INSTITUTE for REPLICATION

No. 15
I4R DISCUSSION PAPER SERIES

Schooling and Labor Market Consequences of School Construction in Indonesia: Comment

David Roodman

January 2023

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 15

Schooling and Labor Market Consequences of School Construction in Indonesia: Comment

David Roodman[1]

[1]*Open Philanthropy, San Francisco/USA*

JANUARY 2023

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

ISSN: 2752-1931

Schooling and Labor Market Consequences of School Construction in Indonesia: Comment

David Roodman*

Abstract

Duflo (2001) exploits a 1970s schooling expansion in Indonesia to estimate the returns to schooling. Under the study's difference-in-differences (DID) design, two patterns in the data—shallower pay scales for younger workers and negative selection in treatment—can violate the parallel trends assumption and upward-bias results. In response, I follow up later, test for trend breaks timed to the intervention, and perform changes-in-changes (CIC). I also correct data errors, cluster variance estimates, incorporate survey weights to correct for endogenous sampling, and test for (and detect) instrument weakness. Weak identification–robust inference yields imprecise, positive estimates. CIC estimates tilt slightly negative.

JEL classification: I2, J31, O15
Keywords: education, wages, reanalysis

* Open Philanthropy, 182 Howard St., #225, San Francisco, CA 94105 (e-mail: david.roodman@openphilanthropy.org). Thanks to Daniel Feenberg at the National Bureau of Economic Research for running and troubleshooting scripts; to Esther Duflo for data, code, and comments; and to Jonathan Morduch, Otis Reid, and Justin Sandefur for comments on earlier drafts. Estimates and inferences were performed with ivreg2 (Baum, Schaffer, and Stillman 2007), reghdfe (Correia 2014), cic (Melly and Santangelo 2015), cmp (Roodman 2011), and boottest (Roodman et al. 2019). Tables and graphs were made with esttab (Jann 2007), coefplot (Jann 2014), palettes (Jann 2022), and blindschemes (Bischof 2017).

I. Introduction

In 1973, the government of Indonesia channeled part of its oil revenue windfall into a project to erect thousands of three-room schoolhouses across its far-flung archipelagic territory. The Instruksi Presiden Sekolah Dasar program (Inpres SD) became one of the largest schooling expansions ever, roughly doubling the country's stock of primary schools in the first six years alone (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. It typifies its moment in intellectual history, creatively exploiting stratagems for causal inference such as difference-in-differences (DID) and natural experiment–derived instrumentation (Angrist and Krueger 1999). And 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 test returns a null result where it ought to; instrument validity is tested for; potential sources of bias are named and addressed. Graphical analysis adds credibility by making the case that the schooling shock manifests in the data with appropriate timing. Few studies so credibly claim to identify the impacts of schooling at scale.

Nevertheless, I reanalyze Duflo (2001) in order to assess how well it improves the measurement of the returns to schooling—improves, that is, is over methods that do not address endogeneity, such as ordinary least squares (OLS) fitting of Mincer (1974) labor functions.

The assessment generates "micro" and "macro" comments. The micro comments pertain to data and technique; most are natural consequences of taking a fresh look after 21 years. Duflo (2001) uses classical variance estimators, which Bertrand, Duflo, and Mullainathan (2004) shows are downward-biased in DID on microdata. The study does not incorporate survey weights, though they turn out to be endogenous, meaning that unweighted regressions are inconsistent (Hausman and Wise 1981). The instruments prove weak, which calls for weak identification–robust inference—an issue better understood now than in 2001. About a tenth of the observations of the Inpres treatment indicator contain transcription errors, which in a few cases generate extreme values.

Modifying data and methods to address the micro comments does not greatly shift the point estimates in Duflo (2001). But it does cast the study's confidence intervals as undersized, especially for returns to schooling. In a representative reduced-form specification, Duflo (2001, Table 4, panel A, column 4) estimates that a unit of treatment (another school per 1,000 children)

increased log hourly wages by 0.0147 ($p = 0.04$). Even though data corrections lift the point estimate to 0.0187, the p value rises to 0.08 after clustering the variance estimate by geographic unit. Additionally incorporating survey weights shifts the point estimate to 0.0167 and increases the p value to 0.29. The story is similar for a related two-stage least squares (2SLS) specification except that since 2SLS estimates are ratios of reduced-form estimates, uncertainty in the denominator—the impact of Inpres SD on schooling attainment—destabilizes the overall result. Duflo (2001, Table 7, panel A1, column 1) puts the marginal return to a year of schooling at 0.075 log points in the hourly wage of wage workers ($p = 0.03$), 95% confidence interval [0.01, 0.14]). In the revised regression, weak identification–robust inference raises the p value to 0.09 and widens the confidence interval to [–0.03, 0.32]. Weighting leads to a preferred point estimate of 0.108, raises the p value to 0.31, and greatly expands the confidence range, to [–0.58, 0.97].

The macro-level comment is that the Duflo (2001) identification strategy does not remove endogeneity as securely as would be hoped. To the extent that OLS estimates of the impact of schooling on labor market outcomes are biased in this data set, that bias can enter the Duflo (2001) results, and not only through weak identification in 2SLS. The mechanism generating this potential bias has three components. The first is a pattern known at least since Mincer (1958) and found in the Indonesia data. In their youth, more-educated workers earn a small multiple of what their less-educated peers earn. Within a cross-section of workers observed at the same time, the ratio rises with age. Mincer (1974) explains this wage scale dilation as an artifact of diminishing returns to experience: more-educated people enter the workforce later, so for them diminishing returns to experience set in later.

The second component of the bias story is the design of Duflo (2001). The study applies DID to data from a follow-up survey fielded in 1995, when the subjects were aged 23–45. To give this cross-section the two dimensions needed for DID, respondents are grouped by place and year of birth. Inpres treatment varies along both dimensions: in the geographic dimension, some regencies and municipalities (the second-level administrative units in Indonesia) received more schools per child; in the temporal dimension, subjects ranged from having been too old to be directly exposed to the schooling expansion to young enough to be fully exposed. When DID is brought to this structure, time's arrow runs from the pre-treatment early-born to the post-treatment late-born—from old to young. With respect to this clock, wages as observed in 1995 *fall* with time since the young earn less. And wages decline more sharply among more-educated workers because their

wage trajectories are steeper. This pattern would be expected even in the counterfactual that Inpres SD never happened. In the terminology of DID, the parallel trends assumption does not hold with respect to pre-treatment schooling level (de Chaisemartin and D'Haultfœuille 2017).

The final component of the bias story is negative selection. Duflo (2001, Table 2) documents that Inpres treatment went disproportionately to localities that were probably poorer, for their native wage-earning men completed less schooling and earned less despite any benefits of Inpres. Since, as just noted, the wage trajectories of the less-educated are shallower, we should observe smaller wage drops in high-treatment/low-education localities (again, as one moves with the study clock from older to younger men within the single survey). Subtracting the low-treatment change (large and negative) from the high-treatment one (small and negative) yields a positive difference in differences. This will enter the Duflo (2001) DID estimates as upward bias. While a common concern about causal interpretation of OLS regressions of wages on schooling is that positive selection upward-biases results, here negative selection occurs to the same effect.

For intuition, imagine an extreme hypothetical: the association of wages with schooling is *zero* among men young enough to have been exposed to Inpres schools but positive among older men. In this hypothetical, Inpres SD's apparent impact in cross-geography comparisons would be zero for the young and, because of negative selection, negative for the old. Benchmarking the first impact against the second would generate a positive difference in differences.

I pursue three strategies to combat this bias. First, I follow up on the Duflo (2001) cohorts later—in 2005, 2010, and 2013–14, as dictated by data availability. Now the subjects are closer to or well within prime working age, and, possibly as a result, less affected by age-differentiated wage scale dilation. Reduced-form impact estimates plunge. And identification in 2SLS regressions weakens further. Perhaps one cause is rising measurement error in later surveys, e.g., in recall of schooling history and of place and year of birth, which determined Inpres exposure.

The second strategy builds on the informal discussion of timing patterns in Duflo (2001) by testing for trend breaks. I do not expect the dilation of the wage scale to accelerate or decelerate *suddenly* at certain ages, so any appropriately timed kinks in time series—in the evolution with respect to age of the cross-geography associations between outcomes of interest and Inpres treatment—would indicate causation by Inpres. Through modest modifications of the Duflo (2001) set-up, the testing introduces a piecewise-linear representation of trends. Surprisingly, it does not strongly confirm an impact of Inpres on schooling attainment, though it does more firmly support

an impact on primary school completion. Again while reasonable support is found for a reduced-form impact on hourly wages of wage workers in 1995, this result does not persist reliably in the later follow-ups. And these results fall to an even more aggressive test, the inclusion of a quadratic time control. The data thus confirm changes in the slopes of some trends in the cross-section with respect to age, but cannot confidently judge whether the bends are sharp, as would be expected after a shock to school construction, or gradual, as would be more likely if other forces dominated.

Finally, I apply the quantile-based changes-in-changes estimator (CIC; Athey and Imbens 2006), which does not assume parallel trends for causal interpretation. Most of the CIC estimates of wage impacts are indistinguishable from zero.

Section II of this paper reviews the methods of Duflo (2001). Section III comments on the data and methodology and incorporates those comments into key regressions in the original. Section IV documents age-differentiated wage scale dilation in Indonesia. Section V follows up later in life. Section VI tests for trend breaks. Section VII applies CIC. Section VIII concludes.

II. The Duflo (2001) specifications

Legal authority to carry out the Inpres SD program flowed from annual presidential instructions, appendices of which list how many schools were to be built in each regency or municipality—administrative units that I will 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. Duflo (2001, note 1) cites a government finding that in this period, actual construction closely matched planned. Using data from the Intercensal Population Survey (SUPAS) of 1995, the study examines how boys from more-treated localities fared in early adulthood: how many years they ultimately stayed in school; whether they held paid employment; and, if so, how much they earned per month and hour.

The study works in the difference-in-differences framework. Classical 2×2 DID benchmarks changes between two periods in a treatment group against changes in a control group. This eliminates confounding from factors whose effects are the same across groups or through time. 2×2 DID can be performing by estimating

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

where δ is the impact parameter; Y_{ijt} is an outcome observed for individual i in 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 and T_t are dummies for $j = 1$ and $t = 1$; ν_j and η_t are place and time fixed effects; and ϵ_{ijt} is a mean-zero error process.

Duflo (2001) elaborates on this specification in a few ways. Since the outcome data come from the single, 1995 cross-section, the observations are molded into a two-dimensional structure by grouping them by year and place of birth, as in the Acemoglu and Angrist (2000) study set in the U.S. The observations of the oldest subjects constitute the pre-treatment section of the data, for Inpres SD launched too late to directly affect them. The youngest subjects generate the post-treatment data. In an intuitive sense, analytical time therefore runs backward, from older to younger people. A separate elaboration is an expansion in the numbers of groups and periods. There were some 290 regencies in Indonesia in 1995.[1] And each year-of-birth cohort constitutes a time period; the included cohorts range in age from 2 to 24 as of 1974. A final elaboration is that most regressions take treatment as continuous, not binary, as planned schools per 1,000 children.

These elaborations lead to the two-way fixed effects (TWFE) model,

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

where t and j index ages and regencies and D_j is now Inpres treatment intensity in regency j. $\widetilde{\mathbf{T}}_t$ is a vector of variables that depend on age in 1974. Duflo (2001) defines $\widetilde{\mathbf{T}}_t$ in several ways. In its least flexible form, $\widetilde{\mathbf{T}}_t$ is the T_t defined earlier, a single dummy for the younger cohorts; in this case, men aged 12–17 in 1974 form the before cohorts and those aged 2–6 the after cohorts, and only these cohorts enter the sample. I will use this set-up for all reduced-form estimates. In its most flexible form, $\widetilde{\mathbf{T}}_t$ is a full set of age dummies, and all cohorts aged 2–24 form the sample. The estimate $\widehat{\boldsymbol{\delta}}$ is then a vector of cohort-specific linear associations between Y_{ijt} and D_j, controlling for the fixed effects. An in-between option that Duflo (2001) favors makes $\widetilde{\mathbf{T}}_t$ a set of birth year dummies, except that the pre-treatment cohorts, aged 12–24 in 1974, are given one dummy, which imposes the constraint that among people too old to have been directly affected by the Y_{ijt} versus D_j association exhibit no trend with respect to age.

For the coefficient vector $\boldsymbol{\delta}$ in (2) to exactly represent causal impacts, the term it multiplies must

[1] The Duflo (2001) data set treats a handful of regencies created by subdivision between the early 1970s and 1995 as distinct clusters with the same values for the baseline variables and treatment instrument. I instead cluster by regency according to boundaries in the early 1970s, except that, for data availability reasons, I accept the consolidation of four pairs of regencies in Central Kalimantan. This choice also affects the definition of birthplace fixed effects.

be exogenous. Two specification choices in Duflo (2001) add plausibility to that possibility. The first is the addition of controls. All are products of baseline variables and age dummies, to allow the baseline variables age-specific associations with the outcome. The augmented specification reads

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

$\mathbf{T}_t$ is a full set of age 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, not in logarithms, a variable whose inclusion the text does not appear to motivate. It gains the fraction of the population enrolled in primary school, to reduce bias from reversion to the mean; and spending on a separate water and sanitation program, which might confound the effects of school construction. I will call the first variant the minimal control set and the second the intermediate.

