INSTITUTE for REPLICATION

Large-Scale Education Reform in General Equilibrium: Regression Discontinuity Evidence from India - Comment

David Roodman

September 2023

I4R DISCUSSION PAPER SERIES

I4R DP No. 70

Large-Scale Education Reform in General Equilibrium: Regression Discontinuity Evidence from India – Comment

David Roodman[1]

[1]*Open Philantrophy, San Francisco, CA/USA*

SEPTEMBER 2023

Editors

Abel Brodeur
University of Ottawa

Anna Dreber
Stockholm School of Economics

Jörg Ankel-Peters
RWI – Leibniz Institute for Economic Research

Large-Scale Education Reform in General Equilibrium: Regression Discontinuity Evidence from India

Comment

David Roodman[*]

Open Philanthropy

August 25, 2023

Abstract

This paper reanalyzes Khanna (2023), which studies labor market effects of schooling in India through regression discontinuity designs. Absent from the data are four districts close to the discontinuity; restoring them cuts the reduced-form impacts on schooling and log wages by 57% and 63%. Using regression-specific optimal bandwidths and a robust variance estimator clustered at the geographic unit of treatment makes impacts statistically indistinguishable from 0. That finding is robust to varying the identifying threshold and the bandwidth. The estimates of general equilibrium effects and elasticities of substitution are not unbiased and have effectively infinite first and second moments.

Introduction

Khanna (2023)—K23 for short—is perhaps the only study to estimate both partial- and general-equilibrium effects of education expansion in a developing country using a high-credibility research design. By exploiting a discontinuity in eligibility, K23 estimates that a primary school funding program in India increased children's future wages by 13.4% per year of additional schooling, and would have increased them 6.6 percentage points more had not the expansion of the supply of skilled workers depressed their wages.

This comment assesses the robustness of those and other findings in K23.

India's District Primary Education Programme (DPEP) received $1.5 billion from the World Bank and other donors in the mid-1990s (Jalan and Glinskaya 2013). It distributed the funds to districts—the second-level administrative units in India—to pay for school construction, teacher training, textbooks, and more. Some 271 of approximately 600 districts participated in the program, thereby increasing their primary education spending by an average of 17.5–20% for 5–7 years (Jalan and Glinskaya 2013, p. 4; Pandey 2000, p. 11). One criterion for funding was that a district's female literacy rate, as measured in the 1991 census, should fall below the national average (Department of Education 1995, p. 4), which was 39.29%. As is often the case, treatment assignment did not perfectly follow the mathematical rule laid down as policy. Available documentation does not confirm that administrators imposed a threshold of exactly 39.29%.[1] And districts could instead qualify by having carried out successful Total Literacy Campaigns. K23 therefore estimates impacts of intention to treat, defining it by the 39.29% cutoff.

The empirical analysis of K23 proceeds in four parts:

---

[1] The official program guidelines refer only to "educationally backward districts with female literacy below the national average" (Department of Education 1995, p. 4).

1. Graphical exploration, emphasizing whether discontinuities in treatment, schooling, and wages are present at the cutoff—the latter two outcomes being observed in 2009–10.

2. Regression discontinuity design (RDD) estimation of reduced-form impacts and fuzzy RDD (FRDD) estimation of the return to schooling.

3. Development and application of a procedure to separately estimate partial- and general-equilibrium effects, as well as elasticities of substitution among workers of different ages and schooling levels.

4. Difference-in-difference (DID) estimation; here the samples include essentially all districts in India and are split according to treatment rather than intention to treat.

This paper comments on the K23 data set construction and estimation methodology. Then it checks how addressing the comments affects results. Among the most consequential comments:

1. Four districts with complete data in the primary sources are missing from the main K23 analysis. All lie just above the female literacy threshold. Including them cuts DPEP's apparent impacts on schooling and wages by about 60%.

2. The conventional standard errors reported for the (F)RDD regressions are invalidated by the mean-squared-error-optimal (MSE-optimal) methods first used to choose the kernel bandwidths.[2]

3. The MSE-optimal bandwidths computed for one specification—for the impacts on schooling among those under 35—are copied to all other specifications. In every case, this widens the estimation bandwidth. These positive deviations from each specification's MSE-optimized bandwidths increase expected endogeneity bias while shrinking the reported (invalid) conventional standard errors.

[2] The MSE-optimal algorithms choose two bandwidths: one for estimation and one for bias correction.

4. Although intention to treat and treatment are assigned by district, and although the follow-up survey disproportionately samples some districts, standard errors are not clustered by district.

5. While K23's method for estimating elasticities of substitution and general equilibrium effects is an important contribution, its results are not "unbiased" nor "precisely" measured (K23, pp. 549, 550). The computations depend on treatment assignment and schooling, both of which are endogenous. And the estimates appear to have effectively infinite first and second moments.

6. The DID estimates do not exploit assertedly exogenous variation, being based on treatment rather than intention to treat. For this reason, DID is not discussed further here.

Overall, the empirical findings of K23 appear fragile. Perhaps this is to be expected, for it is not obvious that a 17.5–20% spending boost for 5–7 years would generate robustly detectable impacts 10–15 years later.

Section 1 of this comment covers the reconstruction of the K23 data set. Sections 2 and 3 revisit the graphical and RDD analyses. Section 4 checks the robustness of the findings from the new, preferred specifications. Section 5 turns to the estimation of general equilibrium effects and elasticities of substitution. Section 6 concludes.

1   Data set reconstruction

I reconstruct key K23 variables from primary sources. The original and new data sets mostly match, with a few notable discrepancies.

I restrict myself to the three data sets in the main analysis:

- *Female literacy rates* by district from the 1991 Census.[3]
- *A list of districts receiving DPEP funding.* K23's source is not

---

[3] District-level primary census abstracts for 1991 are at censusindia.gov.in/census.website/data/census-tables, separately for urban and rural areas.

documented. I use a written answer from a minister to a question posed by a member of parliament in 2001. This source agrees with K23 that 271 districts received the program.[4]

- *Data from the 66th round of India's National Sample Survey (NSS)*, which includes individual-level information on age, schooling attainment, and wage earnings.[5] The survey was fielded in July 2009–June 2010, by which time children who may have been directly affected by DPEP in the mid-1990s were young adults.

The three data sets are linked by district, a process that is complicated by many changes in district names and boundaries over time. Between 1991 and 2009–10, some districts subdivided, producing "single-parent districts." A few were formed from fragments of several antecedents, producing "multi-parent districts." I track the evolution from 1991 to 2001 using Kumar and Somanathan (2016, Tables 7, 8, 9d) and from 2001 to 2009 using Wikipedia and reports from the 2011 census. Like K23, I retain single-parent districts since they do not threaten the consistency of the K23 estimators: if a parent district qualified for intention to treat by virtue of low female literacy, we may class exclusive offspring in the same way. Of the 570 districts in K23 sample, 255 are split, single-parent districts. In the new data set 264 of 599 are. *Multi*-parent districts, in contrast, pose a problem, since some contain territory and population from districts with different intention-to-treat. I exclude them unless all but one parent's contribution of population is *de minimus* or its parents had nearly the same female literacy rate.[6]

---

[4] datais.info/loksabha/question/db0cac20ad912c779f1de1c7b7fd60f3/DISTRICT+PRIMARY+EDUCATION+PROGRAMME.

[5] icssrdataservice.in/datarepository/index.php/catalog/89/data_dictionary.

