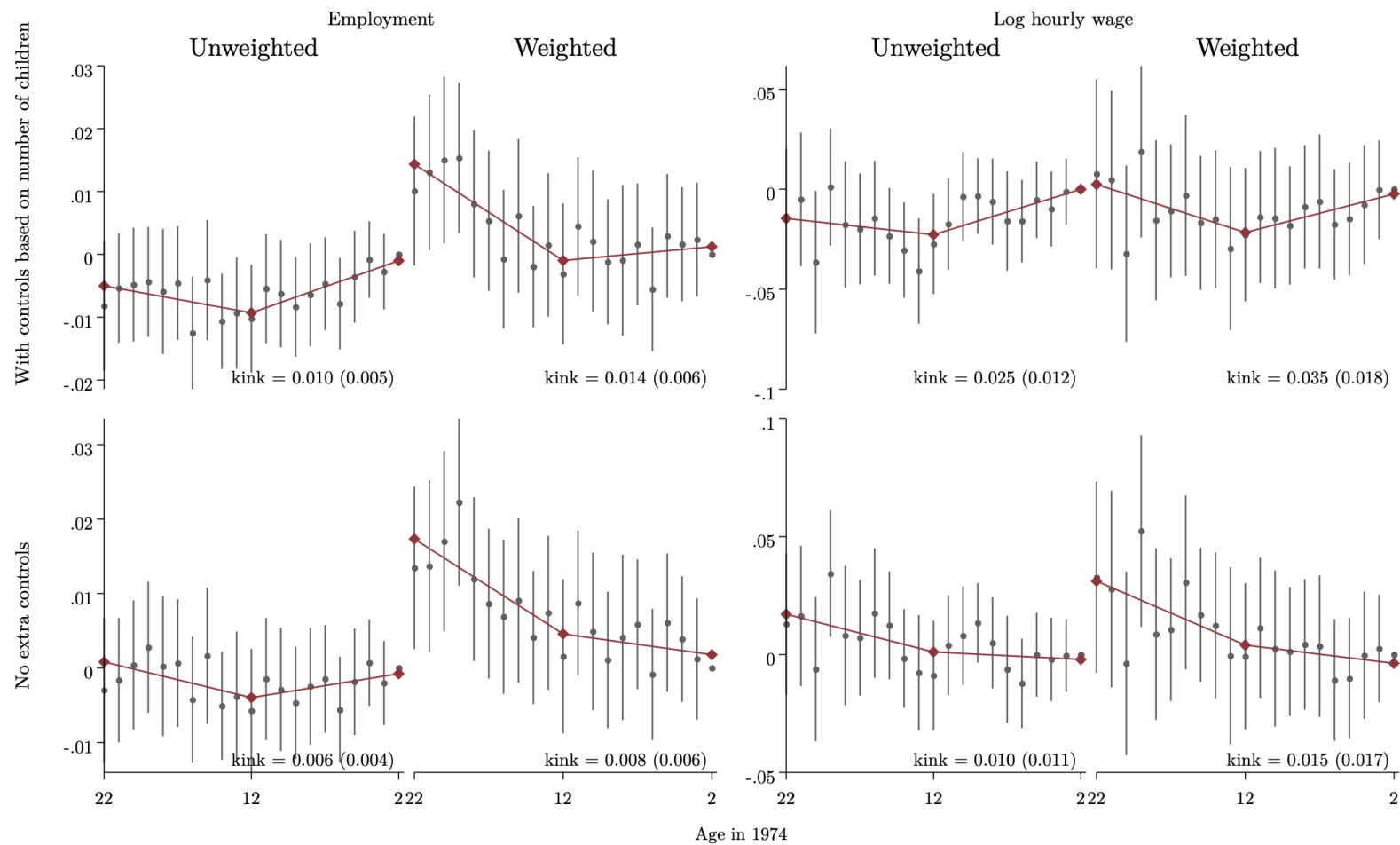


Figure D-2: Coefficients on interactions between age dummies and Inpres intensity in labor outcome regressions, 1995, 2005, 2011–14