The other design element buttressing the exogeneity of $D_j \widetilde{\mathbf{T}}_t$ is the natural-experiment character of Inpres SD. Duflo (2004b, p. 350) emphasizes that one virtue of Inpres SD as a source of identification is that its treatment rule is known: "more schools were built in places with low initial enrollment rates." However, as Duflo (2001, Table 2) documents, the primary school non-enrollment rate is only modestly correlated with treatment intensity. Indeed, after data transcription corrections discussed in section II, there is essentially no linear relation between the treatment indicator D_j and non-enrollment (see Figure 1). The correlation is 0.04 ($p = 0.53$). Regardless, knowing the treatment rule would not make it exogenous.

The greater hope for causal identification lies in the "big push" character of the Inpres funding. It began suddenly in 1973/74, soon after the oil shock of late 1973, approached its maximum in 1977/78, and only slackened after 1983–84 (Suharti 2013, p. 33).[2] Just in the first six years captured in Duflo (2001), the shock roughly doubled the stock of primary schools. This is why Duflo (2001) scans informally for trend breaks in the entries of $\boldsymbol{\delta}$ (taking $\widetilde{\mathbf{T}}_t$ as a full set of birth year dummies). And it is why I formalize the check in section VI.

A final step in the Duflo (2001) empirics is the introduction of 2SLS in order to estimate the impact of one endogenous variable, schooling, on others, including log wages. The 2SLS

[2] Duflo (2001) leaves the impression that the big push ceased after 1978/79. In fact, about half the planned construction was slated for after. The country's primary school gross enrollment rate (the ratio of enrolled children of any age to the population of children of official primary school age) did not plateau until 1987 (Suharti 2013, Figure 2.5). One reason Duflo (2001) stops in 1978/79 is that children only exposed to later-built schools would hardly have attained working age by the 1995 follow-up (Duflo 1999, note 5).

specifications use (3) for the first stage, setting Y_{ijt} to the highest education level attended, denominated in years of schooling (S_{ijt}).[3] They adapt (3) for the second stage, replacing $D_j\widetilde{\mathbf{T}}_t$ with the instrumented variable S_{ijt} and setting Y_{ijt} to an outcome of interest, namely, wage sector participation or log wages. The identifying assumption is thus that after conditioning on controls, the instruments $D_j\widetilde{\mathbf{T}}_t$ are associated with labor market outcomes only through S_{ijt}. $\widetilde{\mathbf{T}}_t$ 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 two. I call these specifications the "instrument by young/old" and "instruments by birth year" variants.

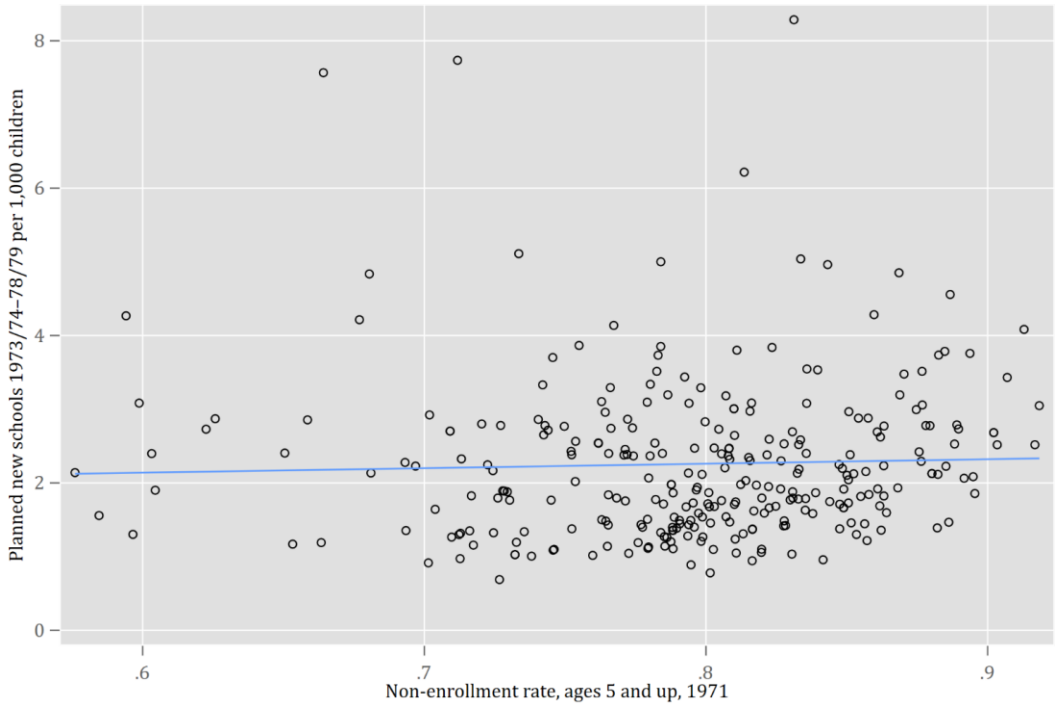

Figure 1. Inpres SD treatment vs. non-enrollment rate by regency

Notes: The plot is based on the corrected data set described in section II. Each data point corresponds to a regency or municipality. The line shows the OLS fit. The denominator for the vertical-axis variable is population aged 5–14 from BPS (1972). The denominator for the horizontal-axis variable is population of age ≥ 5, as provided in the source for the numerator (BPS 1974), where Duflo (2001) appears to take total population from BPS (1972).

III. Revisiting the Duflo (2001) regressions

A. Comments

I comment first on four technical aspects of the Duflo (2001) regressions: clustering of variance

[3] I will call this variable "years of schooling," but it is worth bearing in mind its more precise description as "highest education level attended." 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. Then, in future surveys, the former child could report a lower "highest education level attended" without any reduction in actual years of schooling.

estimates, observation weighting, weak identification, and apparent transcription errors in variables from the 1970s. The technical comments will also inform alternative specifications I introduce later in response to the concern about bias in the application of DID.

Clustering.—The variance estimates in Duflo (2001) are classical: they are not adjusted for the possibilities of heteroskedasticity and intra-regency dependence. Bertrand, Duflo, and Mullainathan (2004) demonstrates the value of the standard Liang and Zeger (1986) clustering correction for DID-based inference on microdata when treatment is constant within clusters.

Endogenous observation weights.—The 1995 SUPAS used a complex survey design, with stratification and clustering. It oversampled low-population regencies, evidently in order to produce comparably accurate statistics for all regencies. The sampling probability must have varied within regencies as well because the sampling weights included in the data sets are not constant within regencies. Most likely sampling was stratified with respect to an index of household assets using data gathered in the 1990 census. The rarer and probably more diverse high-asset households would have been oversampled for precision.[4] The code that Esther Duflo sent me documents that Duflo (2001) incorporates the survey weights in its illustrative 2×2 DID design but not in the main analysis.

One rationale for weighting observations in regressions is to make results more representative for the original setting, in this case Indonesia of 1995. This 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 policy decisions. Setting aside representativeness leaves econometric concerns, about consistency and variance of estimators. All else equal, incorporating unequal weights will increase the variance of estimators. But if the weights are endogenous, using them will remove selection on the basis of the outcome, itself a source of inconsistency (Hausman and Wise 1981; Solon, Haider, and Wooldridge 2015).

To test for the endogeneity of the weights, I add them as controls to representative regressions (Duflo 2001, Table 5, columns 5 and 8). As mentioned, in order to look for trend breaks, Duflo (2001) applies OLS to specializations of (3) in which $\widetilde{\mathbf{T}}_t$ is the full set of cohort dummies; as well,

[4] Surbakti (1995, p. 21, 31) confirms that the National Socio-Economic Survey (SUSENAS) used the same sampling frame, and stratified by "income (or expenditure)." But the 1990 census covered household assets, not money income or expenditure; see international.ipums.org/international/resources/enum_materials_pdf/enum_form_id1990a.pdf. A presentation by the Indonesian statistical agency describes stratifying the SUSENAS sampling on an asset index: sesricdiag.blob.core.windows.net/sesric-site-blob/files/INCOME&CONSUMPTION_Methodology_Susenas_EN.pdf.

$\mathbf{C}_j$ is set to the intermediate control set, and Y_{ijt} is years of schooling or the log hourly wage. Adding the weight variable to these two regressions endows it with large, cluster-robust t statistics, -6.31 and -6.33. This confirms that people with more schooling or higher wages were more likely to be sampled by SUPAS (and thus receive lower sampling weights). Since the sampling probability was not independent of the outcomes, the specifications in the main Duflo (2001) analysis, all unweighted, are inconsistent.

Weak identification.—Since 2001, econometricians have made progress in understanding, detecting, and managing the effects of weak identification in instrumented estimators (Stock and Yogo 2002, Moreira 2003, Kleibergen and Paap 2006). When identification is weak, bias associated with OLS can leak into regressions designed to reduce such bias. Finlay and Magnusson (2019)'s Monte Carlo–based review of inference methods robust to weak identification favors the Anderson-Rubin (AR, 1949) test with a wild-bootstrapped distribution for the test statistic.[5] 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 hypothesis—that the instruments are valid. Rejection on the test is therefore ambiguous: it can indicate rejection of certain coefficient values *or* rejection of instrument validity (Roodman et al. 2019, p. 30). As will be explained, that ambiguity will make the critical conclusions of this reanalysis conservative.

Transcription errors.—Duflo (2001) gathers early-1970s regency-level variables from government sources. For example, from reports on the 1971 census come total population and population aged 5–14 by regency. Both of those are used as denominators for other variables. The latter also enters regressions as a control, being the sole entry in $\mathbf{C}_j$ in the minimal control set.

I reconstruct all the 1970s variables but the water and sanitation spending variable.[6] In all the variables examined, I find some discrepancies between the Duflo (2001) figures and my sources (BPS 1972, 1974; Bappenas 1973–78). The apparent transcription errors affect the Inpres treatment indicator, planned schools per 1,000 children, for about 10% of regencies. For example, perhaps because regencies are listed in different orders in different publications, a few pairs or

[5] Simulations 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.

[6] Page scans of primary sources and an annotated Excel tabulation are at github.com/droodman/Duflo-2001.

quadruplets have their numbers rearranged. When taking ratios of right-skewed variables, transcription errors tend somewhat to generate extreme values, unless taken in logarithms, as when dividing a numerator from the United States by a denominator from Canada. The original and corrected versions of the treatment intensity indicator D_j are correlated 0.81 at the regency level.[7]

Heterogeneous treatment effects.—Recent scholarship (de Chaisemartin and D'Haultfœuille 2020; Goodman-Bacon 2021) has highlighted the counterintuitive way that effect heterogeneity can bias TWFE estimates of mean effects. This methodological critique does *not* appear to apply strongly to Duflo (2001). Effect heterogeneity has the most scope to bias results when treatment is staggered across units. But Inpres SD launched and scaled on about the same schedule in all regencies. Moreover, a check for heterogeneity proposed by Jakiela (2021) produces little evidence of such. See appendix C.

B. OLS/reduced-form results

To check the import of these technical comments, I revise some key regressions in Duflo (2001). I begin with OLS regressions based on (3) that are reported in the paper's Table 4. The regressor of interest is $D_j T_t$, where D_j is regency-level Inpres treatment intensity and T_t is the young/old dummy. Samples are restricted to the young and the old, i.e., men who were aged 2–6 and 12–17 in 1974. When the dependent variable is a schooling outcome, these regressions can be seen as the first stage of the "by young/old" 2SLS regressions to follow. When the dependent variable is a labor market outcome, they are reduced-form regressions. For parsimony, I focus on regressions with the minimal control set. This does not materially affect conclusions.

The new results—after clustering standard errors, weighting observations, and correcting data errors—appear in Table 1. Clustering and weighting each reduce apparent precision. For example, the standard error of 0.00729 in the unweighted reduced-form log hourly wage regression (Duflo 2001, Table 4, panel A) becomes 0.0107 after data corrections and clustering, and 0.0157 with weighting as well.

As a validity check, Duflo (2001) runs a placebo test on this specification, by shifting the comparison from ages 12–17 and 2–6 to 18–24 and 12–17. Neither of the older cohort ranges should

[7] A particular complication affects baseline school attendance. Duflo (2001) appears to take the numerator, people aged 5 and older attending school, from BPS (1974), 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 (1974), perhaps because school attendance was only queried for a subpopulation. For consistency, I therefore use the denominator that accompanies the numerator in BPS (1974), namely, total population aged 5 and above.

have been directly affected by Inpres SD. In the original, the check reassures: the experiment of interest increases schooling and wages while the placebo test does not. Here, I emulate the validity check, and formalize it with a two-tailed Wald test for the null hypothesis that the actual and placebo experiments return the same result. See the last row of Table 1.

Despite the lower (apparent) precision in the new results, there remains little doubt that regencies that received more schools saw bigger increases in schooling; there can be more doubt with regard to wages. In the weighted specification, each unit increment in treatment intensity—another planned school per 1,000 children—increased the probability of completing primary school by 3.45 percentage points, against only 0.41 points in the placebo experiment. The placebo result is indistinguishable from zero and quite distinguishable from the experimental effect ($p = 0.01$, reported in the last row, second column of Table 1). About the same can said for years of schooling, wage sector participation, and log hourly wages, but with progressively less confidence. In the weighted log hourly wage regression, the estimated impact of 0.0167 log points differs from zero at $p = 0.29$ and from the placebo effect at $p = 0.36$.