[6] For each 2009 district, I compute the standard deviation of the parents' literacy rates, weighting by their population contribution. Multi-parent districts are retained only if the standard deviation falls below 1%. Five multi-parent districts clear

Table 1 and the top two-thirds of Figure 1 compare the original and new intention-to-treat and treatment variables with reference to the district boundaries of 2009. The two versions of intention-to-treat agree except that K23 version is marked missing in 15 districts for which the new one is not, and the new one is marked missing (because of multiple parents) in 13 districts where the K23 one is not. The treatment variables disagree on 89 of 618 districts.[7]

The outcome variables taken from the 2009–10 NSS differ slightly in construction. The new schooling attainment variable is 0 years rather than 1 or 2 for individuals self-reporting as "literate without formal schooling." Some 2,000 of 72,000 respondents in the sample report wage earnings from multiple work activities; in the new data, wages are summed from all activities rather than the first-listed one. As well, wage earnings are divided by the fraction of the week worked, to better estimate wage rates as distinct from total wage earnings. On overlap, the original and new age variable match exactly, the female literacy and years of schooling variables are correlated .9998, and the log wage variables are correlated .9676.

More important than the difference in construction of the NSS variables is the difference in sample. Four districts are absent in the K23 data set yet present, with complete observations, in the new data. They can be discerned in the bottom-left of Figure 1. Though geographically dispersed, they are statistically close in one sense. All had female literacy rates just above the formal DPEP eligibility threshold of 39.29%: Aurangabad district in Maharashtra state (at 39.64%), Tamenglong in Manipur (39.68%), Cuddalore in Tamil Nadu (39.70%),

---

that threshold: Ambedkar Nagar, Fatehpur, and Sultanpur in Uttar Pradesh; Jhajjar, Haryana; and Nawanshahr, Punjab. The threshold was 3% in an earlier version of this paper; I thank Gaurav Khanna for critiquing that choice.

[7] Because of district subdivisions, more than 271 districts-of-2009 are classed as treated in the revised data set.

and Latur in Maharashtra (39.74%).[8] All but Tamenglong received DPEP funding, so 3 of 4 represent violations of the program rule that undergirds the K23 regression discontinuity analysis. As would be expected, and as will be seen, adding these districts to the regression samples reduces DPEP's apparent impact on schooling. As a result, they also weaken the first stage of FRDD estimates of the return to schooling. They do so with force because they are close enough to the discontinuity to receive near-maximum weight.

Table 1. Cross-tabulation of original and new intention-to-treat and treatment variables

Intention to treat (female literacy in 1991 below 39.29%)

New	Original			Total
	0	1	Missing	
0	231	0	7	238
1	0	327	8	335
Missing	6	7	32	45
Total	237	334	47	618

Treatment (received DPEP funding)

New	Original			Total
	0	1	Missing	
0	303	4	24	331
1	45	219	16	280
Missing	0	0	7	7
Total	348	223	47	618

Note: Unit of observation is the 2009 district.

[8] The bottom left of Figure 1 shows *five* districts absent in the original and present in the new data set. But the fifth (and westernmost), Gandhinagar, is also excluded from the new analysis because it is multi-parent. And the population-weighted average 1991 female literacy rate of its parents is 65%, which is far from the identifying program threshold.

Figure 1. Overview of original and new data sets

Intention to treat

Treatment

Follow-up

Original data set   New data set
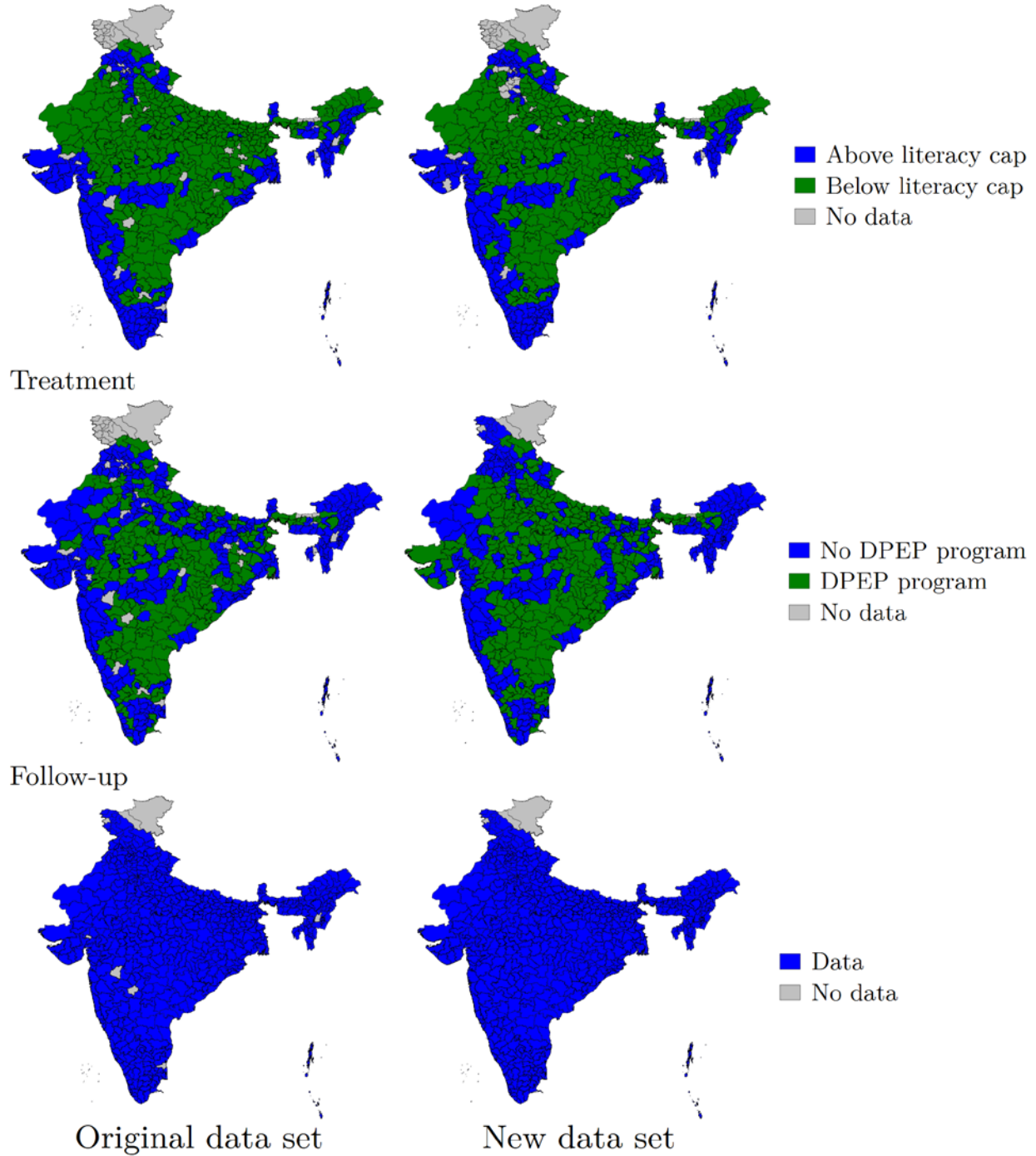
Notes: Original data set is from K23. Sources for the new data set are described in text. "Intention to treat" is whether a district's female literacy rate in 1991 was below the national average of 39.29%. Treatment is whether a district received funding under the District Primary Education Programme. Follow-up is whether the district appears in the extract from the National Sample Survey, 66th round, 2009–10. Underlying map from Minnesota Population Center (2020).