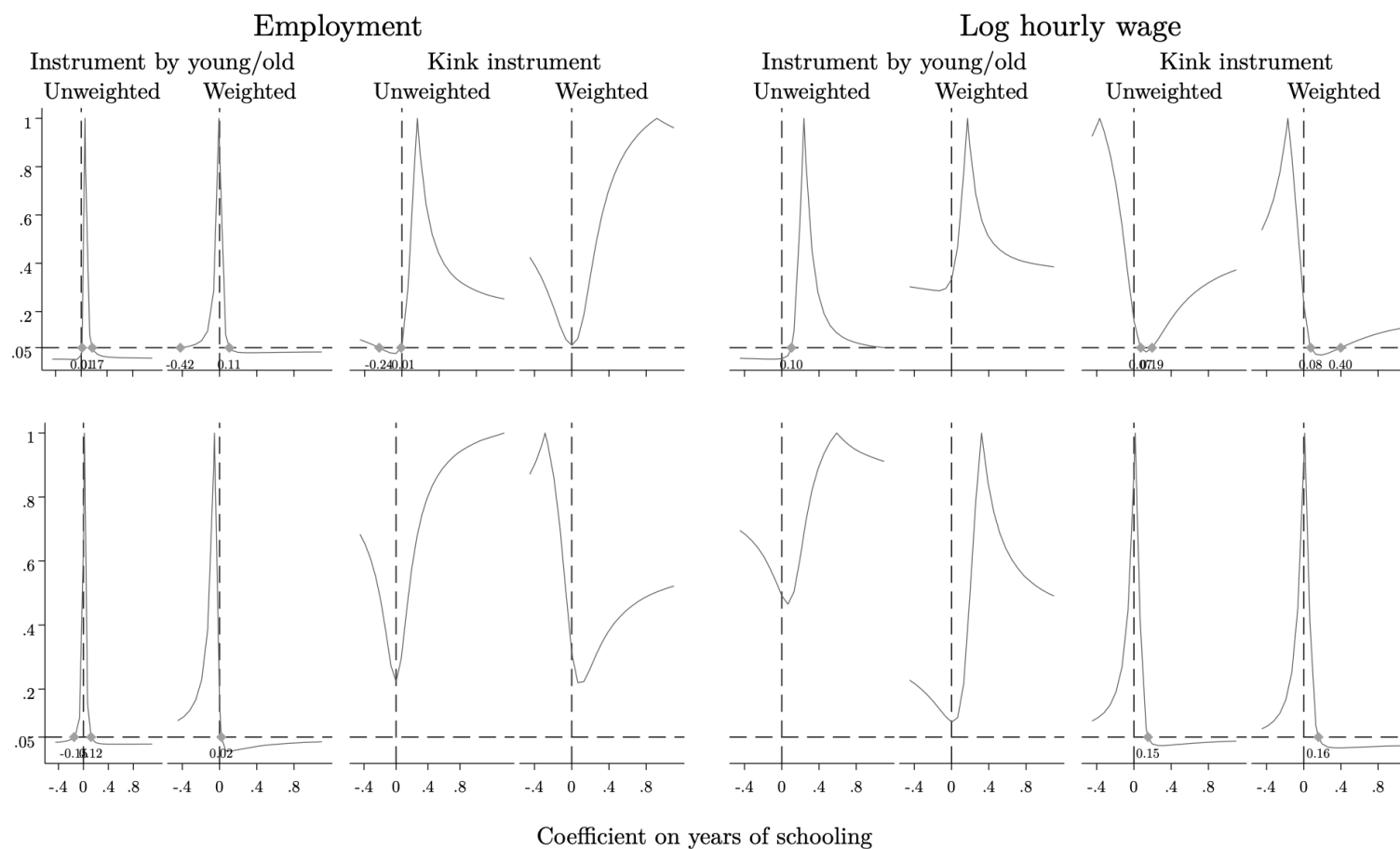


Figure D-3: Confidence curves from wild-bootstrapped tests of the coefficient on schooling in 2SLS regressions in Table D-3

E Returns to primary school completion, 2011–14

Table E-1: 2SLS estimates of the effect of years of primary completion on labor market outcomes, 2011–14