Table 1. Reduced-form TWFE impact estimates in 1995 follow-up

	Primary school completion		Years of schooling		Wage sector participation		Log hourly wage	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
Experiment	0.0247	0.0345	0.159	0.153	0.0207	0.0172	0.0187	0.0167
	(0.00763)	(0.00785)	(0.0581)	(0.0745)	(0.00439)	(0.00633)	(0.0107)	(0.0157)
Observations	31,061	31,061	31,061	31,061	78,470	78,470	31,061	31,061
Placebo	0.00153	0.00408	0.0130	–0.0275	0.000678	0.00536	–0.00280	–0.00351
	(0.00538)	(0.00747)	(0.0614)	(0.0743)	(0.00368)	(0.00490)	(0.00825)	(0.0116)
Observations	30,458	30,458	30,458	30,458	78,488	78,488	30,458	30,458
Experiment = placebo (p)	0.04	0.01	0.14	0.13	0.00	0.19	0.15	0.36

Notes: Experiment samples are restricted to men aged 2–6 or 12–17 in 1974 and placebo samples to ages 12–17 and 18–24. Except for the wage sector participation regressions, samples are also restricted to observations with non-zero wages. Reported coefficient estimates are for $D_j T_t$ where D_j is Inpres schools per 1,000 children and T_t is a dummy for the younger block of cohorts in each sample. All regressions control for year- and regency-of-birth dummies as well as interactions between birth year dummies and number of children aged 5–14. Standard errors are in parentheses, clustered by birth regency.

C. Returns to schooling

Duflo (2001, Table 7) estimates the return to schooling with OLS and 2SLS regressions; their revised counterparts appear here in Table 2. The table's three panels are for outcomes observed in 1995: wage sector participation, the log of monthly wage earnings for participants, and the log hourly wage. OLS regressions are reported on the right. 2SLS regressions "instrumented by young/old" appear on the left; their first stages are the schooling regressions shown in Table 1.

Next come results from overidentified 2SLS, "instrumented by birth year"; recall from section II that these represent the pre-treatment ages 12–24 with a single dummy and the younger ages with individual dummies.

Correcting data and weighting observations hardly affects the OLS estimates of the effects of a year of schooling—here, 2.9 percentage points for wage sector participation, 0.071 log points for monthly wages, and 0.077 log points for hourly wages. Clustering doubles or triples standard errors, which still leaves the estimates rather precise; e.g., that for the weighted hourly wage result is 0.0018.

The revisions affect the 2SLS estimates in more complex ways. As in Duflo (2001, Table 7), an instrument validity check for the overidentified regressions reassures, here through high *p* values on the Hansen (1982) test. This finding reduces the worry about endogenous causation stories. But the reassurance is partial since the test can generate false positives if most or all instruments are invalid. Essentially, the test is premised on enough instruments being strong and valid enough that the second-stage residuals reasonably represent the structural error.

In contrast with the original, for all 2SLS regressions, here I report the Kleibergen-Paap (KP, 2006) measure of instrument strength. Since there is only one instrumented variable, the KP measure is equivalent to an *F* statistic for the hypothesis that all coefficients on the instruments in the first stage are zero (Baum, Schaffer, and Stillman 2007, pp. 21–22). Instrument weakness is unexpectedly prevalent, especially in the overidentified, instruments-by-birth-year regressions. Only the exactly identified regressions, instrumented by young/old, earn KP *F* statistics above 5, where 10 is often taken as a rule-of-thumb minimum. And that only happens when observations are not weighted to adjust for endogenous sampling.[8]

In view of the weak identification, Table 2 reports 95% confidence sets based on the wild-bootstrapped, Anderson-Rubin test. These sets tend to be disjoint or unbounded when the KP *F* statistic is low. Still, a confidence interval of "(–∞, ∞)," which appears several times, does not mean that the regression extracts *no* information. To provide insight into the confidence sets, Figure 2 arrays the associated confidence curves as functions of trial values for the impact of schooling. For each trial value, the test statistic's distribution is simulated using the Wild Restricted Efficient bootstrap

[8] It is counterintuitive that adding instruments weakens their collective strength. The source 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 undesirable and generates bias toward OLS. Evidently here moving to the larger set of instruments, by birth year, increases the undesirable factor more.

(Davidson and MacKinnon 2010).[9] The p value extracted from the bootstrap is then plotted against the trial coefficient. The curve's intersections with $p = 0.05$ are pinpointed where possible to demarcate 95% confidence sets.

The confidence curves show that the Anderson-Rubin tests favor the view that the true coefficient on schooling in the wage regressions is positive. Even where the p value is above 0.05 for large ranges of negative return rates—failing to reject them with 95% confidence—the curve is usually lower there than for positive rates. This inference in favor of a positive impact is of course much weaker than in the original. In my preferred specification for the log hourly wage—exactly identified for instrument strength, weighted for consistency—the 95% confidence interval is [–0.58, 0.97] (last row, second column of Figure 2). The bootstrapped p value for the hypothesis of zero impact is 0.31, as compared to 0.03 in the original.

Adding the extra controls used in Duflo (2001)—variables based on baseline enrollment rates and water and sanitation spending—further lifts confidence curves (results not shown). Moreover, as noted earlier, the Anderson-Rubin procedure jointly tests two assumptions. Possibly some low p values for trial coefficients outside the ranges depicted in Figure 2 arise only from rejection of the instrument validity assumption, not the trial parameter values. In the skeptical frame of this reanalysis, that possibility makes even the wide confidence sets conservative.

To boost instrument strength, and to borrow from Duflo (2004a), I change the treatment variable from years of schooling to the primary school completion dummy. The new variable is a more proximate consequence of the Inpres primary school construction program, though potentially more distal from labor market outcomes. Table A-1 in the appendix shows the results. Since log monthly wages and log hourly wages are highly correlated, I follow Duflo (2001) in focusing on the latter. KP F statistics are much higher now, indicating stronger identification. The preferred regression (column 4) puts the impact of primary school completion on wage sector participation at 44 percentage points (95% confidence interval [13%, 80%]). The estimate of the impact on wages remains imprecise, if still positive, with a point estimate of 0.334 and confidence interval of $[-0.60, \infty)$.

[9] 99,999 replications are performed for each test. Auxiliary weights are Rademacher-distributed, and drawn at the regency level. See Roodman et al. (2019).

Table 2. OLS and 2SLS estimates of the effect of years of schooling on labor market outcomes, 1995

	OLS		2SLS: instrument by young/old		2SLS: instruments by birth year	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
	(1)	(2)	(3)	(4)	(5)	(6)
A. Full sample: participation in wage sector						
Coefficient	0.0328	0.0292	0.173	0.222	0.105	0.0954
	(0.000919)	(0.00109)	(0.0648)	(0.157)	(0.0296)	(0.0371)
95% confidence set			[0.09, 0.58]	(−∞, −0.44] ∪ [0.06, ∞)	[0.05, ∞)	(−∞, ∞)
Overidentification *p*					0.27	0.49
KP *F*			7.94	2.06	1.96	1.00
Observations	152,989	152,989	78,470	78,470	152,989	152,989
B. Wage earners: log monthly wages						
Coefficient	0.0697	0.0709	0.101	0.0801	0.119	0.160
	(0.00150)	(0.00191)	(0.0641)	(0.0910)	(0.0525)	(0.0537)
95% confidence set			[−0.06, 0.39]	(−∞, ∞)	(−∞, ∞)	(−∞, −0.37] ∪ [0.04, ∞)
Overidentification *p*					0.82	0.82
KP *F*			6.83	3.55	1.00	1.43
Observations	61,136	61,136	31,310	31,310	61,136	61,136
C. Wage earners: log hourly wage						
Coefficient	0.0774	0.0768	0.118	0.109	0.106	0.135
	(0.00140)	(0.00175)	(0.0580)	(0.0909)	(0.0476)	(0.0495)
95% confidence set			[−0.03, 0.32]	[−0.58, 0.97]	(−∞, ∞)	(−∞, −0.43] ∪ [−0.01, ∞)
Overidentification *p*					0.70	0.43
KP *F*			7.48	4.20	1.08	1.57
Observations	60,663	60,663	31,061	31,061	60,663	60,663

Notes: Each panel shows results for a different dependent variable. Coefficients are for years of schooling. Instruments "by birth year" are interactions between D_j and dummies for ages 2–24 as of 1974, except that age ≥12 defines one category. Young/old regressions are restricted to ages 2–6, 12–17. All regressions control for year- and regency-of-birth dummies as well as interactions between birth year dummies and number of children aged 5–14 in a regency in 1971. Standard errors are in parentheses, clustered by birth regency. Confidence sets are wild-bootstrapped with 99,999 replications, from a Wald test for the reduced-form regressions and Anderson-Rubin for the instrumented. Overidentification *p* is from the Hansen test for instrument validity. "KP *F*" is the Kleibergen-Paap measure of instrument strength.

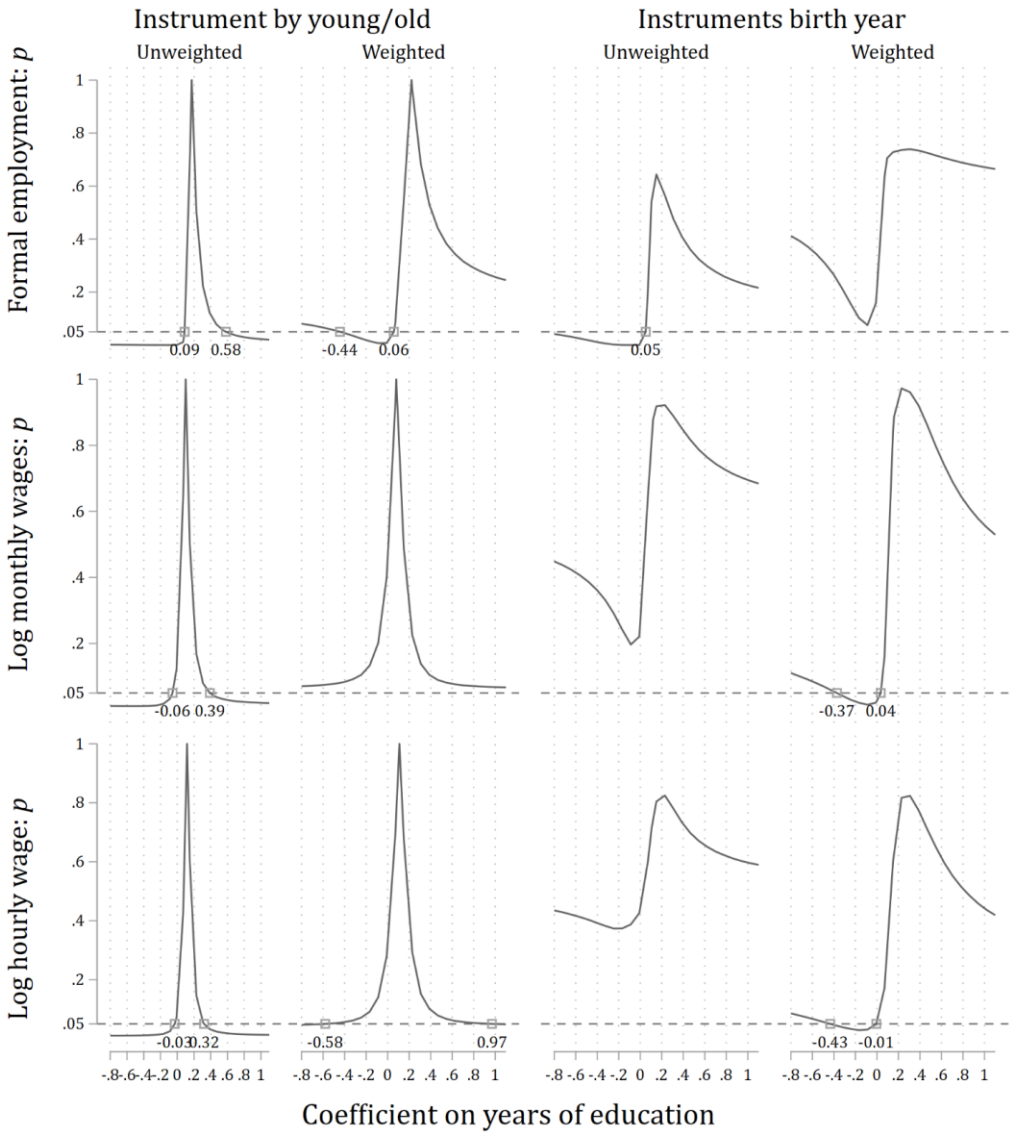

Figure 2. Confidence curves from wild-bootstrapped Anderson-Rubin tests of the coefficient on years of schooling in 2SLS regressions in Table 2

Notes: Each plot in this figure shows, in the context of a specific log hourly wage regression, the p value for the null that the instruments 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 2SLS results in the right two-thirds of Table 2.

IV. Bias from age-differentiated wage scale dilation

An important and common pattern is present in the Indonesia data involving wages, ages, and schooling. In their early 20s, educated wage workers earn a small multiple of what their less-educated peers earn. But, within a cross-section such as the 1995 SUPAS data set in Duflo (2001), the multiple rises with age. Figure 3 demonstrates. Each contour is a local polynomial–smoothed

fit to the association between the log hourly wage and age in a particular schooling stratum in 1995. All regressions are restricted to wage-earning men and incorporate survey weights. As one scans the figure from left to right, the wage scale dilates.

