

AP: Duflo (2001)

Alessandro Caggia

June 2025

Abstract

1. Application: Duflo (2001)

2. 1.1. Research Questions

- 3. Effect of school infrastructure on future wages (through higher
4. education)

5. 1.2. The Natural Experiment

- 6. Massive primary school construction program, started in 1973
7. in Indonesia: between 1973–1974 and 1978–1979 more than
8. 60,000 schools were built.

9. 1.3. Data

- 10. Uses 1995 Census, more than 150,000 individuals. This is A SIN-
11. GLE CROSS SECTION.
- 12. Only 60,000 work for a wage (others are self-employed).

13. 1.4. Treatment Assignment

- 14. Number of schools to be built proportional to number of unen-
15. rolled primary school-age children in 1972.

16. Sources of Exogenous Variation

17. Brilliant: COHORT PLAYS THE ROLE OF TIME! Two key sources:

18. 1. District variation: High intensity districts = more schools 19. given population (residuals from school allocation regression).

- 20. Duflo estimates the following school allocation regression:
$$\log(\text{Schools}_d) = \alpha + \beta_1 \ln(\text{Children}_d) + \beta_2 \ln(1 - \text{Enr.Rate}_d) + \varepsilon_d$$

21. Then, for each district d , the residual ε_d captures how
22. many more or fewer schools the district received com-
23. pared to what the regression predicts:

- 24. – If $\varepsilon_d > 0$: the district received **more schools than
expected**.
- 25. – If $\varepsilon_d < 0$: the district received **fewer schools than
expected**.

26. 2. Cohort variation: children aged >12 in 1974 unaffected (al- 27. ready out of primary school); starting from 1974 The younger a 28. child was in 1974, the more it benefited from the program be- 29. cause spent more time in the new schools.

31. Simple Difference 1: High vs Low Districts

- 32. Why not just estimate a simple difference: comparing schooling
33. in high vs in low intensity districts?
- 34. Biased: high-intensity districts had lower education achievement pre-program. So T and C groups are not comparable (se-
35. lection bias). But they idea is: they are different there is selection
36. bias but if I control for such selection we are fine. in this case
37. you would need to control for a lot of unobservables and en-
38. rollment achievement pre-program (which we don't have a good
39. measure of).

Simple Difference 2: Young vs Older Children

- 41. Young = 1 if age 2–6 in 1974, 0 if age 12–17.
42. Why not just estimate simple difference: comparing school-
43. ing for young vs older children? (Even without the program,
44. younger cohorts typically get more schooling than older ones
45. (due to economic growth, reforms, etc))
46. Biased due to secular upward trend in education over time. If
47. the problem had been: t1 cohort are more males, you control
48. and you are fine. the problem is: t1 cohort has more education
49. which is also your outcome.
50. you could use time effects to control for this but you need a lot
51. of time periods
52. Unbiased only if educational attainment would have evolved
53. similarly for both groups absent the program (education young
54. cohort would have had without the program similar to education old cohort would have had without the program).
55. 56.

1.5. Final methodology

$$\text{DID} = [\mathbb{E}(S | \text{Young} = 1, \text{High}) - \mathbb{E}(S | \text{Young} = 0, \text{High})] \\ - [\mathbb{E}(S | \text{Young} = 1, \text{Low}) - \mathbb{E}(S | \text{Young} = 0, \text{Low})]$$

- 57. the second term in the difference is time! this is a normal DiD
58. with the first term you remove the bias coming from being in
59. better district
60. with the second you remove the secular trend
61.

1.6. Results

62. exactly showing what we have discussed above theoretically
63.

Table 3—MEANS OF EDUCATION AND LOG(WAGE) BY COHORT AND LEVEL OF PROGRAM CELLS

	Years of education			Log(wages)		
	Level of program in region of birth			Level of program in region of birth		
	High (1)	Low (2)	Difference (3)	High (4)	Low (5)	Difference (6)
<i>Panel A: Experiment of Interest</i>						
Aged 2 to 6 in 1974	8.49 (0.043)	9.76 (0.037)	-1.27 (0.057)	6.61 (0.0078)	6.73 (0.0064)	-0.12 (0.010)
Aged 12 to 17 in 1974	8.02 (0.053)	9.40 (0.042)	-1.39 (0.067)	6.87 (0.0085)	7.02 (0.0069)	-0.15 (0.011)
Difference	0.47 (0.070)	0.36 (0.038)	0.12 (0.089)	-0.26 (0.011)	-0.29 (0.0096)	0.026 (0.015)
<i>Panel B: Control Experiment</i>						
Aged 12 to 17 in 1974	8.02 (0.053)	9.40 (0.042)	-1.39 (0.067)	6.87 (0.0085)	7.02 (0.0069)	-0.15 (0.011)
Aged 18 to 24 in 1974	7.70 (0.059)	9.12 (0.044)	-1.42 (0.072)	6.92 (0.0097)	7.08 (0.0076)	-0.16 (0.012)
Difference	0.32 (0.080)	0.28 (0.061)	0.034 (0.098)	0.056 (0.013)	0.063 (0.010)	0.0070 (0.016)