	Employment				Log hourly wage			
	Instrument by young/old		Kink instrument		Instrument by young/old		Kink instrument	
	(1)	(2)	(3)	(4)	(6)	(7)	(8)	(9)
<i>A. With controls based on number of children aged 5–14</i>								
Coefficient	0.11 (0.14)	–0.14 (0.11)	0.73 (0.73)	0.84 (0.50)	1.41 (0.66)	0.34 (0.38)	–6.48 (11.1)	–3.03 (4.95)
KP <i>F</i>	[–0.19, 0.48] 15.7	[–0.40, 0.07] 24.5	(–∞, ∞) 3.09	[–0.00, ∞) 7.30	[0.26, ∞) 10.5	(–∞, ∞) 18.3	(–∞, –0.26] 0.40	(–∞, ∞) 0.56
<i>B. Without extra controls</i>								
Coefficient	–0.21 (0.22)	–0.32 (0.15)	0.45 (0.85)	0.42 (0.47)	–1.02 (1.75)	–1.15 (0.88)	–0.85 (3.15)	1.04 (3.45)
KP <i>F</i>	(–∞, 0.21] 7.25	(–∞, –0.08] 15.3	(–∞, ∞) 1.68	(–∞, ∞) 6.06	(–∞, ∞) 1.62	(–∞, 0.14] 6.62	(–∞, ∞) 0.51	(–∞, ∞) 0.41
Data corrections	✓	✓	✓	✓	✓	✓	✓	✓
Clustered	✓	✓	✓	✓	✓	✓	✓	✓
Weighted		✓		✓		✓		✓
Pre-trend control			✓	✓			✓	✓
Observations	382,383	382,383	705,366	705,366	113,297	113,297	188,949	188,949