Mincer (1974) observes a similar pattern in U.S. data from 1960 and connects it to a theory of causation from schooling to earnings. The result is the standard Mincer labor function, which has the log wage linear in years of schooling, and quadratic in years of work experience, with diminishing returns. The wage scale widens because more-educated people are later to enter the workforce, later to start accumulating experience, and later to hit the diminishing returns to experience.

The Mincer model has been critiqued and elaborated on empirical and theoretical grounds. As for empirics, Lemieux (2006, p. 127) concludes "that the Mincer equation remains an accurate benchmark…provided that it is adjusted by….including a quartic function in potential experience [and] allowing for a quadratic term in years of schooling." We should therefore expect that the Inpres shock plays out against a background of broad trends of higher polynomial order than is implied by the standard Mincer labor function.[10] As for theory, a central concern in the measurement of returns to schooling is that the causal relationship between schooling and wages is complex, so that that reverse- or third-variable causation biases results from OLS estimation of the Mincer model. If such bias is present, a 2SLS specification would tend to absorb it to the extent that *it* is biased toward OLS.

The empirical fact of age-differentiated wage scale dilation provides the starting point of a plausible story for bias in DID as conducted in Duflo (2001). Table 3, panel A, of Duflo (2001) documents that regencies that received higher treatment produced (wage-earning) men with less schooling on average: 8.02 versus 9.40 years of schooling among those aged 12–17 in 1974 and 8.49 versus 9.76 for those aged 2–6. This negative selection causes natives of high-treatment areas to congregate disproportionately near the lower curves in Figure 3. Within this group, the before-after contrast in wages—moving from older to younger cohorts—should be relatively small, as the lower curves are flatter. Among workers from less-treated, more-schooled regencies, the change should be larger, as their wage curves slope downward more as one scans from right to left. The difference in those differences—the small, negative change for high-treatment regencies minus the

[10] The standard Mincer model is $\ln w = \ln w_0 + rS + \beta_1 X + \beta_2 X^2$ where w is earnings, w_0 is a constant, S is schooling, X is potential experience, and $r, \beta_1 > 0, \beta_2 < 0$. Indonesians typically start school at age 7, so we estimate $X = A - S - 7$, where A is age. The return to schooling is $\partial \ln w/\partial S = r - \beta_1 S - 2\beta_2(A - S - 7)$. The evolution of that with respect to A is characterized by $\partial(\partial \ln w/\partial S)/\partial A = -2\beta_2$, which is constant: returns to schooling rise linearly with age. Introducing higher-order powers of S and X into the wage equation as advocated by Lemieux (2006) breaks this linearity.

larger, negative change for low-treatment ones—is positive. It should be expected to produce positive DID estimates in the counterfactual of no Inpres SD treatment. de Chaisemartin and D'Haultfœuille (2017), supplement 3.3.1, makes essentially this point, starting from the Mincer model and equating treatment to schooling attainment rather than Inpres treatment intensity. Here, the bias story is extended by linking from schooling through negative selection back to the Inpres intensity indicator.

If age-differentiated wage scale dilation is primarily a feature of youth, then following up on the Duflo (2001) cohorts later in life would reduce its biasing influence. To explore this possibility, Figure 4 plots the wage scale at each age, defined as the slope estimate from an age-specific OLS regression of the log hourly wage on schooling. The age bounds of the Duflo (2001) sample—2 and 24 in 1974—are marked. Wage scale dilation proceeds steadily within the Duflo (2001) sample, perhaps slowing at the older end. It stops and then reverses among even older men. Deferring follow-up can be visualized as shifting the study window marked in the figure to the right. *If* the wage scale follows the same pattern in later cross-sections, then delaying follow-up would reduce or even flip the bias in DID.

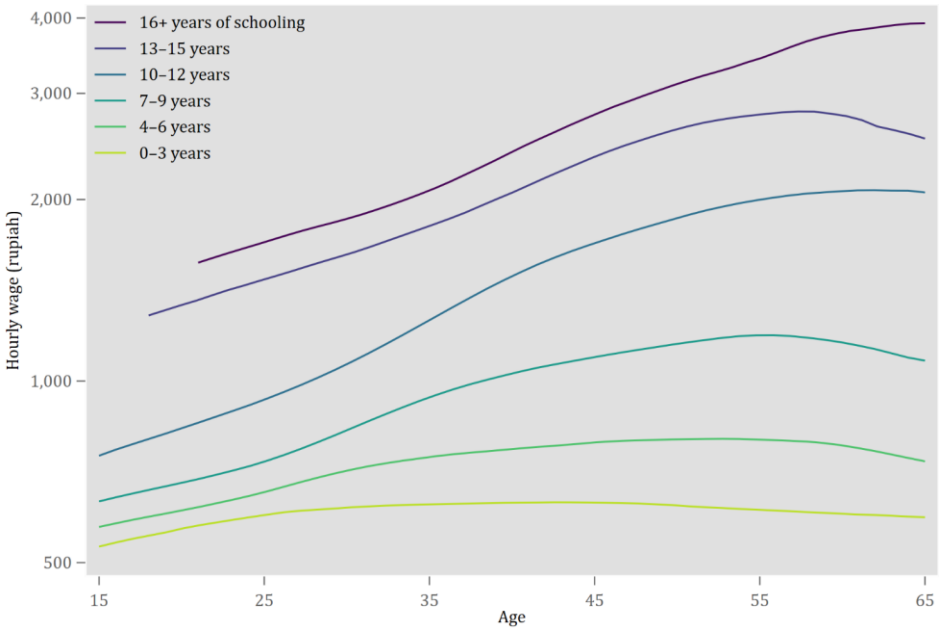

Figure 3. Smoothed fits of hourly wage to age, by years of schooling, wage-earning Indonesian men in 1995

Notes: Survey weights are incorporated.

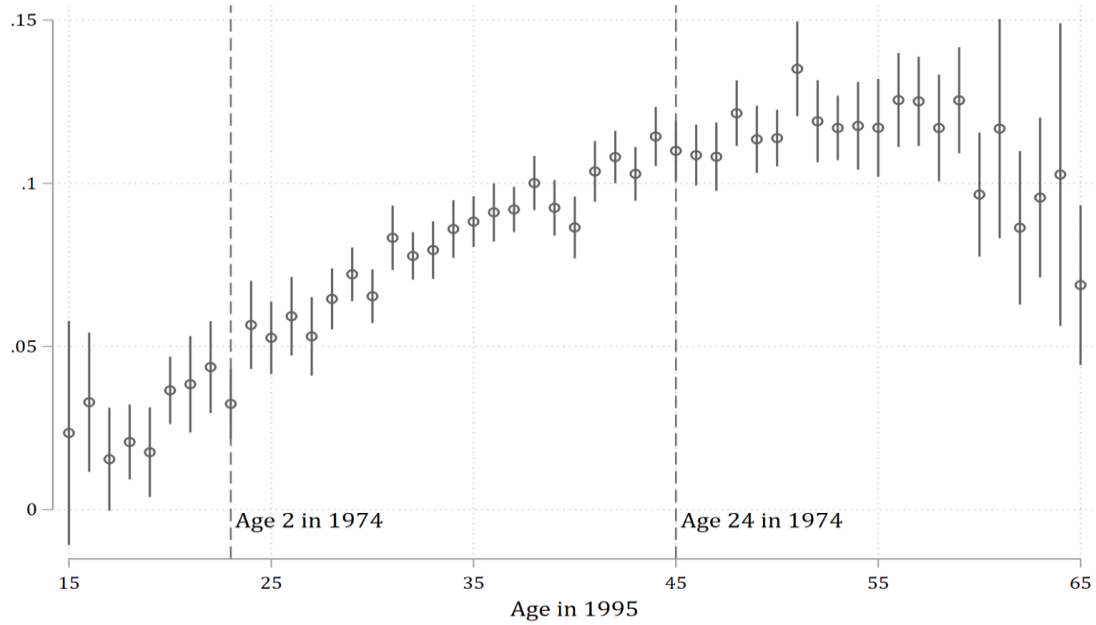

Figure 4. Wage scale dilation by age among Indonesian men in 1995

Notes: This graph shows the slope of the linear association between the log hourly wage and years of schooling for men at each year of age in the 1995 data. Survey weights are incorporated. The plotted 95% confidence intervals are clustered by birth regency. The slope rises steadily within the range examined in Duflo (2001): ages 2–24 in 1974. This wage scale dilation undercuts the parallel trends assumption needed for causal interpretation of difference-in-differences results. It wanes and even reverses for older workers, on the right.

V. Following up later

The Indonesian statistical agency has long conducted an array of national surveys, with distinct foci such as labor and welfare, and questionnaires that have changed over time. The choice of follow-ups used here is constrained by which rounds of which surveys asked the requisite questions, and by the availability of the answers—the data—to researchers. The constraints are tight: no available survey matches the 1995 SUPAS in capturing regency of birth and current earnings from the wage sector. Facing imperfect options, I follow up using:

- *The 2005 intercensal survey (SUPAS)* data hosted by the IPUMS project (MPC 2020), which provides regency of birth and wage sector participation but not wage earnings. I impute wages in 2005 with a model calibrated to the 1995 SUPAS data. The model consists of a weighted OLS regression of reported log hourly wage on interactions between, on the one hand, dummies for birth regency, occupation, industry, and urban residence and, on the other, powers of order 0–4 in age.
- *The 2010 labor survey (SAKERNAS)*, which covers wage sector participation and wages, but not regency of birth. Following Duflo (2004a)'s work with earlier SAKERNAS rounds, I use regency of work instead of birth. Since the linkage between the two is eroded by

migration and commuting, I copy Duflo (2004a) in excluding the regencies of the Jakarta megalopolis.

- *The 2013 and 2014 household socioeconomic survey (SUSENAS)*, which provides regency of birth and net earnings from all sources, including self-employment. The survey asks for the typical monthly earnings in a person's main line of work, regardless of whether a respondent has earned that rate recently. Duflo (2001, Table 7, panel B2) uses 1993 SUSENAS earnings data in a robustness test. Here, data from the 2013 and 2014 rounds are used.[11]

I first check whether following up later indeed reduces age-differentiated wage scale dilation. The phenomenon consists in age interacting with schooling to predict wages. So to check for it, I regress the preferred earnings variable from each survey on schooling, age, and their product. To maximize relevance to the main specifications, I emulate them by clustering, weighting, and including the same controls. And I restrict once more to ages 2–6 and 12–17 in 1974. The estimated coefficients on the interaction term are presented graphically in leftmost pane of Figure 5. The coefficient is significantly above zero in all follow-ups but drops by about half after 1995. Age-differentiated wage scale dilation declines but does not disappear.

The OLS/reduced-form results from the later follow-ups are presented graphically in the rest of Figure 5. As in previous reduced-form regressions, the displayed coefficients apply to $D_j T_t$. The upper plots of the figure present the mystifying result that the impact of school construction on schooling attainment declines with age *within the same birth year cohorts*. Possibly the passage of time adds noise to the recollection of place and year of birth, which determine Inpres treatment intensity. Such noise would generate attenuation bias.

Impact estimates also fall for labor market outcomes. These drops are less consistent with attenuation bias in that the estimated impact of labor participation in 2005 has the same magnitude as that for 1995, but opposite sign. The changes across surveys in estimated labor market impacts *do* correspond well with the changes in wage scale dilation. All three plots along the bottom of Figure 5 document a plunge between 1995 and 2005 and a partial recovery in 2010. In the 2013–14 follow-up, which provides information only on log wages, the estimated impact again moves in same direction as wage dilation—downward—though only slightly (compare the green and orange spikes at the bottom of the figure). The parallel zigzagging between wage scale dilation and the reduced-form impact estimates suggests that the former is a major source of the Duflo (2001)

[11] Hsiao (2022) also uses the 2011 and 2012 editions, but they are not available at this writing.

results. Still, that the estimated impacts on schooling outcomes evolve in a somewhat similar way, which the wage-scale dilation story does not predict, makes the theory less than airtight.

The lack of clear impacts on schooling in later follow-ups further weakens the first stages of the 2SLS regressions in the later follow-ups. See Table 3. The bulk of this table reports regression from one follow-up at a time, as before. The final panel consolidates the three post-1995 earnings samples, for a higher-powered replication outside the original 1995 sample. The consolidation is blunt in that the clustering regency and dependent variable are defined within each sample just as in the survey-specific follow-ups. The clustering regency is place of birth in 2005 and 2013–14, place of work in 2010; the dependent variable is imputed log hourly wage in 2005, log hourly wage in 2010, and log typical monthly wages in 2013–14. Of the 24 KP F statistics in the table, only one surpasses 4, which itself is considered low. As before, switching the treatment variable to primary school completion strengthens identification but still generates imprecise results (Table A-2).