2   Graphical analysis

K23 begins the empirics with graphs, most made with the rdplot program for Stata (Calonico, Cattaneo, and Titiunik (CCT) 2015). The program plots scatters of a dependent variable against a running variable after averaging the first within bins of the second. By default, the bins are chosen by a theoretically grounded, data-driven algorithm, and cover the full range of the running variable. The program then overlays polynomials fits, quartic by default, to the unbinned data on each side of the threshold. The user may override the defaults governing the binning, the polynomial fits, and sample. Indeed, in K23, polynomial fits are quadratic rather than quartic, bin counts are hand-coded, and the sample is sometimes restricted.

Figure 2 below shows the effects of reverting to the rdplot defaults and switching to the revised data set. Plots in the first column are exact replications of the original; their dependent variables are DPEP program status as of 2001 and years of schooling and log wage earnings as of 2009–10. Samples for the latter two are restricted to within 20 points of the threshold. The plots leave the impression of discontinuities with the predicted sign in all three variables. In the case of wages, the impression depends heavily on the curve fits (bottom left pane). The dots alone do not clearly depict a discontinuity. The second column merely rescales the plots in the first, vertically and horizontally, in order to facilitate comparison with those that follow. The third reverts to rdplot's defaults, with one caveat: instead of accepting the default "variance-mimicking" method for choosing the number of equally spaced bins, the alternative is invoked, which asymptotically minimizes the mean squared error with respect to the regression function. This algorithm more closely approximates the K23 binning, while still avoiding manual overrides. These changes do not reduce the appearance of a discontinuity for DPEP assignment (top row). Nor do they definitively alter the picture for schooling. But they do erase the impression of an impact on wages.

The final column of Figure 2 switches to the new data set, in particular restoring four absent districts; this narrows the apparent discontinuity for years of schooling as well (middle row, last column).

Thus, to the extent that the K23 discontinuity plots support an inference of significant program impacts, it is worth bearing in mind that this impression derives from overriding defaults and the absence of certain observations near the threshold.

Figure 2. Global polynomial fits of program assignment, schooling attainment, and log wages to female literacy
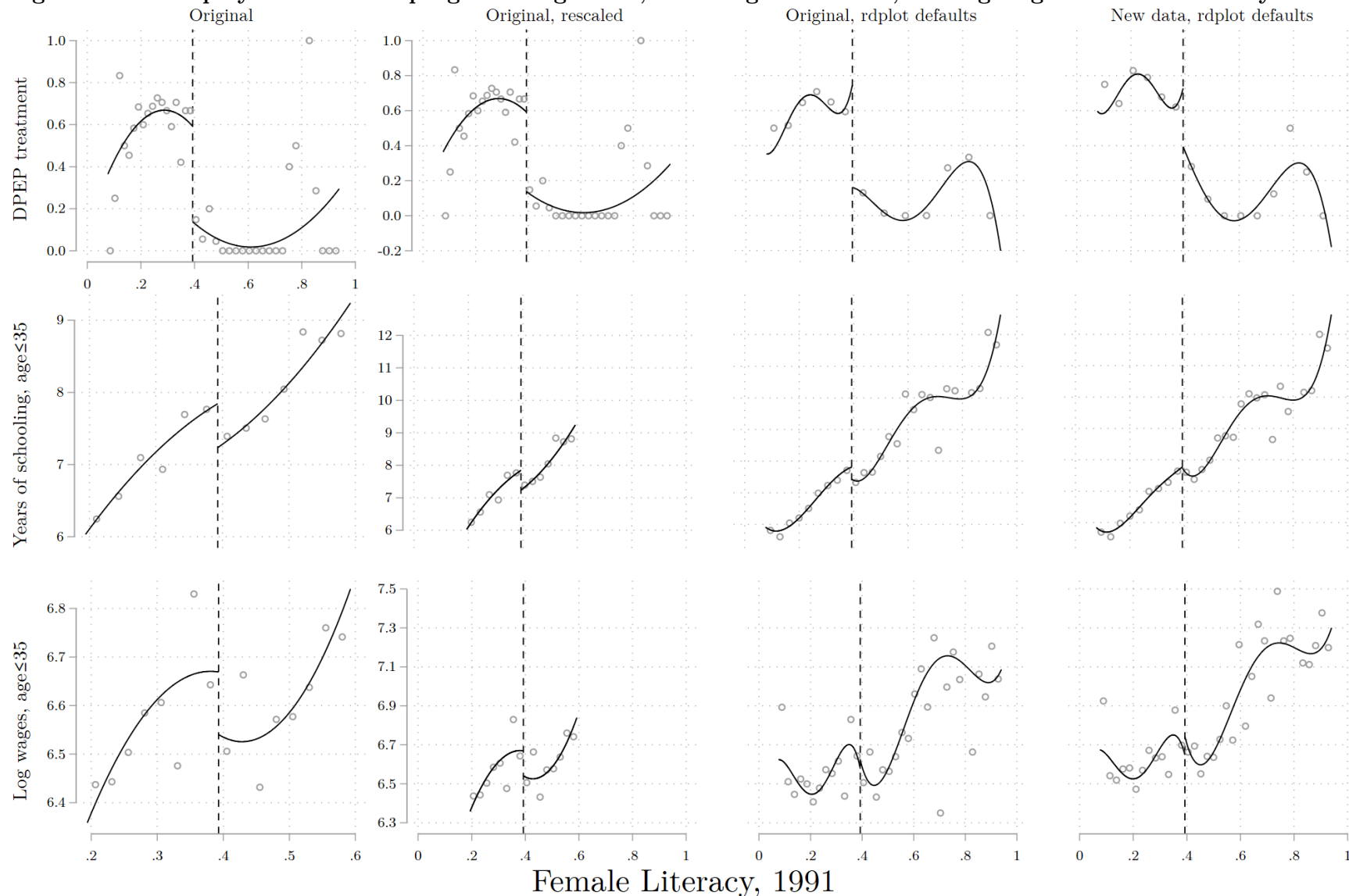
Notes: Each plot is made by the rdplot program for Stata (Calonico et al. 2017). Observations are averaged within bins before scatter-plotting. Polynomials are fit to the un-binned data on each side of the 39.29% female literacy threshold. Plots in the first column exactly match ones in Khanna (2023, Figures 1 and 2). They entail overriding rdplot defaults for the order of the polynomial fit (2 instead of 4) and the number of bins (chosen by hand rather than by a data-driven algorithm); and, in rows 2 and 3, restricting the sample to within 20 points of the cut-off. Plots in the second column are rescaled versions of those in the first. Plots in the third are made by accepting rdplot's choice of mean-

13

3 Regression discontinuity estimates

The K23 RDD regressions estimate the impact on schooling and wages of trans-
iting the 39.29% female literacy threshold. FRDD regressions esentially take ra-
tios of those results in order to estimate the return to schooling.[9] This section
comments on the K23 approach, then modifies the regressions in ways motivated
by the comments as well as by the data revisions described in section 1.

K23 performs RDD and FRDD with the rdrobust program for Stata (CCT
2014b; Calonico et al. 2017), which implements the "robust RDD" method of
CCT (2014a), as revised in Calonico, Cattaneo, and Farrell (2018) and Calonico
et al. (2019). By default, the CCT procedure follows a sequence:

1. Given a specification, a bandwidth is chosen that balances bias and vari-
   ance in order to asymptotically minimize mean-squared error (MSE).
   (Raising the bandwidth increases sample size, adding precision; lowering
   it reduces the scope for bias by hewing closer to the discontinuity that is
   the basis for credible causal identification.) The conditioning information
   for the bandwidth optimization includes a sample; a response variable; a
   running variable; a threshold; the order of the polynomials in the run-
   ning variable to be controlled for on each side of the threshold; a kernel
   type such as uniform or triangular; any observation weights; and a
   choice of variance estimator, such as clustered.