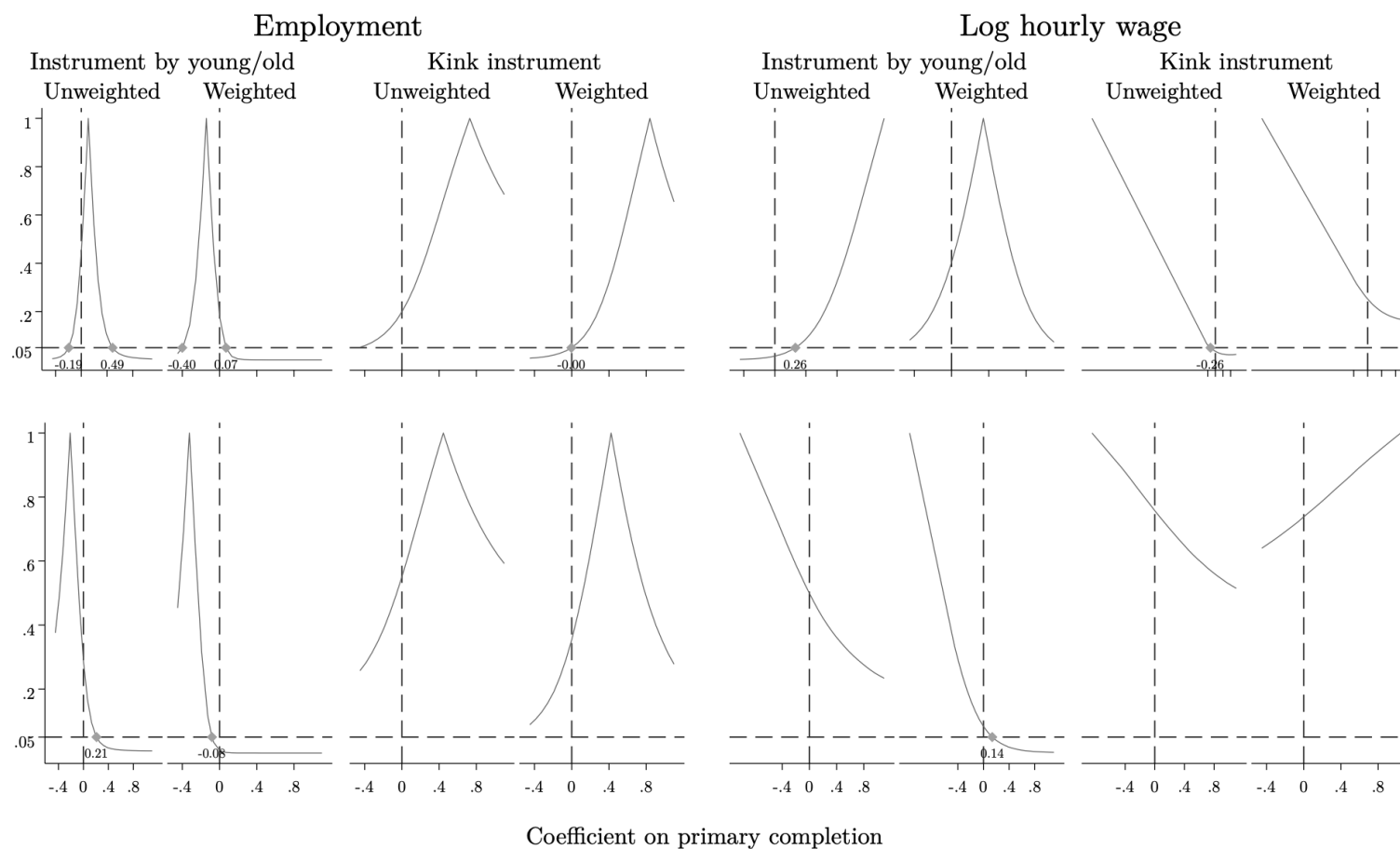


Figure E-1: Confidence curves from wild-bootstrapped tests of the coefficient on primary completion in 2SLS regressions in Table E-1

F Quantile-based estimation

The changes-in-changes (CIC) estimator of Athey and Imbens (2006) is defined for the 2×2 set-up. For a concrete example of its mechanics, suppose a person in a treatment group is observed before treatment to earn a wage of 1000 rupiah per hour. Suppose that wage rate would place the person in the 25th percentile of the *control* group's distribution for the same, pre-treatment period. Finally, suppose that the 25th percentile of the control group's post-treatment distribution is 2000 rupiah per hour. Then this treatment-group subject would contribute an observation of 2000 to the counterfactual distribution for the *treatment* group in the post-treatment period. The gap between a given quantile of this counterfactual distribution and the observed quantile is an estimate of the impact of treatment at that point in the distribution.

CIC is designed for continuous outcomes, which here restricts us to wages. (The estimator's results for schooling attainment, which is quantized by year, are degenerate.) Of necessity I apply CIC within a 2×2 set-up, with the same definitions of the high- and low- treatment groups and pre- and post-treatment cohorts as in the Duflo (2001) 2×2 DID. I incorporate the needed controls—dummies for year of birth, regency of birth, and year of survey—using the method of Melly and Santangelo (2015). I estimate in the 1995 sample, the 2011–14 sample, and the two together.

Estimates by decile appear in Figure F-1, along with bootstrap-based 95% confidence intervals. p values from tests of various null hypotheses synthesizing the results are gathered in Table F-1. Especially when not weighting, the results for 1995 exhibit a dramatic swing toward negative impacts in the lower deciles. However, the Cramér–von Mises tests in the first two columns of in Table F-1 do not strongly reject the null of no impact at all deciles ($p = 0.30$ without weights and 0.59 with); perhaps the test is under-powered. At any rate, in the larger, pooled follow-up of 2011–14—in prime earnings years—there is a stronger suggestion of *positive* impact at lower deciles. The null of a positive impact at all deciles is rejected only at $p = 0.74$ when weighting, while the null of negative impact at all deciles is rejected at $p = 0.07$.

The finding of a benefit for poorer people is consistent with the story that Inpres increased schooling for the less-well-off, those for whom the marginal year of schooling was in primary school. The slight hint of negative impacts at higher deciles might reflect the confounding from secondary school expansion discussed in section 5.

de Chaisemartin and D'Haultfœuille (CH, 2017) introduces a variant of CIC that offers a quantile-based way to estimate returns to schooling, as distinct from reduced-form estimation as just performed. Indeed, that paper illustrates its methods on the Duflo (2001) data, and estimates the wage return to a year of schooling at 0.099 log points (standard error 0.017). However, the CH Wald CIC estimator requires a control arm within which expected treatment—years of schooling—is stable over time, as well as arms in which expected treatment strictly increases or decreases. To accommodate this requirement, CH discards Duflo (2001)'s Inpres-based identification strategy. CH instead groups regencies by whether average schooling rose, fell, or stayed about the same between the age 12–17 and age 2–6 cohorts. In the instrumental variables perspective, this instrument schooling using a trichotomous factor variable that itself depends on schooling and thus is about as presumptively endogenous. As is demonstrated in appendix G, the combination of this procedure for forming supergroups and the Wald CIC estimator introduces

endogeneity bias. I therefore do not view the CH results for schooling in Indonesia as much more informative as to causal impacts than OLS-type results.