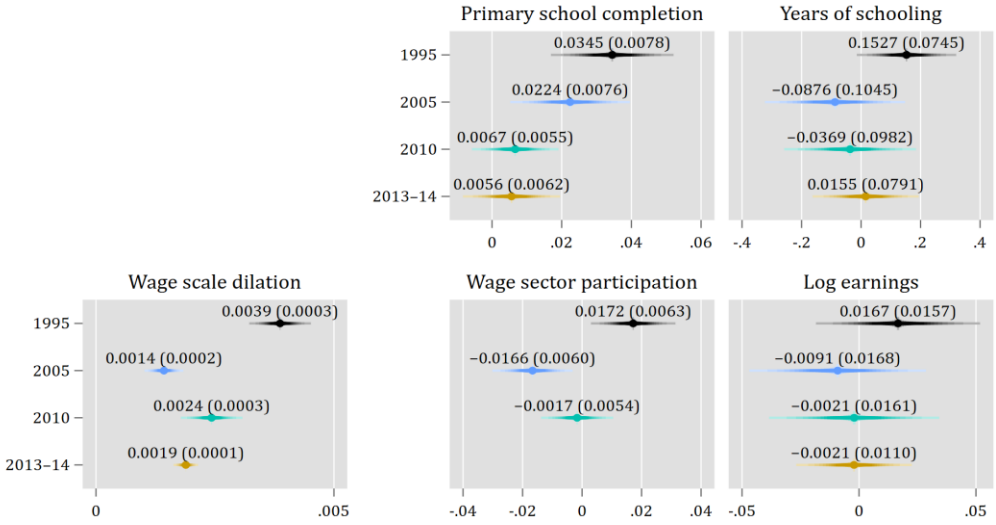

Figure 5. Age-differentiation in wage scale dilation and reduced form impact estimates by follow-up

Notes: The leftmost pane plots estimates of the coefficient on years of schooling × age in a log wage regression, controlling for schooling and age. Point estimates are labeled, along with standard errors in parenthesis. Horizontal spikes represent 95% confidence intervals. The other panes show estimates of coefficients on $D_j T_t$ where D_j is Inpres treatment intensity in birth regency and T_t is a dummy for being age 2–6 in 1974. Underlying regressions include dummies for birth year and birth regency, and interactions between birth regency dummies and number of children aged 5–14 in a regency in 1971. "Log earnings" is the log hourly wage in 1995 and 2010, the imputed log hourly wage in 2005, and log of typical monthly earnings in 2013–14. Samples are men aged 2–6 and 12–17 in 1974. Estimates incorporate survey weights and cluster by birth regency.

Table 3. Estimates of the effect of years of schooling on labor market outcomes, post-1995

	OLS		2SLS: instrument by young/old		2SLS: instruments by birth year	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
	(1)	(2)	(3)	(4)	(5)	(6)
A. Full sample, 2005: participation in wage sector						
Coefficient	0.0410	0.0413	–0.126	–0.519	–0.0247	–0.0215
	(0.000552)	(0.000654)	(0.198)	(0.898)	(0.0302)	(0.0253)
95% confidence set			(−∞, ∞)	(−∞, −0.06] ∪ [0.24, ∞)	(−∞, −0.00] ∪ [0.23, ∞)	(−∞, −0.08] ∪ [0.39, ∞)
Overidentification *p*					0.05	0.02
KP *F*			0.72	0.39	1.27	1.84
Observations	159,205	159,205	83,927	83,927	159,205	159,205
B. Wage earners, 2005: imputed log hourly wage						
Coefficient	0.0581	0.0571	0.259	0.104	0.0459	0.0480
	(0.000889)	(0.00104)	(0.939)	(0.166)	(0.0475)	(0.0324)
95% confidence set			(−∞, ∞)	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)
Overidentification *p*					0.82	0.57
KP *F*			0.06	0.71	1.18	0.95
Observations	40,264	40,264	22,074	22,074	40,263	40,263
C. Full sample, 2010: participation in wage sector						
Coefficient	0.0380	0.0372	0.0739	–0.0240	0.0913	0.0434
	(0.000573)	(0.000610)	(0.0447)	(0.0832)	(0.0279)	(0.0310)
95% confidence set			[−0.02, 1.09]	(−∞, ∞)	(−∞, −0.46] ∪ [−0.00, ∞)	(−∞, −0.26] ∪ [−0.03, ∞)
Overidentification *p*					0.67	0.32
KP *F*			4.39	2.19	1.92	1.35
Observations	153,864	153,864	83,190	83,190	153,864	153,864
D. Wage earners, 2010: log hourly wage						
Coefficient	0.102	0.105	0.519	0.0559	0.174	0.0729
	(0.00150)	(0.00223)	(0.709)	(0.358)	(0.0480)	(0.0466)
95% confidence set			(−∞, ∞)	(−∞, ∞)	(−∞, −0.16] ∪ [0.14, ∞)	(−∞, 0.05] ∪ [0.24, ∞)
Overidentification *p*					0.29	0.14
KP *F*			0.41	0.14	1.23	0.99
Observations	35,992	35,992	21,435	21,435	35,992	35,992
E. Income earners, 2013–14: log typical monthly earnings						
Coefficient	0.0499	0.0523	–0.00937	–0.137	0.0236	0.0956
	(0.000827)	(0.000808)	(0.0653)	(1.127)	(0.0294)	(0.0532)
95% confidence set			(−∞, ∞)	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)
Overidentification *p*					0.30	0.63
KP *F*			3.14	0.04	1.61	0.86
Observations	221,041	221,041	125,047	125,047	221,041	221,041
F. Combined B, D, and E						
Coefficient	0.0562	0.0583	0.0203	0.232	0.0672	0.125
	(0.000790)	(0.000717)	(0.102)	(0.811)	(0.0348)	(0.0537)
95% confidence set			(−∞, ∞)	(−∞, ∞)	(−∞, −0.05] ∪ [−0.00, ∞)	(−∞, −0.17] ∪ [0.02, ∞)
Overidentification *p*					0.32	0.62
KP *F*			1.54	0.09	1.23	1.07
Observations	297,297	297,297	168,557	168,557	297,297	297,297

Notes: Notes to Table 2 apply. Each panel shows results from a different combination of dependent variable and survey. Regressions in panel F consolidate the post-1995 earnings samples and add year dummies as controls.

VI. Testing for trend breaks

The second strategy for reducing bias from age-differentiated wage scale dilation is to revise the Duflo (2001) specifications to focus more sharply on the timing of the schooling supply shock.

Even if wage scale dilation plays a role in the impact estimates in Duflo (2001), there is little reason to think that it could explain trend breaks. We should not expect the wage scale to dilate *suddenly* at some age that coincides with the onset of Inpres SD.

Duflo (2001) analyzes timing somewhat informally. It begins by estimating (3) under the choices that Y_{ijt} is schooling, $\widetilde{\mathbf{T}}_t$ is a full set of birth cohort dummies, and $\mathbf{C}_j$ holds the intermediate control set. The paper's Figure 1 plots the coefficient estimates $\widehat{\boldsymbol{\delta}}$ much as in Figure 4 above. Each estimate quantifies the cohort-specific linear association between Inpres treatment intensity in the 1970s and schooling attainment as reported in 1995. The text observes that the "coefficients fluctuate around 0 until age 12 and start increasing after age 12….As expected, the program had no effect on the schooling of cohorts not exposed to it, and it had a positive effect on the schooling of younger cohorts" (Duflo 2001, p. 801). To my eyes, any trend break in that figure is small enough relative to the noise that a straight line might fit nearly as well as a kinked one, especially after one clusters standard errors.

The most rigorous way to test for a trend break at age 12 would be to apply a regression kink design, which would fit a polynomial in age to data ranges on either side of the proposed kink point, and allow both a jump and a slope change. However, the coarse quantization of the running variable, years of age, and the noise in the cohort-specific coefficient estimates in Duflo (2001), Figure 1, indicate that such a method would have little power. Moreover, it appears reasonable to assume, as Duflo (2001) implies, that there would be no jump. Because of the multiyear ramp-up, expected Inpres exposure effectively depends continuously on age.

I therefore think locally while fitting globally. I fit a piecewise-linear contour to the evolution of all entries in $\widehat{\boldsymbol{\delta}}$ from age 2 to age 24. The left pane of Figure 6 illustrates the idea by superimposing this fitted model on the same coefficient estimates and confidence intervals as in the Duflo (2001) figure. To construct the model, I modify $\widetilde{\mathbf{T}}_t$ in (3). It becomes a vector with two terms: a time trend t and a spline term that emerges at age 12:

$$\max(0, 12 - t) \tag{4}$$

where t is age in 1974. Recall from (3) that the regressors of interest are $D_j\widetilde{\mathbf{T}}_t$. In OLS/reduced-form regressions, both components of $D_j\widetilde{\mathbf{T}}_t$ enter as controls. In 2SLS, the first enters as a control and the second instruments schooling.

This set-up is not above debate. On the one hand, the particular modeling choice is arbitrary at

the margin and can be read as overly demanding of the data. The true and unknown Inpres impact trends did not turn cleanly and exactly at age 12. Different schools opened at different times. Some 13-year-olds of 1974, here cast as pre-treatment subjects, attended the new schools.[12] On the other, choosing a simple framework that is loosely, implicitly preregistered by the original text's description of a trend emerging around age 12 reduces the scope for multiple model testing, which would invalidate inference. On balance, I think it best to choose one framework that is reasonable and transparent.

In the kinked model fit depicted in the left pane of Figure 6, the slope is estimated to jump by 0.0094 at age 12. Following that steeper trend forward brings us to men aged 2–6 in 1974, which is the treatment cohort in many Duflo (2001) specifications. The deviation from pre-trend implies that the two-to-six cohort, spent an average of $8 \times 0.0094 = 0.075$ additional years in the classroom for each planned Inpres school per 1,000 children. (The average lag from age 12 to ages two to six is eight years.) I call this impact estimate τ and document it in the upper-left of the plot. Under the original paper's classical variance estimator, the associated standard error is 0.046, which is also displayed. This point estimate is about three times larger than the closest corresponding estimate in the original, 0.0147 (Duflo 2001, Table 4, panel A, column 4). The t statistic here is a bit smaller, at 1.62 instead of 2.02, but still favors the existence of a trend break in schooling attainment around age 12 (pending technical revisions below).[13]

While pointing up the relevance of Mincer's standard labor function, section IV gave reason to expect that broad patterns of higher polynomial order are at play in the relationship between schooling and wages. The same may hold for the relationship between age and schooling. These possibilities lead to the question of whether the linear spline term $D_j t$ suffices to control for ambient cross-sectional trends with respect to age. If not, then in the piecewise-linear model just presented, gradual curvature generated by non-Inpres factors could load onto the Inpres-inspired spline term, generating spurious results. To illustrate the concern, I add a parabolic fit to the cohort-specific coefficient estimates (see the right of Figure 6). This fit is made after replacing the term (4) in $\widetilde{\mathbf{T}}_t$ with t^2. Though the piecewise-linear and quadratic models carry different structural

[12] At the crest of the Inpres surge in the early 1980s, Indonesia's primary school gross enrollment ratio, which is the ratio of the number of enrolled children of *any* age to the number of all children of official age, temporarily surpassed 120% (Suharti 2013, Figure 2.5).

[13] A quinquennial rhythm in SUPAS data may add noise to Figure 6. People with little schooling are more likely to report being born in a year ending in 0 or 5, evidently because they are not sure when they were born. The enumerator's manual discusses this problem (international.ipums.org/international-action/source_documents/enum_instruct_id1995a_tag.xml) It is not obvious how this distortion affects the age-specific linear associations between schooling and Inpres treatment plotted in the figure. But the estimates for ages 24, 19, 14, 9, and 4 are mostly higher than those on either side.

implications—pointing to the impact of a discrete, crisply launched program or to longer-term patterns—the two fits look very similar. The 1995 SUPAS data can hardly distinguish between them. Appendix A formalizes this characterization. When the evidence struggles to adjudicate between competing models, that shifts inferential burden onto the reader's priors. E.g., to the extent that one is confident that the true model for the non-Inpres trend is linear, one should favor the structural break reading.

Figure 7 brings the OLS/reduced-form analysis in the left pane of Figure 6 to more outcomes and all follow-ups. The first row confirms that the upward trend in the association between primary school completion and Inpres treatment accelerates with timing ascribable to the intervention. Though consistent, as in Figure 5 above, the effect weakens in later follow-ups.

It is unclear under the spline-based standard whether Inpres SD affected total schooling attainment even in 1995 (second row of Figure 7). The estimated impact is positive in all four follow-ups, but with little statistical significance. In the revised treatment the t statistic for 1995 is only 0.53. This finding resonates with the weakness found earlier in the first stages of the 2SLS specifications, as well as with the Duflo (2001, p. 804) finding that Inpres treatment *reduced* secondary school progression even as it increased primary school progression. Possibly the burgeoning primary school system diverted resources and even students from the secondary school system, blunting the impact on total schooling.

Also surprising is that while the reduced-form impacts on labor market outcomes are more statistically significant, the signs are not consistent. The wage earnings trend bends upward in the 1995 data and in 2013–14, but downward in 2005 and 2010 (last two rows of Figure 7). Because of its larger sample, the 2013–14 data dominates in the consolidated post-1995 fit.

Appendix B documents the results of challenging this piecewise-linear model with a quadratic time control. Just as in the right half of Figure 6, the data are usually unable to choose clearly between the two models, which can make it hard to be confident in the existence of any trend breaks associated with the arrival of Inpres SD. In the few instances where one model wins, it is the quadratic.