2. Since these MSE-minimizing bandwidths "are by construction not valid
   for conventional confidence intervals" (CCT 2014a, p. 2295), "robust"
   standard errors are computed (CCT 2014a).

[9] The FRDD estimates can be thought of loosely as local average treatment ef-
fects (LATEs; Imbens and Angrist 1994)—impacts at the thresholds among those
who attended school longer because of DPEP—with the caveat that the DPEP pro-
gram may have been large enough to generate GE effects. GE effects undercut the
Stable Unit Treatment Value Assumption (SUTVA), i.e., the premise that different
individuals' outcomes are independent.

3. The bias resulting from the MSE minimization procedure is estimated and subtracted from the point estimate and confidence interval bounds. The increases the variance of the point estimate, but improves inference in the sense of generating confidence intervals with better size.

K23 departs from this sequence in two ways. First, it stops at step 1. It uses the CCT bandwidth selection algorithm as an input to conventional RDD (as well as the Imbens and Kalyanaraman (2012) bandwidth algorithm, as a check). It then reports conventional standard errors. Cattaneo, Idrobo, and Titiniuk (2019, p. 59) calls this combination "methodologically incoherent." Second, K23 overrides the default behavior of the estimation software by using the same bandwidths for all regressions—that is, all outcomes (years of education, completion of upper primary school, log earnings) and all subgroups (young and old). Because the MSE-optimized bandwidths of all the other combinations are smaller, as will be shown, these bandwidth overrides should be expected to increase bias and reduce apparent variance, as manifested in smaller conventional standard errors.

Two other methodological choices bear mentioning.[10] First, while clustering standard errors by 2009 district in DID regressions (which are not reanalyzed here for lack of exogenous identifying variation), K23 does not cluster in RDD regressions. Yet both are subject to established arguments for clustering in regressions on microdata when the treatment is assigned at the geographic level (Bertrand, Duflo, and Mullainathan 2004; Bartalotti and Brummet 2017; Abadie et al. 2023). Abadie et al. (2023) additionally motivates clustering when, as here, the complex survey design disproportionately samples some districts. Not clustering can be expected to lead to undersized confidence intervals.[11] Second,

---

[10] Roodman (2022) makes these same points with respect to Duflo (2001).

[11] Since the running variable, female literacy, is continuous and unique for each 1991 district, I cluster by the running variable as an expedient to clustering by 1991

sampling weights are not incorporated. Yet the NSS sampling program is presumptively endogenous to the outcomes studied in K23. In the second stage, households were stratified by assets and economic activity (in rural areas) or per-capital expenditure (in urban areas).[12] Not factoring in the weights provided by the statistical agency in order to correct for this endogenous sampling renders the estimators inconsistent (Hausman and Wise 1981; Solon, Haider, and Wooldridge 2015). On the other hand, as will be seen, the distribution of the weights is so skewed that incorporating the weights can greatly increase the variance of the estimators. Because of the potentially steep bias-variance tradeoff, I often report results with and without weighting.

The methodological comments above and the data comments in section 1 motivate revisions to the K23 analysis. I replicate key RDD estimates, then introduce the changes in isolation, then make all at once. The results appear in Table 2. There are panels for the RDD schooling and log wages regressions as well as FRDD returns-to-schooling regressions. The panels are also split by age, with separate estimates for young and old, i.e., under and over 35.

The first column replicates results in K23 Table 1. These suggest that living in a district just below DPEP's female literacy eligibility threshold lifted schooling attainment among under-35s by .72 years (standard error .199) and log wage earnings by .112 (.031) as of 2009–10. The FRDD estimate of the return to schooling is .155 log points per year (.043). Impacts on the old, in the lower half of the column, are statistically close to zero.

Next, the specifications are modified by allowing rdrobust to optimize the bandwidth in order to minimize the mean-squared error of the estimator, which

district. Kolesár and Rothe (2018) warns against clustering by the running variable when it is discrete. The warning does not apply here because the running variable takes some 200 values within a typical K23 bandwidth of 0.1, and because clustering here increases standard errors rather than decreasing them as in the degenerate example of Kolesár and Rothe.

[12] microdata.gov.in/nada43/index.php/catalog/124/sampling

is the default behavior (column 2). The schooling estimate for the young is essentially unchanged by construction, as its bandwidth is used in all the K23 regressions.[13] But all other bandwidths shrink. The reduced-form wage regression sees its bandwidth go from .103 to .033 and its point estimate collapse from .112 to −.043.[14] In the sample of old workers, the bandwidth for the reduced-form wage regression contracts to .037 and the point estimate falls from −.011 in K23 to −.188, which is highly significant despite a doubling in the standard error. At face value, these results suggest that DPEP did not raise pay for the young yet decreased pay for the old. More certainly, they demonstrate that the K23 wage impact estimates are sensitive to bandwidth choice.

The third column reverts to K23's bandwidth and instead switches to the CCT robust, bias-corrected confidence intervals, for the sake of more reliable inference. For presentational consistency, these confidence intervals are reported in the table as bias-corrected point estimates along with standard errors. If one is interested purely in point estimates, this new, bias-corrected value for the schooling-among-the-young regression (first row, third column) is not obviously superior to the MSE-optimized estimate (first column); for while bias correction should reduce bias, it increases variance more than enough to drive the estimate away from the estimated MSE minimum. However, in all other regressions (later rows), the bias-corrected estimates in the third column are more strongly preferred, for, as argued above, their non-bias-corrected comparators in the first column are *not* (asymptotically) MSE-minimizing and indeed likely contain excess bias. If one's priority is inference, e.g., with regard to the null of zero impact, then the main thing to note is that the robust, bias-corrected confidence

---

[13] The bandwidth changes slightly because I switch to the 2017 version of rdrobust.

[14] For consistency with the next section, standard errors in these regressions are computed using the jackknifed heteroskedasticity-robust estimator rather than the default nearest-neighbor method. The latter would produce unstable results in the FRDD regressions in the next section.

intervals still tend to reject that null, if not quite as strongly as in K23.

Column 4 clusters the standard errors by district; they double or triple.

Column 5 instead factors in survey weights. The weights' distribution is extremely right-skewed, with minimum .7, maximum 99,134, skewness 4.7, and kurtosis 42.6. The literature provides ad hoc procedures for trimming weights (Kerckhove, Mohadjer, and Krenzke 2014). I employ a procedure that performs well in Potter and Zheng (2015)'s simulations, which is to trim high values to the median weight plus 5 times the interquartile range. Weighting largely bolsters the K23 findings for the young, while surprising with estimates about as large for the old: seemingly, intention-to-treat raised schooling by .485 years and log wages by .16 among the old. I am uncertain whether to favor weighting when the weights are so skewed, the trimming ad hoc, and the results hard to explain.