- 64. • **Col:** difference due to the fact that high program areas are
65. worse. So they gain +0.47 and we partialled out the intrinsically
66. lower level (8.02 vs 9.40)
- 67. • **Row:** difference due to overall increases in education within
68. regions over time. Out of 0.47 we remove 0.36 of common trend!
- 69. • **DID:** effect of the program, differences across districts between
70. 26 year-old children and 1217 year-old children.
- 71. • 0.12 additional years of education on average, 0.026 difference
72. in log wages.

- 73 • **Placebo test:** DID for two groups whom we know did not benefit — the estimated effect is indeed statistically 0 (aged 1217 in
 74 1974 vs aged 18-24).

76 **1.7. Wald-DID Estimator: now we need to use the IV**

$$\frac{\text{DID of } Y}{\text{DID of } D} = \frac{0.026}{0.12} = 0.217$$

- 77 • School construction as IV to estimate the returns to education.
 78 • Equivalent to a 2SLS of Y (wages) on **district dummies, cohort dummies, education attainment**, using the interaction
 79 of time and group dummies as instrument.
 80 • de Chaisemartin and D'Haultfuille (2016) show that in a heterogeneous treatment effect setting the **Wald-DID estimator this**
 81 **gives a weighted average of Wald-DIDs.** The weights $w_{g,t}$
 82 depend on the strength of the first stage (i.e., how much the instrument shifts treatment) and the size of each group-time cell.
 83 Therefore, the 2SLS estimate gives more weight to comparisons
 84 where the instrument is more powerful, and thus identifies a
 85 weighted average of local treatment effects.

Regression Equivalent for Each DID with More Than Two Groups
 Practically, Duflo estimated the did through the following regression

$$\text{DID: } S_{ijk} = c + T_i \gamma + G_j \delta + (G_j \times T_i) \rho + \varepsilon_{ijk}$$

- 89 • S_{ijk} : schooling (first stage) or wages (reduced form) for individual i in region j , born in year k .
 90 • G_j : dummy for being in high intensity district.
 91 • T_i : dummy for being in young cohort.
 92 • a_j : district of birth fixed effect, captures differences across districts (instead of G_j).
 93 • β_k : cohort of birth fixed effect, captures differences across cohorts (instead of T_i).
 94 • C_j : vector of region-specific variables for district j .
 95 • P_j : intensity of the program in district of birth j .

note how the results are lower in table 4

$$\frac{0.0270}{0.188} \approx 0.144$$

Source of Difference	Table 3 (Basic DID)	Table 4 (Regression/IV)
0) Functional form	Non Parametric	Linear and additive, but can incorporate richer controls and continuous treatment
1) Controls for cohort (birth year)	Cohorts grouped (Young = 2-6, Old = 12-17) coarse, no specific birth-year FE	Includes full year-of-birth fixed effects (β_k) fine-grained control
2) Program intensity measure	Binary high/low intensity dummy (discrete treatment variable)	Continuous: number of schools per 1,000 children captures variation in exposure intensity
3) Treatment variation	Group-level variation (young/old GE high/low region); no within-region variation	Individual-level variation in treatment intensity (based on region \times year of birth), leading to finer comparisons
4) Controls for region and cohort differences	Only 4 group means; controls are coarse: region = high/low, cohort = young/old	Includes full region fixed effects (a_j) and cohort fixed effects (β_k), absorbing all systematic region and cohort variation. Allows also for time varying controls!

99 **1.8. Identification threats**

- 100 • **Mean reversion bias.**

101 School construction was targeted to regions with low past education levels. **Risk:** These regions may improve over time even
 102 without the program, biasing the estimated effect upward.

103 **Mitigation:** Control for pre-treatment characteristics (e.g., 1971
 104 enrollment, sanitation); EVEN BETTER (what we are doing):

105 compare cohorts within region (young vs old) to remove time-
 106 invariant region-level bias (if there was a rebound this has char-
 107 acterized both young and old!). This is ok if reversion is in
 108 anything but education.

109 • **Confounding program shocks.**

110 The school construction coincided with national shocks (e.g.,
 111 oil windfalls, suppression of school fees). National shocks are
 112 absorbed by cohort/time fixed effects!