I next perform 2SLS in order to formally estimate returns to schooling within this framework. This step is analogous to taking the ratio of a τ reported for the log hourly wage in the fourth row of Figure 7 to a τ for schooling in the second row. The results are in Table 4. As should be expected in light of the unclear impact on schooling, the KP F statistics are again very low and the weak-

identification-robust confidence sets very wide. Switching the instrumented treatment variable from years of schooling to primary completion hardly changes the story: all confidence sets are unbounded in at least one direction. (See Table A-3.)

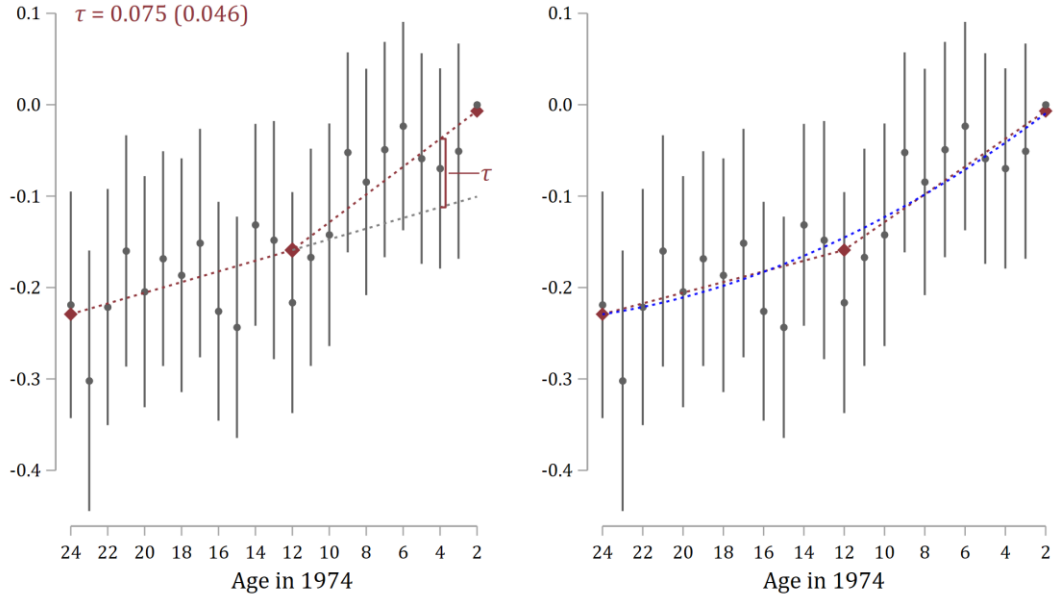

Figure 6. Coefficients on the interactions of age in 1974 and program intensity in region of birth in the schooling equation, with piecewise-linear fit

Notes: Both plots depict, in grey, the coefficient estimates and 95% confidence intervals in Duflo (2001), Figure 1, except that the age-2 rather than age-24 cohort is the omitted reference cohort, which shifts all values down by 0.22. The coefficients are for interactions between age dummies and Inpres treatment intensity (planned new schools per child) in an OLS regression of the log hourly wage that includes regency-of-birth dummies and the intermediate control set (interactions between age dummies and both the number of children and the population enrollment rate in the birth regency in 1971). Confidence intervals are computed from classical variance estimates. The fits of two more restrictive models are superimposed. On the left, the plotted terms are replaced by interactions between two linear spline terms and treatment intensity. The gap labeled τ is an impact estimate implied by this fit. On the right, the linear spline fit is joined by a quadratic fit, which is estimated by regression on interactions between treatment intensity and the 1st and 2nd powers of age.

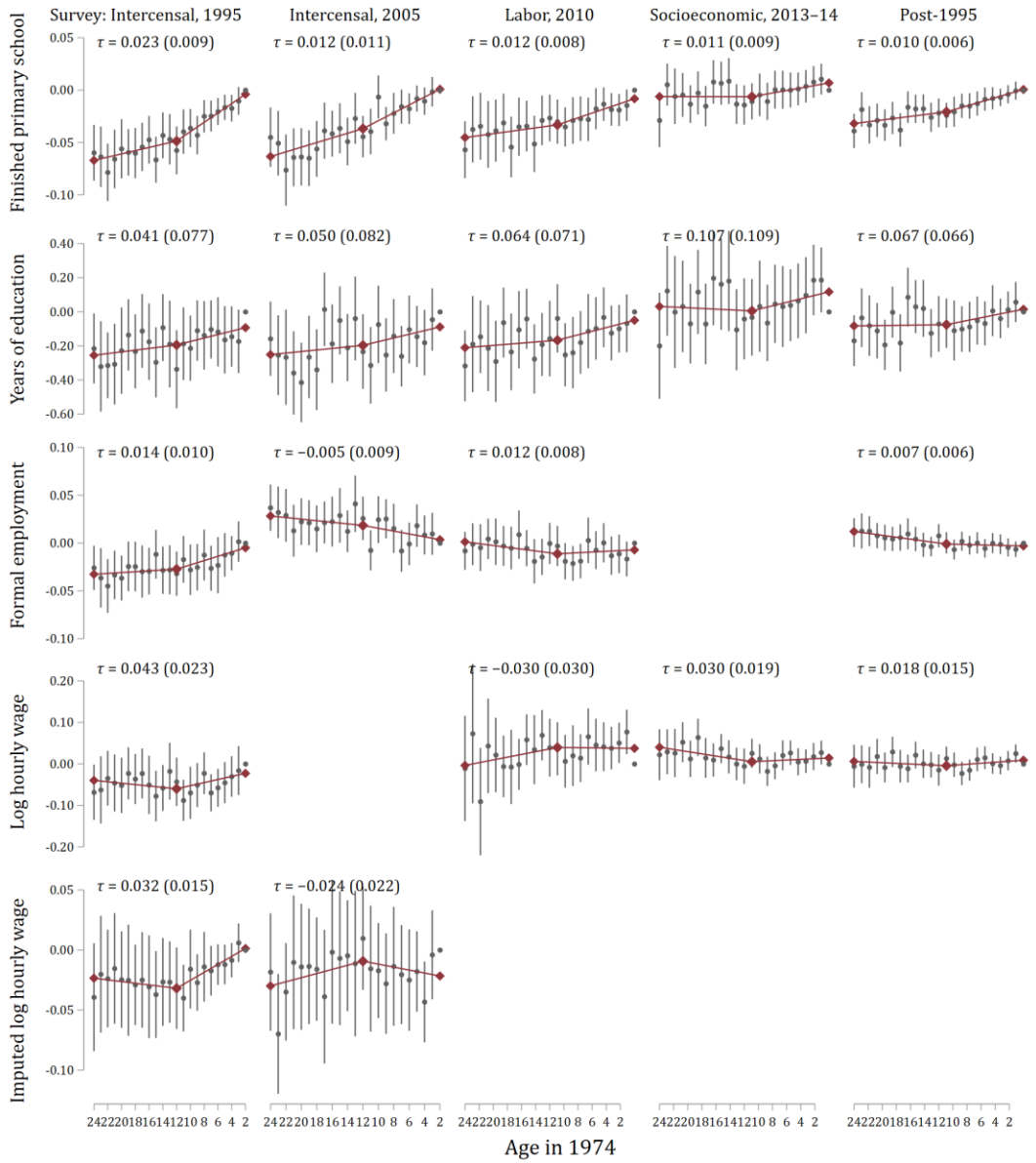

Figure 7. Coefficients on interactions between age dummies and Inpres intensity in region of birth in weighted OLS/reduced-form regressions of outcomes observed in various surveys

Notes: Plots are constructed as in the left half of Figure 6 (which see), except that standard errors are clustered by regency. In 2013–14, "wages" are not wages from formal employment in the last month, but typical monthly net income from main work activity, which may be self-employment. The post-1995 regressions pool the data from 2005, 2010, and 2013–14.

Table 4. Spline-based 2SLS estimates of the effect of years of schooling on labor market outcomes

	Linear time control		Quadratic time controls	
	Unweighted	Weighted	Unweighted	Weighted
A. Full sample, 1995: participation in wage sector				
Coefficient	0.385	0.389	0.156	0.00766
	(0.465)	(0.856)	(0.129)	(0.0831)
95% confidence set	(−∞, −0.24] ∪ [0.07, ∞)	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)
KP *F*	0.73	0.21	1.56	1.55
Observations	152,989	152,989	152,989	152,989
B. Wage earners, 1995: log hourly wage				
Coefficient	0.167	0.173	0.162	0.0895
	(0.109)	(0.0934)	(0.225)	(0.109)
95% confidence set	(−∞, ∞)	[−0.02, ∞)	(−∞, ∞)	(−∞, ∞)
KP *F*	2.11	4.40	0.40	2.50
Observations	60,663	60,663	60,663	60,663
C. Full sample, 2005: participation in wage sector				
Coefficient	–0.180	–0.171	0.239	0.220
	(2.434)	(0.493)	(0.336)	(0.350)
95% confidence set	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)
KP *F*	0.01	0.22	0.40	0.38
Observations	159,205	159,205	159,205	159,205
D. Wage earners, 2005: imputed log hourly wage				
Coefficient	–0.159	25.14	–0.209	–0.295
	(0.496)	(4037.2)	(0.490)	(0.723)
95% confidence set	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)
KP *F*	0.28	0.00	0.34	0.27
Observations	40,263	40,263	40,263	40,263
E. Full sample, 2010: participation in wage sector				
Coefficient	0.151	0.321	0.232	0.135
	(0.185)	(0.867)	(0.154)	(0.127)
95% confidence set	(−∞, ∞)	(−∞, ∞)	(−∞, −0.15] ∪ [0.08, ∞)	(−∞, ∞)
KP *F*	0.56	0.12	1.55	1.09
Observations	153,864	153,864	153,864	153,864
F. Wage earners, 2010: log hourly wage				
Coefficient	0.683	0.283	0.107	0.0509
	(3.576)	(0.365)	(0.141)	(0.0757)
95% confidence set	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)
KP *F*	0.03	0.47	0.93	3.81
Observations	35,992	35,992	35,992	35,992
G. Income earners, 2013–14: log typical monthly earnings				
Coefficient	0.230	1.842	0.106	0.145
	(0.214)	(13.99)	(0.0572)	(0.123)
95% confidence set	(−∞, ∞)	(−∞, ∞)	[0.00, 0.61]	(−∞, ∞)
KP *F*	1.19	0.02	5.33	1.78
Observations	221,041	221,041	221,041	221,041
H. Combined D, F, G				
Coefficient	0.0861	0.0834	0.353	–0.527
	(0.0427)	(0.0712)	(0.635)	(2.197)
	[0.00, 0.25]	(−∞, −0.31] ∪	(−∞, ∞)	(−∞, ∞)
95% confidence set		[−0.15, ∞)		
KP *F*	8.06	3.43	0.26	0.08
Observations	297,297	297,297	297,297	297,297

Notes: Notes to Table 2 apply, except that all regressions gain $D_j t$ as a control, and the quadratic ones $D_j t^2$ as well; and the sole instrument is $D_j \cdot \max(0,12 - t)$ where t is age in 1974.

VII. Quantile-based estimation

Wage scale dilation preserves *order*: while the university graduates pull away from primary school dropouts as they age, their quantiles within the wage distribution change less. Applying a DID-type method that constructs the treatment counterfactual using quantiles could immunize the results against wage scale dilation.

A leading example of quantile-based DID is the changes-in-changes (CIC) estimator of Athey and Imbens (2006). It 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.

While the robustness of CIC to violations of the parallel trends assumption gives it an advantage over the 2SLS estimators employed just above, CIC the disadvantage of not harvesting information about *timing*. As a result, in this context, its source of treatment variation, Inpres SD allocation, is not as obviously exogenous.

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. The young and old cohorts are defined as usual. 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. As in Duflo (2001)'s 2×2 DID, I weight observations. For consistency with previous sections, I incorporate Duflo (2001)'s minimal control set, using the method of Melly and Santangelo (2015). Estimates of impacts on earnings in all four follow-ups appear in Figure 8 along with bootstrap-based 95% confidence intervals.

The CIC results indicate that Inpres SD had little clear, systematic effect on earnings. In contradiction with the linear spline–based estimation, the 2010 follow-up now produces the strongest signs of positive impact, in the high quantiles. The p values for from a Cramér–von Mises test for no impact at any percentile in the 1995, 2005, 2010, 2013–14, and combined post-1995 follow-

ups are 0.62, 0.09, 0.43, 0.08, and 0.69.[14] The hypothesis of *positive* impacts at all percentiles is generally rejected more confidently, with p values of 0.28, 0.07, 0.64, 0.01, and 0.22. The hypothesis of all negative impacts is hardly rejected, with p =0.72, 0.54, 0.16, 0.69, 0.92. In sum, the results suggest that Inpres SD reduced (wage) earnings in 2005 and 2013–14 and perhaps increased wages in 2010. The pattern of signs here is not the same as in the linear spline–based results (last two rows of Figure 7), but the inconsistency across follow-ups is familiar.

de Chaisemartin and D'Haultfœuille (CH, 2017) introduces a fuzzy variant of CIC that offers a quantile-based way to estimate returns to schooling. Indeed, that paper illustrates its methods on the Duflo (2001) data. It estimates the log wage return to a year of schooling at 0.099 (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 instrumentation 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 instruments schooling using a trichotomous factor variable that is itself a function of schooling and thus is about as presumptively endogenous. As is demonstrated in appendix B, 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.

[14] Using the same Melly and Santangelo (2015) "cic" package, I also estimate impacts on years of schooling. The results are degenerate, perhaps because schooling is a discrete variable, whose coarse quantization hampers quantile-based methods.