Columns 6 and 7 instead transfer the K23 specification to the new data set. The slight modifications of the schooling and wage variables, along with a minor sample expansion made possible by more complete linkage of 2009 districts to their 1991 parents, make little difference (column 6). However, additionally restoring the four districts absent from K23's NSS data, the ones bunched just above the threshold, dramatically reduces apparent impacts on the young (column 7). The schooling estimate drops from .720 to .308 years and the log wage estimate from .112 to .042. The FRDD estimate of the return to schooling only falls from .155 to .136. But its standard error doubles, to .088.

The last two columns of Table 2 show the effect of combining these changes—robust, clustered standard errors, regression-specific bandwidths, the new data set in full. Regressions are unweighted in the penultimate column and weighted in the last. All results are hard to distinguish from zero.

Figure A–1 and Table A–1 in the appendices report results from more combinations of data and method. Overall, the K23 RDD results are not robust to changes that ought to bring improvements in bias, consistency, and inference.

Table 2. Replication and revision of Khanna (2023) RDD regressions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Regression-specific bandwidth	Robust, bias-corrected		Sampling weights	New data without 4 districts absent in original	New data in full	All changes but weights	All changes
	Original			Clustered					
Impact of intention-to-treat on schooling: young									
Estimate	.720***	.708***	.674***	.720*	.692**	.727***	.308	.240	.150
	(.199)	(.208)	(.235)	(.435)	(.296)	(.197)	(.189)	(.505)	(.563)
Bandwidth	.103	.101	.103	.103	.103	.103	.103	.112	.141
Observations	10,175	10,067	10,175	10,175	10,175	10,221	10,485	12,332	15,245
Impact of intention-to-treat on log wages: young									
Estimate	.112***	−.043	.062*	.112	.202***	.064**	.042	−.106	−.042
	(.031)	(.058)	(.037)	(.102)	(.045)	(.029)	(.027)	(.112)	(.129)
Bandwidth	.103	.033	.103	.103	.103	.103	.103	.067	.069
Observations	10,175	3,963	10,175	10,175	10,175	10,221	10,485	7,360	7,614
Impact of schooling on log wages: young									
Estimate	.155***	.099**	.096*	.155	.292***	.088**	.136	.030	.298
	(.043)	(.045)	(.051)	(.131)	(.106)	(.035)	(.088)	(.336)	(.524)
Bandwidth	.103	.085	.103	.103	.103	.103	.103	.099	.117
Observations	10,175	8,569	10,175	10,175	10,175	10,221	10,485	10,344	12,760
Impact of intention-to-treat on schooling: old									
Estimate	−.086	−.075	−.367	−.086	.485	.0043	−.147	−.387	.609
	(.218)	(.262)	(.258)	(.692)	(.322)	(.206)	(.198)	(.689)	(.854)
Bandwidth	.103	.078	.103	.103	.103	.103	.103	.119	.102
Observations	11,293	8,895	11,293	11,293	11,293	12,631	13,005	15,971	12,962
Impact of intention-to-treat on log wages: old									
Estimate	−.011	−.188***	−.085*	−.011	.16***	−.050	−.037	−.111	.076
	(.037)	(.067)	(.044)	(.148)	(.055)	(.033)	(.031)	(.143)	(.158)
Bandwidth	.103	.037	.103	.103	.103	.103	.103	.081	.095
Observations	11,290	4,907	11,290	11,290	11,290	12,632	13,006	10,597	11,965
Impact of schooling on log wages: old									
Estimate	.129	.589	.556	.129	.332**	−11.8	.255	.286	.185
	(.303)	(.696)	(.358)	(1.01)	(.166)	(578)	(.251)	(.651)	(.236)
Bandwidth	.103	.063	.103	.103	.103	.103	.103	.112	.144
Observations	11,290	7,601	11,290	11,290	11,290	12,631	13,005	15,172	19,130

Notes: Each cell reports results from a distinct regression. Those in column 1 exactly match results in Khanna (2023, Table 1, columns 1 and 2). Subsequent columns introduce one modification at a time, and then, in the final two columns, most or all at once. Column 2 switches to the CCT (2014) robust, bias-corrected method. Column 3 incorporates regression-specific optimal bandwidths. Column 4 clusters standard errors by district. Column 5 introduces survey weights. Columns 6–7 switch to the new data set, except that column 6 excludes four districts near the threshold also excluded in K23. Columns 8–9 combine all changes except that column 8 does not weight. Standard errors in parentheses. *$p < .1$. **$p < .05$. ***$p < .01$.

4 Systematic robustness testing

The previous section showed that the estimates produced by the K23

specifications are sensitive to the omission or inclusion of four districts close to
the identifying cutoff, whose absence is not explained. A signed report by
Gaurav Khanna on an earlier version of this paper does not precisely explain
that absence, but demonstrates that the sensitivity is an aspect of a larger phe-
nomenon. There are other ways to drop districts near the threshold that lead to
results similar to those in K23. One may drop just the district nearest the cutoff;
or exclude all within .4 percentage points of the cutoff ("donut RDD"); or ex-
clude only split, single-parent districts in the donut hole; or switch from the tri-
angular to the uniform weighting kernel in order to deemphasize districts nearest
the cutoff.[15]

The generalization of the sensitivity finding poses a conundrum. On the one
hand, it may point to an empirical regularity, which deserves investigation. On
the other, RDD promises clean identification precisely through a focus on the
observations closest to the threshold, at least when fitting with local polynomials
as Gelman and Imbens (2019) and Cattaneo, Idrobo, and Titiniuk (2019) recom-
mend. Such observations should only be deleted according to rules that are econ-
ometrically grounded and minimally arbitrary. The modifications listed above
arguably do not meet those criteria. For instance, the usual rationale for donut
RDD, that the running variable was plausibly manipulated near the cutoff (Al-
mond and Doyle 2011; Barreca, Lindo, and Waddell 2016; Catteneo, Idrobo, and
Titiniuk 2019) does not hold. Census takers in 1991 almost certainly did not ma-
nipulate their counts in anticipation of a donor-driven education finance pro-
gram not yet formulated. Even if they foreseen the program, they could not
have known what national average would emerge from the final census

---

[15] While the Khanna report does not challenge most of the methodological com-
ments in section 3 above, it does not incorporate them either. That is, MSE-mini-
mizing bandwidths are not computed anew for each regression when changing the
sample, dependent variable, or kernel; conventional, non-robust standard errors are
reported; and standard errors are not clustered by district.

tabulation. And as argued in section 1, the 255 split districts in the K23 sample do not by their presence undermine the consistency of FRDD; the case for selectively deleting some is also therefore weak. Finally, the use of the uniform kernel, especially when coupling it with a bandwidth optimized for the triangular, seems questionable. The triangular kernel is the default in the dominant RDD package, rdrobust, "because when used in conjunction with a bandwidth that optimizes the [MSE], it leads to a point estimator with optimal properties" (Cattaneo, Idrobo, and Titiniuk 2019, p. 37). The only remaining rationale for these sample modifications is to explore specification space in an ad hoc way.

A more rigorous way to engage with the influence of "inliers" starts with recognizing that the identifying threshold used in the econometrics may not match the administrative threshold. We can compute that 129,752,482 of 330,286,606 Indian females over age 6—39.29%—were marked as literate in the 1991 census. It does not follow that the officials administering DPEP cleaved the field at precisely that percentage. To my knowledge, official documents do not specify a number, instead referring only to "the national average" (Planning Commission 1994; Department of Education 1995). Plausibly, administrators implemented a "39% or less" rule in such a way that even 39.99% qualified. That would have put the cutoff at 40%. Or perhaps in a game of administrative telephone, or in the face of the difficulty of rejecting districts that almost qualified under a 39% rule, "below 40%" mutated into "40% or below." That would have lifted the threshold to 41%. If the true and modeled thresholds differ, that would cause some observations nearest the cutoff to be misclassified on intention to treat, and could add substantial, influential measurement error. It is worth checking for sensitivity to varying the identifying threshold.