113 • **GE effects.**

114 SUTVA? (more educated all around, bad outcome for control bc
 115 mroe competition in the amrket) scalability?

116 *Risk:* likely depends on localized inpmlemtnation and labor mo-
 117 bility

118 **■ Extensions: Cohort-Specific Estimation (Event Study)**

119 • **Estimate effects for each cohort separately.**

120 No need to define arbitrary young vs. old groups. Instead, inter-
 121 act program intensity with cohort dummies.

$$S_{ijk} = c + a_j + \beta_k + \sum_{l=2}^{23} (P_j \cdot d_{il}) \gamma_l + \sum_{l=2}^{23} (C_j \cdot d_{il}) \delta_l + \varepsilon_{ijk}$$

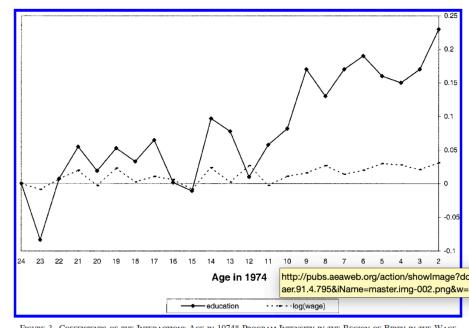


FIGURE 3. COEFFICIENTS OF THE INTERACTIONS AGE IN 1974* PROGRAM INTENSITY IN THE REGION OF BIRTH IN THE WAGE AND EDUCATION EQUATIONS

Figure 1. Coefficient on interactions

123 **■ Duflo (2001): Returns to Education — Identification Threats**

124 **Exclusion Restriction Assumption**

125 • **Key requirement:** The instrument (school construction inten-
 126 sity \times cohort exposure) must affect wages *only through* educa-
 127 tion (enrollment: **the goal is to estimate the causal effect of years**
 128 **of education on wages, using school construction as an instru-**
 129 **ment.**)

130 • **Violation concern:** If school construction affects wages
 131 through other channels (e.g., school quality, local economic de-
 132 velopment), the exclusion restriction fails. **WHATEVER THAT**
 133 **IS NOT SCHOOLING YEARS**

134 **1. School Construction May Affect School Quality**

- 135 • **Mechanism:** The program may increase:

- Teacher hiring,
- Class size reduction,
- Classroom quality or resources,
- Peer quality.

136 • **Risk:** These effects may increase wages *independently* of educa-
 137 tion levels.

138 • **Implication:** Violates the exclusion restriction; bias in IV esti-
 139 mate.

145 **Empirical Check: DID on School Quality Measures**

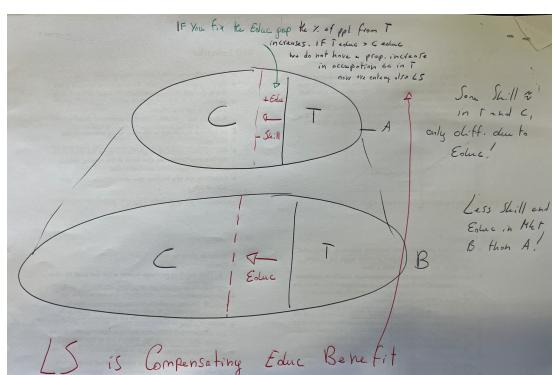
- 146 • Duflo examines whether the program changed **teacher/pupil**
 147 **ratios** using a difference-in-differences (DID) design.
 148 • **Result (Table 6):** No significant effect of the program on
 149 teacher/pupil ratios.
 150 • **Interpretation:** Limited evidence that school quality im-
 151 proved alongside school quantity.

152 **1.8.1. Placebo Tests on Subpopulations Without Education Margin**

- 153 • **Strategy:** Identify subgroups for whom school construction
 154 should *not* affect education quantity (exckyde tnat there was a
 155 braod effect of ecoomic devekopmen etc):
 156 – Individuals with more than 9 years of schooling (already
 157 beyond primary),
 158 – Individuals in dense districts (already high school access).
 159 • **Finding:** No wage effects in these subsamples.
 160 • **Implication:** Suggests the wage effect operates *only* via in-
 161 creased education.

162 **Selection Bias: Full vs Observed Job Market**

163 Heckman selection model: our sample is made up of self selected in-
 164 dividuals. In our case: individuals on average If we observed wages
 165 for the full population (including self-employed and non-wage work-
 166 ers), the DID would correctly capture the average wage effect of edu-
 167 cation for all individuals. In our case, however, wages are observed
 168 only for wage earners. This marekt is made up by the most educated
 169 people. Before treatment it was made up by the most educated peo-
 170 ple in the T group and in the C group (more skilled? more motivated).
 171 Now, compliers among the treated gain extra educaiton and enter
 172 the market. This expands the observed sample to include marginal
 173 individuals in the treated group, while the control group remains
 174 positively selected. Some of the marginal individuals are entering
 175 bc ceteris paribus in skills they lacked educ, because now they com-
 176 pensate low skill with more educ! **THE CONTROL GROUP IS**
177 NOT TH VALID CONTERFACTUAL OF THE TRETAMETN
178 GROUP!!!!



179 Need (reasonable) assumptions on the distribution of human cap-
 180 ital. The rest is just pure intuition. **Expect downward Bias!!!!**