Table F-1: Rejection p values for various null hypotheses about the CIC results

Null	1995		2011–14		1995, 2011–14	
No effect, all deciles	0.30	0.59	0.00	0.23	0.003	0.03
Constant effect	0.46	0.30	0.00	0.27	0.00	0.01
Positive effect, all deciles	0.24	0.26	0.56	0.74	0.43	0.55
Negative effect, all deciles	0.51	0.68	0.00	0.07	0.00	0.004
Weighted	✓		✓		✓	

Notes: All p values are from Cramér–von Mises tests.

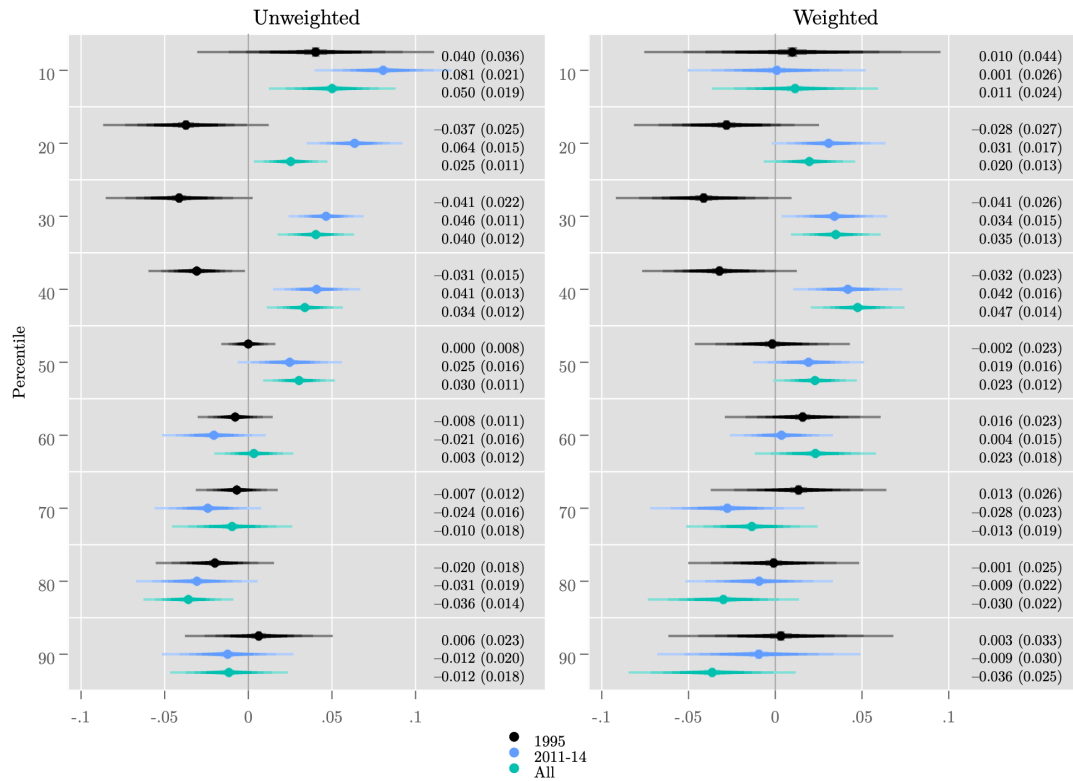


Figure F-1: Changes-in-changes estimates of impact of high Inpres treatment on log hourly wage at various percentiles

Notes: The plot shows, at various percentiles, estimates of the impact of high Inpres treatment on earnings, along with bootstrapped 95% confidence intervals. Point estimates and standard errors are printed on the right. Results are from the CIC-with-controls estimator of Melly and Santangelo (2015). The definitions of before and after periods and low- and high-treatment groups are the same as in the Duflo (2001) 2×2 DID regressions. The controls are dummies for regency of birth, year of birth, and year of survey.