Figure 3 examines the basic question of threshold placement by plotting the fraction of districts with female literacy in $[.10, .11), [.11, .12)$, etc., that received DPEP funding. The threshold most consistent with the data is .41. The treatment rate does not decline as one approaches .41 from the left. But it plunges to

0 for the 22 districts between .41 and .43, returns to a lower level between .43 and .49, then falls back to 0.[16] Since it is plausible on priors that the administrative threshold was a whole percentage, Bayesian reasoning puts nontrivial weight on .41 as the value closest to the truth.

One potential drawback of shifting the econometric threshold to, say, .41 is that, while it may undermine consistency even as it adds realism. Officials may have tuned the administrative threshold in response to traits of marginal districts, such as poverty or political power. That would undermine the local exogeneity of intention to treat. To examine this concern, I perform an invalidation test, checking for discontinuities in several predetermined, district-level variables at the candidate thresholds of .3929, .4, and .41. In particular, I check for discontinuities in district population, area, number of residential houses, number of households, and the share of those over age 6 engaged in agricultural work; all come from the same district-level primary census abstracts as the female literacy figures.[17] See Table 3. Each check is performed via an RDD regression, using the robust, bias-corrected standard errors for inference (Catteneo, Idrobo, and Titiniuk 2019, §5.1). The unit of observation is the 1991 district. Under the null that intention to treat is locally exogenous at each candidate threshold, districts just on either side will not differ systematically on any traits. The high p values in Table 3 fail to reject that null.

Having established the viability of alternative econometric thresholds, I next examine the sensitivity of preferred estimates in the previous section to varying the threshold. As in column 8 of Table 2, these estimates are computed on the full sample with regression-specific, MSE-optimized bandwidths and district-

---

[16] All of the DPEP districts with literacy near .8 are in Kerala; they qualified on the alterative criterion of having run successful Total Literacy Campaigns (Pandey 2000, p. 15).

[17] The census abstracts decompose population by age only to the extent of reporting the 6-and-under total, separately for males and females.

clustered, robust, bias-corrected point estimates and confidence intervals. The estimates do not incorporate survey weights. (Results with survey weights do not differ qualitatively and are omitted for brevity.) The upper left pane of Figure 4 starts by testing for evidence of an impact of intention to treat on treatment. The estimated impact on treatment peaks, in size and significance, when anchoring at .41. The lower panes in the first column of the figure confirm the robustness of the previous section's finding of no clear, positive impact on schooling or wages—even when using the higher-powered .41 threshold.

In similar fashion, I next check for sensitivity to varying the bandwidth while fixing the threshold at .3929 or .41. Like the threshold, the MSE-optimized bandwidth may be estimated with error since the CCT algorithm is only guaranteed by theory to work asymptotically. The bandwidth also affects the relative influence of the most central observations.[18,19] The bandwidth sensitivity results appear in the second and third columns of Figure 4. In both, vertical lines now show the MSE-optimized bandwidths. No surprises emerge. Regardless of bandwidth, intention-to-treat is more strongly related to treatment under the .41 definition than the .3929 definition. Impacts of intention to treat on other variables are never strongly positive. The same goes for the impact of schooling on wages.

---

[18] Khanna's report on a previous version of this paper also checks sensitivity to bandwidth choice, but with non-robust, non-clustered variance estimates.

[19] Ideally, these tests would set the bandwidth for the estimation step while allowing rdrobust to choose the bandwidth for the bias-correction step through a data-driven algorithm. However, if one sets the estimation bandwidth to a specific value, rdrobust assigns the same value to the bias-correction bandwidth. To increase realism, I therefore manually assign both. I make the ratio between the two the same as when both are chosen automatically by the CCT algorithm.

Figure 3. Fraction of 2009 districts receiving DPEP treatment
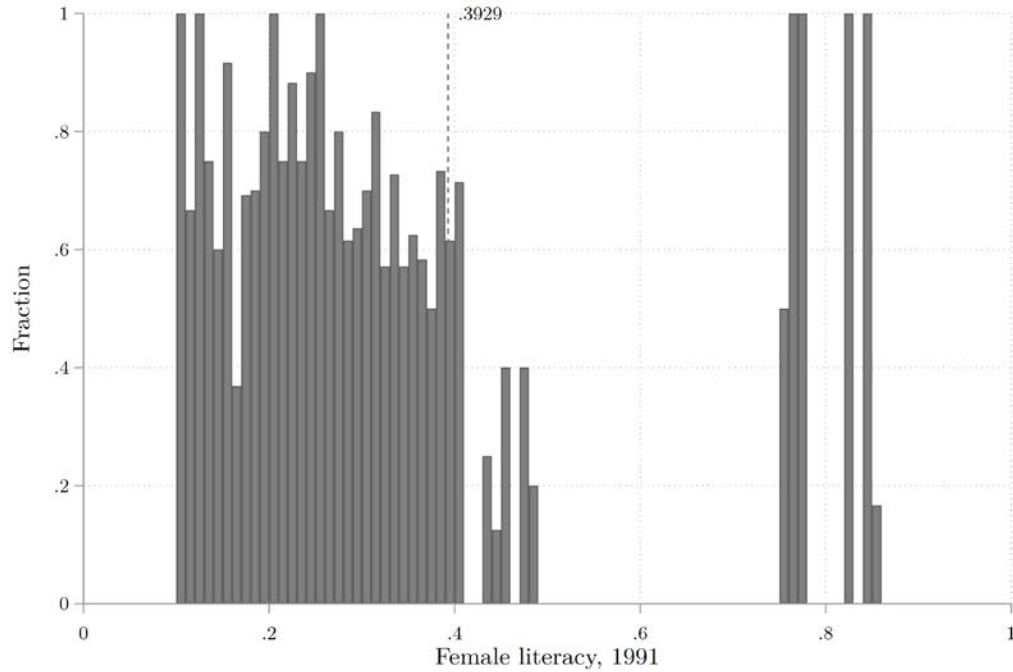
Note: Excludes multi-parent districts using the criterion in footnote 6.

Table 3. Tests for discontinuities in predetermined variables at various female literacy thresholds

Female literacy threshold	Population	Area	Houses	Households	Agricultural worker share
.3929	.93	.57	.69	.71	.40
.4	.79	.49	.80	.78	1.00
.41	.43	.83	.55	.57	.61

Notes: Each cell reports a p value from a robust RDD estimate of the change in a district-level variable observed in 1991 census, at the indicated female literacy threshold. The unit of observation is the 2009 district. The denominator for last the column is number of individuals over age 6. The robust "HC$_3$" variance estimator is used. Excludes multi-parent districts.