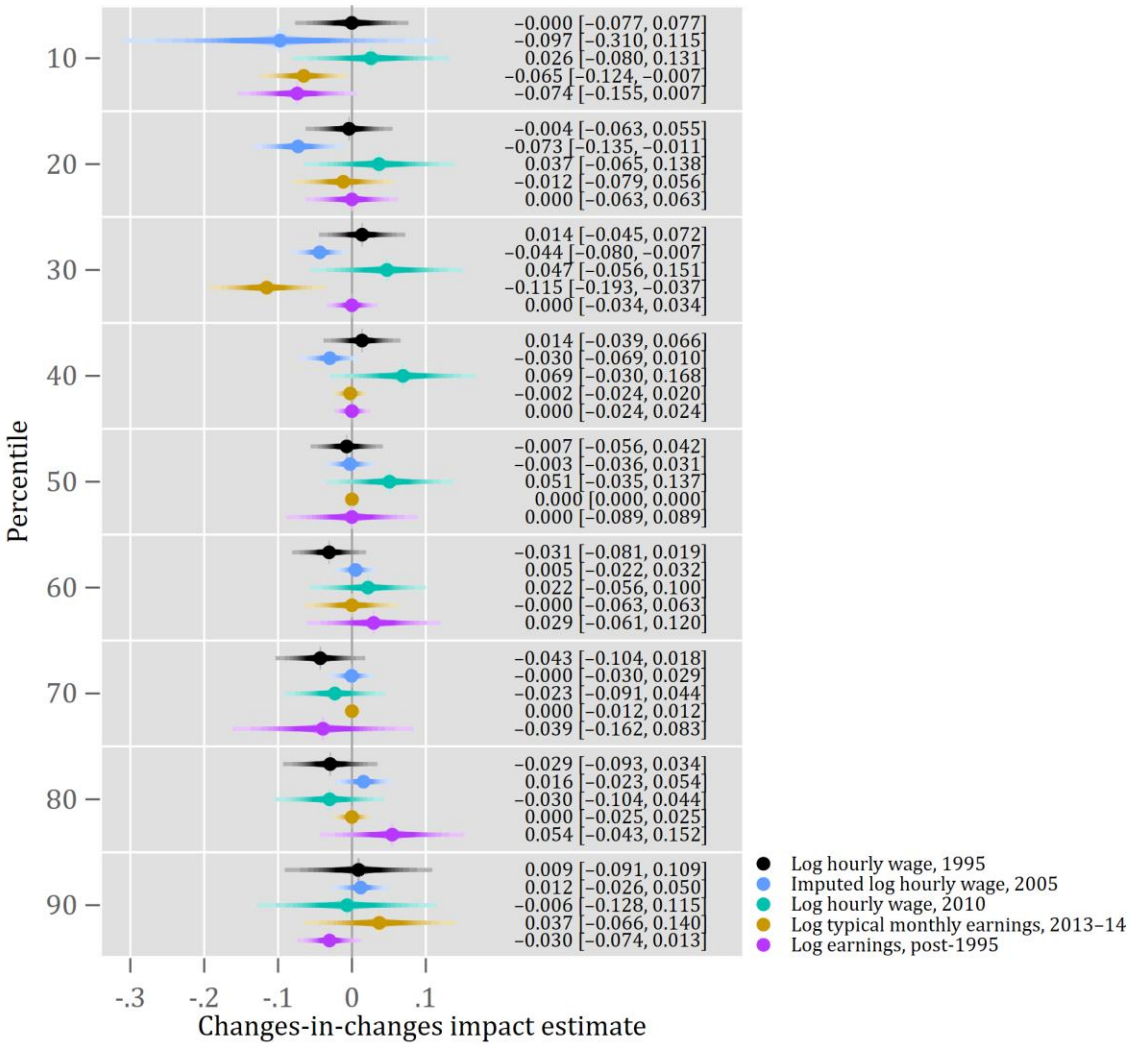

Figure 8. Changes-in-changes estimates of impact of high Inpres treatment at various percentiles of log wage or earnings

Notes: The figure shows, at various percentiles, estimates of the impact of high Inpres treatment on earnings in four follow-ups and bootstrapped 95% confidence intervals. Point estimates and confidence bounds are labeled 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 and use of survey weights all mimic the Duflo (2001) 2×2 DID regressions. The control set is the Duflo (2001) minimal control set, based on interactions between age dummies and number of children aged 5–14 in 1971 in the birth regency. The post-1995 regression consolidates the three later samples and adds year dummies as controls.

VIII. Conclusion

The question animating this reanalysis is not whether schooling and wages are positively related at the individual level, nor even whether the first substantially affects the second. It is, rather, how well the quasi-experimental literature has succeeded in removing potential biases in OLS-type evidence on that association. This reanalysis raises substantial questions about whether Duflo (2001) makes progress in this respect. The study's design allows bias to enter the estimates through nonparallel trends. In addition, the impact of the natural experiment on schooling is small enough

that weak identification further biases 2SLS toward OLS. A model with quadratic trends fits the data at least as well as the piecewise-linear model, which makes it hard to judge whether Inpres SD caused step changes in time series—which, if present, would constitute the most compelling signs of impact. CIC, designed for contexts where the parallel trends assumption is violated, generates estimates that are mostly indistinguishable from zero.

How is the evidence best read? It is entirely plausible that a large primary school construction campaign boosted primary school completion, as the piecewise-linear estimator finds in all follow-ups. It is less clear that the program affected total schooling or adult earnings. Such findings do not shine through clearly in the various approaches taken here.

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

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.

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.

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.

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.

Bappenas. 1973–78. *Pedoman Pelaksanaan Bantuan Pembagunan Sekolah Dasar Menurut Instruki Presiden Republik Indonesia.* Jakarta.

Baum, Christopher F., Mark E. Schaffer, and Steven Stillman. 2007. "Enhanced Routines for Instrumental Variables/Generalized Method of Moments Estimation and Testing." *Stata Journal* 7 (4): 465–506.

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.

Biro Pusat Statistik (BPS). 1972. *Sensus Penduduk 1971: Diperintji Menurut Propinsi dan Kabupaten/Kotamadya*.

BPS. 1974. *Sensus Penduduk 1971: Series E*.

Bischof, Daniel. 2017. "New Graphic Schemes for Stata: plotplain and plottig." *Stata Journal* 17 (3): 748–59.

Correia, Sergio. 2014. "REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects."

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.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2017. "Fuzzy Differences-in-Differences." *The Review of Economic Studies* 85 (2): 999–1028.

de Chaisemartin, Clément de, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96.

Duflo, Esther. 1999. *Essays in Empirical Development Economics*. Dissertation. Massachusetts Institute of Technology. https://dspace.mit.edu/handle/1721.1/9516.

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.

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.

Duflo, Esther. 2004b. "Scaling up and Evaluation." In *Annual World Bank Conference on Development Economics*, 341–69. http://hdl.handle.net/10986/14889.

Finlay, Keith, and Leandro M. Magnusson. 2019. "Two Applications of Wild Bootstrap Methods to Improve Inference in cluster-IV Models." *Journal of Applied Econometrics*.

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225 (2): 254–77.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50 (4): 1029–54.

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.

Hsiao, Allan. 2022. "Educational Investment in Spatial Equilibrium: Evidence from Indonesia." May 15. https://allanhsiao.github.io/files/Hsiao_schools.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.

Jann, Ben. 2014. "Plotting Regression Coefficients and Other Estimates." *Stata Journal* 14 (4): 708–37.

Jann, Ben. 2022. "Color Palettes for Stata Graphics: An Update." *University of Bern Social Sciences Working Papers*, April.

Kleibergen, Frank, and Richard Paap. 2006. "Generalized Reduced Rank Tests Using the Singular Value Decomposition." *Journal of Econometrics* 133 (1): 97–126.

Lemieux, Thomas. 2006. "The 'Mincer Equation' Thirty Years After Schooling, Experience, and

Earnings." In *Jacob Mincer: A Pioneer of Modern Labor Economics*, edited by Shoshana Grossbard, 127–45. Springer US.

Liang, Kung-Yee, and Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73 (1): 13–22.

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

Mincer, Jacob. 1958. "Investment in Human Capital and Personal Income Distribution." *The Journal of Political Economy* 66 (4): 281–302.

Mincer, Jacob A. 1974. *Schooling, Experience, and Earnings*. National Bureau of Economic Research.

Minnesota Population Center (MPC). Integrated Public Use Microdata Series, International: Version 7.3. Minneapolis, MN: IPUMS, 2020.

Moreira, Marcelo J. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica: Journal of the Econometric Society* 71 (4): 1027–48.

Roodman, David. 2011. "Fitting Fully Observed Recursive Mixed-Process Models with cmp." *Stata Journal* 11 (2): 159–206.

Roodman, David, Morten Ørregaard Nielsen, James G. MacKinnon, and Matthew D. Webb. 2019. "Fast and Wild: Bootstrap Inference in Stata Using boottest." *Stata Journal* 19 (1): 4–60.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50 (2): 301–16.

Stock, James H., and Motohiro Yogo. 2002. "Testing for Weak Instruments in Linear IV Regression." Technical Working Paper Series. National Bureau of Economic Research.

Suharti. 2013. "Trends in education in Indonesia." In Daniel Suryadarma & Gavin W. Jones, eds., *Education in Indonesia*. ISEAS Publishing.

Surbakti, Pajung. 1995. *Indonesia's National Socio-Economic Survey: A Continual Data Source for Analysis on Welfare Development*. Central Bureau of Statistics.

Appendices

A. Substituting primary school completion for years of schooling as the treatment

This appendix reports results from regressions cited in the main text, which take primary school completion rather than years of schooling as the instrumented treatment variable.

Table A-1 Estimates of the effect of primary school completion on labor market outcomes, 1995

	OLS		2SLS: instrument by young/old		2SLS: instruments by birth year	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
	(1)	(2)	(3)	(4)	(5)	(6)
A. Full sample: participation in wage sector						
Coefficient	0.171	0.148	0.680	0.443	0.635	0.450
	(0.00555)	(0.00697)	(0.171)	(0.153)	(0.133)	(0.118)
95% confidence set			[0.39, ∞)	[0.13, 0.80]	[0.26, ∞)	[−0.17, ∞)
Overidentification *p*					0.75	0.84
KP *F*			21.24	28.83	2.72	4.29
Observations	152,989	152,989	78,470	78,470	152,989	152,989
B. Wage earners: log hourly wage						
Coefficient	0.456	0.439	0.758	0.483	0.596	0.439
	(0.0121)	(0.0134)	(0.457)	(0.453)	(0.391)	(0.354)
95% confidence set			[−0.13, ∞)	[−0.50, ∞)	(−∞, ∞)	[−0.07, ∞)
Overidentification *p*					0.68	0.05
KP *F*			10.49	19.31	2.58	2.97
Observations	60,663	60,663	31,061	31,061	60,663	60,663

Notes: Notes to Table 2 apply except that the instrumented variable is now primary school completion.

Table A-2 Estimates of the effect of primary school completion on labor market outcomes, post-1995

	OLS		2SLS: instrument by young/old		2SLS: instruments by birth year	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
	(1)	(2)	(3)	(4)	(5)	(6)
A. Full sample, 2005: participation in wage sector						
Coefficient	0.174	0.173	–0.317	–0.507	–0.252	–0.329
	(0.00426)	(0.00485)	(0.266)	(0.198)	(0.157)	(0.113)
95% confidence set			(−∞, 0.22]	(−∞, −0.14]	(−∞, −0.01]	(−∞, ∞)
Overidentification *p*					0.05	0.02
KP *F*			7.17	17.83	2.52	4.75
Observations	159,205	159,205	83,927	83,927	159,205	159,205
B. Wage earners, 2005: imputed log hourly wage						
Coefficient	0.347	0.338	0.257	–0.407	0.338	–0.138
	(0.0101)	(0.00973)	(0.635)	(0.788)	(0.582)	(0.585)
95% confidence set			(−∞, ∞)	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)
Overidentification *p*					0.89	0.56
KP *F*			9.19	8.76	1.80	2.01
Observations	40,264	40,264	22,074	22,074	40,263	40,263
C. Full sample, 2010: participation in wage sector						
Coefficient	0.167	0.162	0.532	–0.0797	0.735	0.00576
	(0.00434)	(0.00537)	(0.342)	(0.256)	(0.345)	(0.191)
95% confidence set			[−0.12, ∞)	[−0.72, 0.60]	[−0.02, ∞)	[−0.29, ∞)
Overidentification *p*					0.57	0.15
KP *F*			7.20	10.34	1.26	2.00
Observations	153,864	153,864	83,190	83,190	153,864	153,864
D. Wage earners, 2010: log hourly wage						
Coefficient	0.720	0.720	3.197	–0.310	1.457	0.145
	(0.0226)	(0.0293)	(2.234)	(2.419)	(0.694)	(0.709)
95% confidence set			[−0.00, ∞)	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)
Overidentification *p*					0.27	0.28
KP *F*			2.18	1.48	1.24	1.14
Observations	35,992	35,992	21,435	21,435	35,992	35,992
E. Income earners, 2013–14: log typical hourly earnings						
Coefficient	0.571	0.594	–0.0857	–0.378	0.114	0.481
	(0.0104)	(0.0100)	(0.591)	(2.107)	(0.342)	(0.648)
95% confidence set			(−∞, ∞)	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)
Overidentification *p*					0.25	0.29
KP *F*			6.36	0.83	2.00	0.89
Observations	221,041	221,041	125,047	125,047	221,041	221,041
F. Combined B, D, and E						
Coefficient	0.594	0.615	0.118	0.414	0.232	0.621
	(0.0104)	(0.00940)	(0.612)	(0.921)	(0.404)	(0.522)
95% confidence set			(−∞, ∞)	(−∞, ∞)	(−∞, −0.51] ∪ [−0.10, ∞)	[0.26, ∞)
Overidentification *p*					0.09	0.21
KP *F*			7.12	3.94	1.83	1.54
Observations	297,297	297,297	168,557	168,557	297,297	297,297

Notes: Notes to Table 2 and Table 3 apply. Each panel shows results from a different combination of dependent variable and survey.