G Simulation of de Chaisemartin and D’Haultfœuille (2017)’s Wald CIC

Appendix F asserts that a combination of procedures in de Chaisemartin and D’Haultfœuille (CH, 2017)— gathering groups into three supergroups, then performing “Wald CIC”—will produce biased impact estimates if treatment is endogenous. (In the CH application to Duflo (2001), “treatment” refers not to the basis of Duflo (2001)’s instruments, planned Inpres schools per child, but to schooling attainment, which is indeed presumed endogenous.) This appendix justifies the concern with a simulation.

The simulated data come from 100 groups, indexed by j , each with 100 subjects, indexed by i . Each subject is observed once, with equal probability in time period 0 or 1. The data generating process is

$$S_{ijt} = \lceil Z_{jt} + s_{jt} + u_{ijt} \rceil$$

$$Y_{ijt} = \beta S_{ijt} + v_{ijt}$$

$Z_{jt} \sim \text{i.i.d. } \mathcal{N}(0, 1)$ is an exogenous, standard-normal, group-level component of schooling. s_{jt} is a group-level time trend such that $s_{j0} \equiv 0$ and $s_{j1} \sim \text{i.i.d. } \mathcal{N}(0, 1)$ for all j ; it determines whether treatment in each group tends to rise, fall, or stay about the same. $u_{ijt}, v_{ijt} \sim \mathcal{N}(0, 1)$ are error terms with cross-correlation ρ , which constitutes the endogeneity. Schooling, S_{ijt} , is discretized with the ceiling operator because Wald CIC is defined only when treatment has finite support. Y_{ijt} is wages.

To collect the groups into supergroups characterized by downward, flat, or upward trends in treatment, CH suggest using a χ^2 test with a high p value, 0.5, to judge whether average treatment changes in each group between pre- and post-treatment periods. I apply that procedure to each simulated data set. Then I apply three estimators: 2SLS-based Wald DID on the original grouping, using the perfect instrument Z ; Wald DID instrumenting with supergroup dummies; and CH’s Wald CIC on the supergroups. The first estimator serves to benchmark the other two.

I run two sets of 100 simulations. In the first, $\rho = 0$, making the treatment S_{ijt} exogenous.²⁸ In the second, $\rho = 0.95$, making treatment highly endogenous. In both, the impact parameter β is 0.

Results from the simulations appear in Table G–1. They confirm that the Wald DID and CIC estimators run on the supergroups are unbiased only when treatment is exogenous. Then, the estimators produce nearly identical distributions centered on the correct value for β , zero. Introducing substantial endogeneity does not harm the Wald DID regressions with the perfect instrument, but lifts the mean estimates away from the true value of zero, by a standard deviation, when estimating from the CH supergroups.

²⁸A few simulations fail when, in the Wald CIC, the support of the constructed counterfactual post-treatment outcome distribution for the treatment groups does not cover that for the observed outcome distribution, so that changes at one or more quantiles cannot be estimated.

Table G-1: Average supergroup size and point estimates from DID-type estimators on simulated data

	$\rho = 0$		$\rho = 0.95$	
	Mean	Standard deviation	Mean	Standard deviation
Number of groups in supergroups defined by trend in treatment				
Decreasing	46.9	5.1	46.6	5.2
Flat	7.5	2.5	8.1	5.4
Increasing	45.6	4.9	45.8	4.9
Estimate of β				
Wald DID on groups	0.001	0.025	0.001	0.025
Wald DID on supergroups	0.001	0.021	0.020	0.020
Wald CIC on supergroups	0.001	0.022	0.018	0.023
Simulations	98		99	

Notes: True estimand in all simulations is 0. Simulated data sets consist of 100 groups with 100 subjects. Following de Chaisemartin and D'Haultfœuille (2017), groups are sorted into supergroups based on an F test of whether their average schooling rises or falls, using a threshold p value of 0.5. Wald DID is DID-type 2SLS on the original groups, instrumenting with the known, exogenous component of treatment. Wald DID and Wald CIC on supergroups are de Chaisemartin and D'Haultfœuille (2017) estimators and take the "flat" supergroup as the control group.