Figure 4. Sensitivity of results from preferred unweighted specification to varying threshold and bandwidth
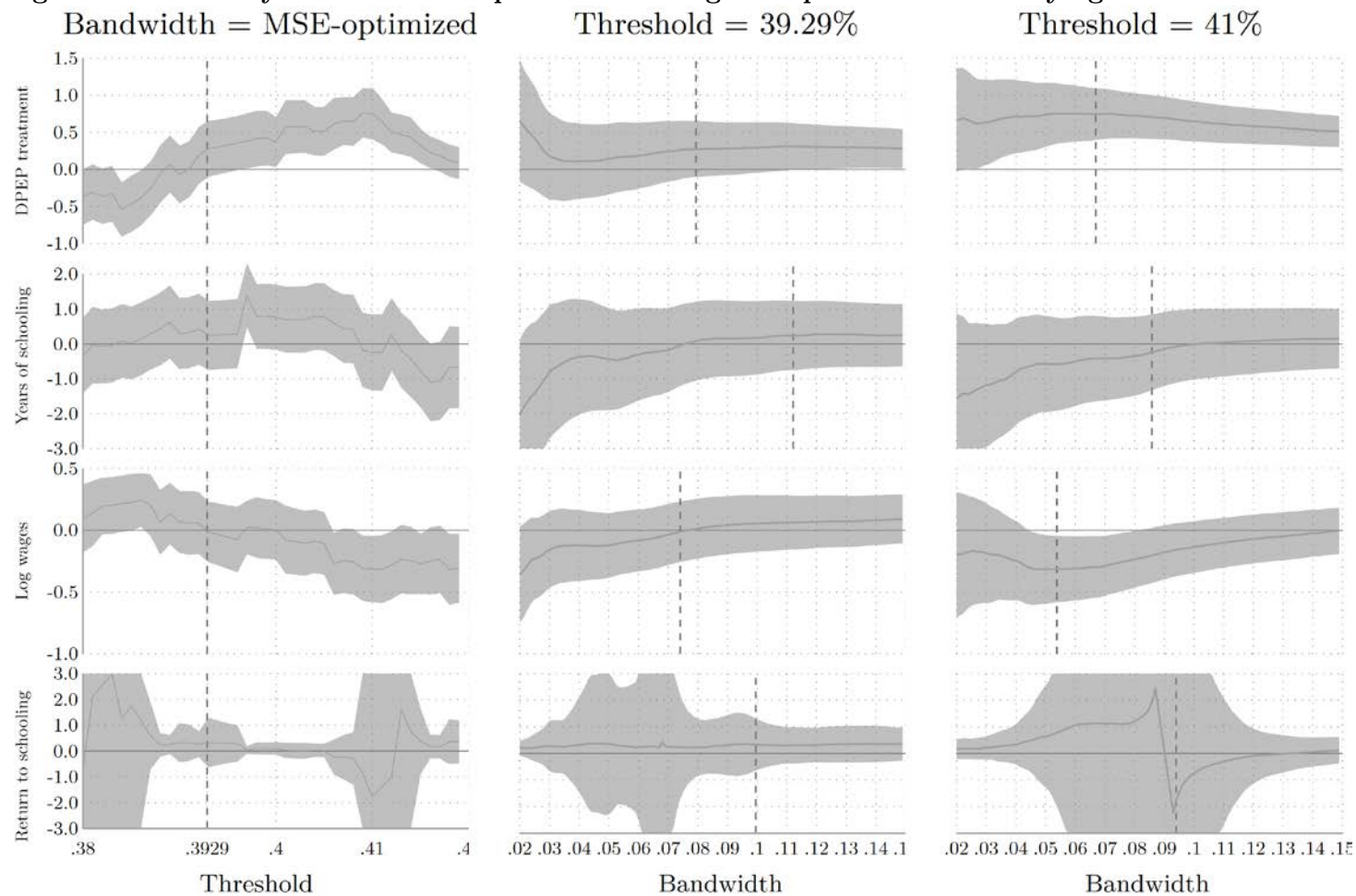
Notes: Each pane shows how the estimated impact on the indicated outcome varies with the threshold used to define intention to treat or the bandwidth. Grey regions show 95% confidence intervals. Results in first three rows are sharp RDD while those in the last are fuzzy RDD. The unit of observation is the district in the first row and the individual in the other rows. In column 1, the identifying threshold is varied, and, at each value, the MSE-optimal bandwidth is used (CCT 2014a). Vertical lines mark the national female literacy rate in 1991. In columns 2 and 3, the threshold is fixed as indicated and the bandwidth is varied. The vertical lines now show the MSE-optimized bandwidth for each specification. Because rdrobust will not separately estimate the bias-correction bandwidth when the estimation bandwidth is manually set, the ratio between the two is chosen to be the same as for when neither is manually set. For legibility, vertical ranges are clipped to within ±3.

5 General equilibrium effects and elasticities of substitution

K23 leverages the large-scale, RDD-framed natural experiment to estimate general equilibrium (GE) effects of increased schooling as well as elasticities of substitution in production among workers of different age and skill groups. Appendix B comments on the methodology for translating RDD results into estimates of GE effects. This section largely leaves the method unchanged and checks for robustness to the modifications to the underlying regressions made in section 3.

One intuitive way to estimate GE effects runs as follows. Suppose we observe that in untreated districts, young, skilled workers—those below age 35 with at least eight years of schooling—attended school 5 years longer than young, *unskilled* workers, and earn 50% more. Then we can estimate the return to schooling across this skilled-unskilled split within untreated districts, call it $\tilde{\beta}_0$, at $50\%/5 = 10\%$ per year. And suppose that the same statistic for treated districts ($\tilde{\beta}_1$) is 8%. Paralleling K23, we interpret the 2-percentage-point difference across the treatment divide ($\Delta_D\tilde{\beta}$), as an estimate of the GE effect, $\Delta_D\beta$. Increasing the supply of young, skilled workers in treated districts reduces their wages, and thus the realized return to schooling, by 2 points.

The K23 method for estimating these effects is less direct. It is designed so that certain FRDD estimates may be plugged into the formulas, with the aim of eliminating bias. And it extends to estimating elasticities of substitution between workers of different skill levels and ages (σ_E and σ_A). See this paper's appendix B. The FRDD regressions differ subtly from those performed in the previous section: now the instrumented treatment variable is not schooling, but DPEP treatment status. This is intended to allow endogeneity-free estimation of differences in wages and schooling across the treatment/non-treatment split.

A point developed in appendix B is that most of the empirical inputs to the K23 GE formulas, including those FRDD results, contain endogeneity after all. While the K23 method is a practical contribution, its outputs are not as

"unbiased" or "precisely estimate[d]" as asserted in the paper (pp. 549, 550).

The K23 public data and code archive documents the estimation of $\tilde{\beta}_0$, $\tilde{\beta}_1$, and $\Delta_D\tilde{\beta}$, but not σ_E and σ_A, which makes the latter hard to reproduce. The robustness testing reported here therefore emphasizes the wage effects. See Table 4, which approximates Table 2 in format. As in K32, bootstrap p values based on 1,500 nonparametric resamples are reported for the one-sided hypothesis that each parameter has sign opposite that estimated; these are shown bracketed. In addition, bootstrap standard errors appear in parentheses.

Results in the first column of Table 4 almost exactly match those in K23's Table 3; the computation of $\Delta_D\tilde{\beta}$ differs slightly because K23 reports medians of 1,500 bootstrap estimates rather than full-sample estimates. Among the young, the return to skill in non-program districts is put at 19.9% in non-program villages and 13.4% in program districts, about one-third less. The standard errors for $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are surprisingly large given that their bootstrap p values (for the null that they are negative) are below .01; this conjunction of results indicates that the bootstrap distribution of these parameter estimates is highly, positively skewed.

Modifying the FRDD regressions that provide input to use the MSE-optimized bandwidth for each regression—and in each bootstrap replication—rather than using the same, hard-coded bandwidths throughout—essentially erases all precision (column 2).[20]

On the other hand, if we instead move to CCT's bias-corrected point estimate this hardly affects results (column 3).