Table A-3 Spline-based 2SLS estimates of the effect of primary school completion on labor market outcomes

	Linear time control		Quadratic time controls	
	Unweighted	Weighted	Unweighted	Weighted
A. Full sample, 1995: participation in wage sector				
Coefficient	1.117	0.599	1.652	0.136
(standard error)	(0.582)	(0.482)	(1.400)	(1.440)
95% confidence set	[0.29, ∞)	[−0.34, ∞)	(−∞, ∞)	(−∞, ∞)
KP *F*	6.84	6.43	1.11	0.37
Observations	152,989	152,989	152,989	152,989
B. Wage earners, 1995: log hourly wage				
Coefficient	1.255	1.475	0.978	1.461
(standard error)	(0.867)	(0.909)	(1.181)	(1.958)
95% confidence set	(−∞, ∞)	[−0.12, ∞)	(−∞, ∞)	(−∞, ∞)
KP *F*	3.42	5.32	2.24	1.20
Observations	60,663	60,663	60,663	60,663
C. Full sample, 2005: participation in wage sector				
Coefficient	0.918	–0.568	0.957	1.920
	(9.467)	(0.970)	(0.931)	(3.408)
95% confidence set	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)
KP *F*	0.02	1.15	2.18	0.39
Observations	159,205	159,205	159,205	159,205
D. Wage earners, 2005: imputed log hourly wage				
Coefficient	–0.467	–1.071	–4.146	–2.104
	(0.934)	(1.197)	(12.14)	(2.620)
95% confidence set	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)
KP *F*	6.18	4.45	0.13	1.11
Observations	40,263	40,263	40,263	40,263
E. Full sample, 2010: participation in wage sector				
Coefficient	1.489	0.863	5.168	5.761
	(1.855)	(0.971)	(8.620)	(23.83)
95% confidence set	(−∞, ∞)	(−∞, ∞)	[0.80, ∞)	(−∞, ∞)
KP *F*	0.66	1.55	0.33	0.06
Observations	153,864	153,864	153,864	153,864
F. Wage earners, 2010: log hourly wage				
Coefficient	–4.632	–3.450	24.07	1.479
	(16.19)	(6.276)	(362.9)	(2.775)
95% confidence set	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)
KP *F*	0.15	0.50	0.00	0.89
Observations	35,992	35,992	35,992	35,992
G. Income earners, 2013–14: log typical monthly earnings				
Coefficient	2.255	4.532	1.252	1.590
	(1.883)	(6.601)	(0.692)	(1.291)
95% confidence set	(−∞, ∞)	(−∞, ∞)	[0.04, ∞)	[−0.60, ∞)
KP *F*	1.88	0.41	6.13	2.42
Observations	221,041	221,041	221,041	221,041
H. Combined D, F, G				
Coefficient	1.221	1.042	1.783	2.059
	(0.680)	(0.976)	(1.760)	(2.656)
95% confidence set	[0.02, ∞)	(−∞, ∞)	(−∞, ∞)	(−∞, ∞)
KP *F*	6.71	3.46	1.59	0.74
Observations	297,297	297,297	297,297	297,297

Notes: Notes to Table 4 apply.

B. Quadratic spline fits to cohort-specific associations between treatment instrument and outcomes

Figure B-1, below, is a version of Figure 7 in the main text, designed to formally compare piece-wise-linear and polynomial models for the evolution across cohorts of the association between Inpres treatment intensity (D_j) and various outcomes. Relative to Figure 7, the plots gain quadratic

fits in blue. To formally test whether one model fits better in each case, omnibus regression are run that include the distinctive terms of both. Each plot is annotated with cluster-robust p values from Wald tests that the coefficient on each model's distinctive term is zero. For example, the upper-left pane of the figure shows that in a regression including $D_j t$, $D_j t^2$, and $D_j \cdot \max(0,12-t)$, the p value for $D_j t^2$, is 0.09 (shown in blue) while that for the distinctive term in the piecewise-linear model, $D_j \cdot \max(0,12-t)$, is 0.83 (in red).

In most cases, the tests do not signal strong preference for one model over the other, in the form of a high p value for one and a low p value for the other. When they do, as in the example just given, the smooth quadratic model is favored over the trend break model.

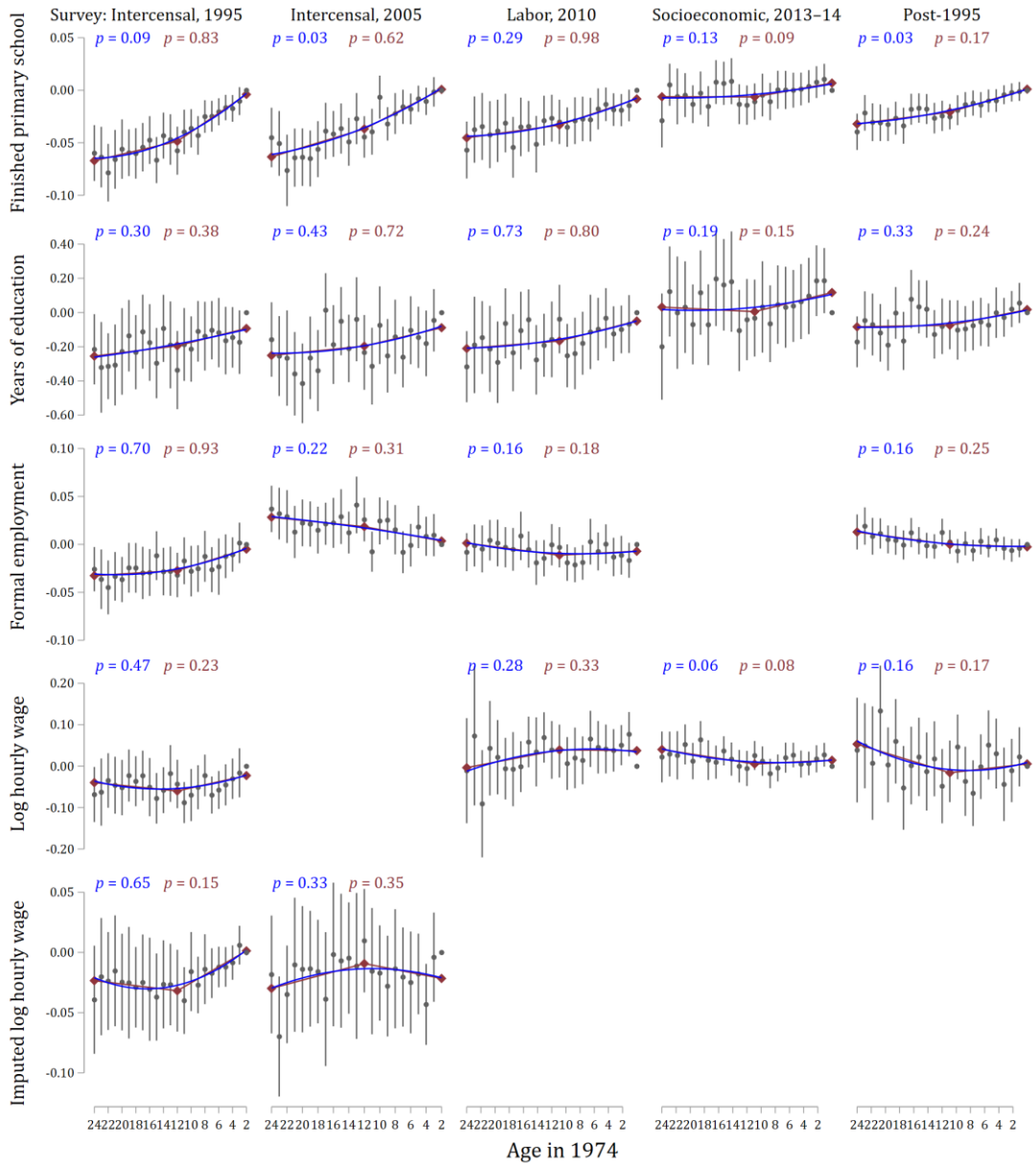

Figure B-1. Coefficients on the interactions between age dummies and program intensity in region of birth in weighted OLS regressions of outcomes observed in various surveys, with piecewise-linear and quadratic fits

Notes: Plots are constructed as in the right half of Figure 6 (which see), except that standard errors are clustered by birth regency. Quadratic fits and associated *p* values are in blue. Those for piecewise linear fits are in red.

C. Simulation of de Chaisemartin and D'Haultfœuille (2017)'s Wald CIC

Section VII asserts that a combination of procedures in de Chaisemartin and D'Haultfœuille (CH, 2017)— gathering groups into three supergroups, then performing Wald CIC—will produced 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

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

$Z_{jt} \sim$ 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 $s_{j1} \sim$ 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.

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.[15] In the second, $\rho = 0.95$, making treatment highly endogenous. In both, the true impact parameter β is 0.

Results from the simulations appear in Table C-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 by a standard deviation away from the true value of zero when estimating from the CH supergroups.

[15] A 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 C-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.

D. 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. 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 preferred variant of a Duflo (2001) regression. In the upper right corner of Table 1 appear results from a reduced-form regression of the log hourly wage in 1995 on the treatment intensity indicator $D_j T_t$. The sample is restricted to men aged 2–6 and 12–17 in 1974. The minimal Duflo (2001) control set is included. Standard errors are clustered by birth regency; observations are weighted; data corrections are incorporated.

I modify the Jakiela test in two minor ways. Since treatment is not binary, I split the sample according to whether the treatment indicator $D_j T_t \neq 0$, i.e, age in 1974 is 2–6, or $D_j T_t = 0$, i.e.,

age is 12–17. And since the representative regression includes other controls, I partial them out too. Figure D-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. The linear fits for treated and untreated, whose slopes constitute estimates of mean treatment effect, are 0.170 and 0.164. The *p* value for the null hypothesis that they are the same is 0.88.

This check for treatment heterogeneity therefore finds none.

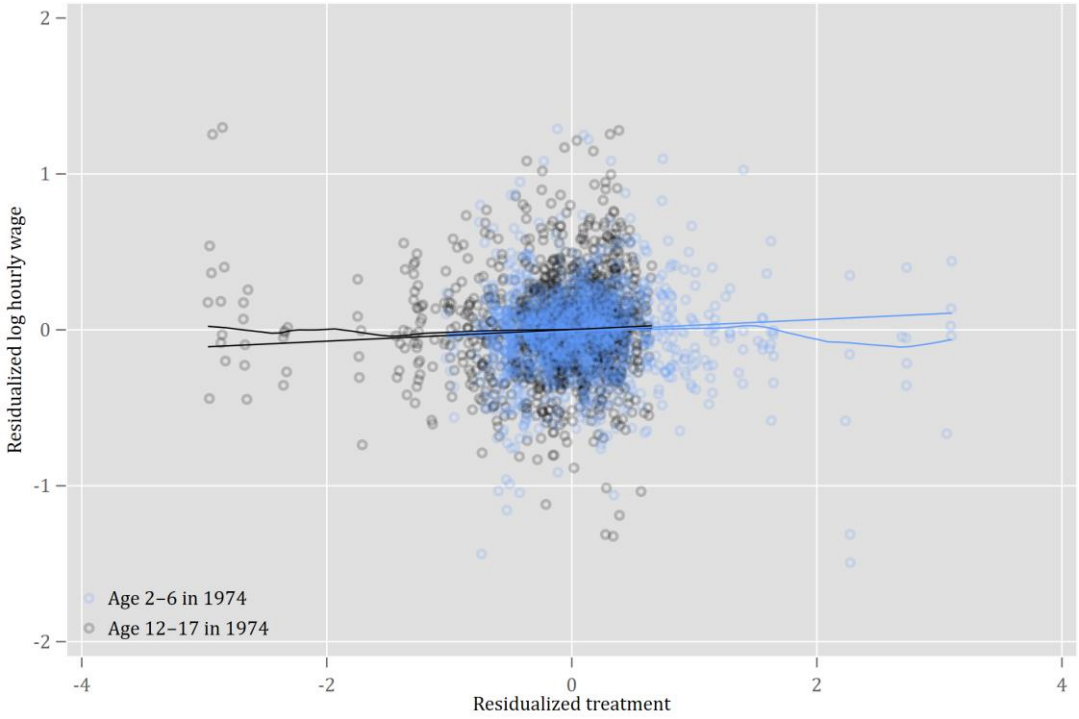

Figure D-1. Residualized outcome vs. residualized treatment in revised Duflo (2001) reduce-form wage regression, age and birth regency