---

[20] By default, rdrobust computes the heteroskedasticity-robust standard errors using residuals with respect to a fit to nearest neighbors. This method tends to fail here when both the dependent variable, DPEP assignment, and the intention-to-treat are constant within districts. Then, the residuals are all zero. This in turn destabilizes the search for optimal bandwidths, when bandwidths are not pre-specified as in the K23 regressions. In this case I switch to the "HC$_3$" variance estimator (MacKinnon and White 1985).

Clustering the bootstrap sampling by district (Field and Welsh 2007) increases the standard errors even more than the first change (using regression-specific bandwidths). Incorporating sampling weights (column 5) also produces large standard errors. Instead moving to the new data while excluding the four districts absent in K23 (column 6) broadly preserves the original results, increasing the estimate of $\tilde{\beta}_0$ from .199 to .306, and the GE effect estimate from a third of $\tilde{\beta}_0$ to half. But including the four absent districts once more explodes the standard errors (column 7). Weighted results produce even more extreme standard errors and are omitted for the sake of space.

Appendix B proposes revisions to the K23 method of estimating GE effects and elasticities of substitutions. These include reverting to the intuitive method described above, since it not obviously more biased by endogeneity. Column 9 of Table 4 layers these revisions onto the methods and data of column 8. Also reported are new estimates of σ_E and σ_A. The standard errors are now huge. All the statistics in this table, as ratios of variables estimated with error, have infinite first and second moments as a matter of theory—and apparently as a matter of practice too.

Table 4. Estimates of general equilibrium effects of DPEP on wages, and of elasticities of substitution

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
						New data without 4 dis-		All changes	All changes but weights,
		Regression-specific	Robust, bias-cor-		Sampling	tricts absent	New data	but	revised
	Original	bandwidth	rected	Clustered	weights	in K23	in full	weights	method
Skill premium per year of schooling, non-DPEP districts ($\tilde{\beta}_0$)	.199	−.0955	.179	.199	−.445	.306	−1.35	−12.1	.00077
	(.39)	(5.5)	(.38)	(15)	(16)	(2.1)	(137)	(4287)	(476)
	[.002]	[.2]	[.012]	[.2]	[.42]	[.0073]	[.24]	[.56]	[.24]
Skill premium per year of schooling, DPEP districts ($\tilde{\beta}_1$)	.134	−.17	.103	.134	−.305	.159	−.812	−13.7	.00064
	(.38)	(5.5)	(.38)	(15)	(15)	(1.5)	(136)	(4476)	(276)
	[.0067]	[.11]	[.057]	[.26]	[.29]	[.032]	[.26]	[.46]	[.39]
Difference ($\Delta_D\tilde{\beta}$)	−.0646	−.0742	−.0762	−.0646	.139	−.147	.542	−1.53	−.00013
	(.032)	(.1)	(.038)	(1.1)	(2.9)	(1.1)	(10)	(547)	(549)
	[.01]	[.047]	[.011]	[.15]	[.59]	[.015]	[.2]	[.35]	[.37]
Elasticity of substitution, skill groups (σ_E)	4.24								.0198
									(120852)
									[.55]
Elasticity of substitution, age groups (σ_A)	5								−13.1
									(2372)
									[.62]

Notes: "Original" estimate of $\Delta_D\tilde{\beta}$ differs slightly from that in K23 (Table 3), which reports medians of 1,500 bootstrap estimates rather than full-sample estimates. Original values for σ_E and σ_A are from notes to K23, Table 3. Definitions of subsequent columns, except for the last, correspond to those in Table 2. Estimates in the last column are derived with a revised method described in appendix B. σ_E and σ_A are computed only in this case because K23 method for computing them is not precisely documented in text or code. Bootstrap standard errors, based on 1,500 replications, are in parentheses. In clustered estimates, the data are resampled by district. Bootstrap p values for one-sided tests that a parameter has the sign opposite that estimated are in brackets.

6 Conclusion

K23 highlights how general equilibrium dynamics affect the distribution of the impacts of government initiatives. It proposes a method for distinguishing par- tial- and general-equilibrium effects in the context of RDD-based program evalu- ation. And it applies these ideas to the case of an externally financed education program in India in the 1990s. This comment emphasizes the empirics. The esti- mates of impacts, GE effects, and elasticities of substitution are not robust to changes in data and method that ought to improve inference. The preferred esti- mates presented here of impacts on schooling and wages are hard to distinguish from zero. These findings are robust to varying the identifying threshold and the

bandwidth. Since the Indian program only increased school spending 17.5–20% for 5–7 years in affected districts, the difficulty of detecting impacts should perhaps not surprise.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2022. "When Should You Adjust Standard Errors for Clustering?" *Quarterly Journal of Economics* 138 (1): 1–35. https://doi.org/10.1093/qje/qjac038.

Almond, Douglas, and Joseph J. Doyle. 2011. "After Midnight: A Regression Discontinuity Design in Length of Postpartum Hospital Stays." *American Economic Journal: Economic Policy* 3 (3): 1–34. https://doi.org/10.1257/pol.3.3.1.

Barreca, Alan I., Jason M. Lindo, and Glen R. Waddell. 2016. "Heaping-Induced Bias in Regression-Discontinuity Designs." *Economic Inquiry* 54 (1): 286–93. https://doi.org/10.1111/ecin.12225.

Bartalotti, Otávio, and Quentin Brummet. 2017. "Regression discontinuity designs with clustered data." In *Advances in Econometrics: Vol. 38—Regression Discontinuity Designs: Theory and Applications*, eds. M. D. Cattaneo and J. C. Escanciano, 383–420. Emerald. https://doi.org/10.1108/S0731-905320170000038017.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *Quarterly Journal of Economics* 119 (1): 249–75. https://doi.org/10.1162/003355304772839588.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. 2018. "Coverage Error Optimal Confidence Intervals for Local Polynomial Regression." *arXiv [econ.EM]*. http://arxiv.org/abs/1808.01398.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik. 2017. "rdrobust: Software for Regression-Discontinuity Designs." *Stata Journal* 17 (2): 372–404. https://doi.org/10.1177/1536867X1701700208.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik. 2019. "Regression Discontinuity Designs Using Covariates." *Review of Economics and Statistics* 101 (3): 442–51. https://doi.org/10.1162/rest_a_00760.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014a. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6): 2295–2326. https://doi.org/10.3982/ecta11757.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014b. "Robust Data-Driven Inference in the Regression-Discontinuity Design." *The Stata Journal* 14 (4): 909–46. https://doi.org/10.1177/1536867X1401400413.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2015. "Optimal Data-Driven Regression Discontinuity Plots." *Journal of the American Statistical Association* 110 (512): 1753–69. https://doi.org/10.1080/01621459.2015.1017578.

Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. 2019. "A Practical Introduction to Regression Discontinuity Designs: Foundations." In *Elements in Quantitative and Computational Methods for the Social Sciences.* Cambridge University Press. https://doi.org/10.1017/9781108684606.

Department of Education, Ministry of Human Resource Development, India. 1995. *District Primary Education Programme: Guidelines.* https://scribd.com/document/410912918/DPEP-guidelines-pdf.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4): 795–813. https://**doi.org/10.1257/aer.91.4.795**.

Field, C. A., and Alan H. Welsh. 2007. "Bootstrapping Clustered Data." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 69 (3): 369–90. https://doi.org/10.1111/j.1467-9868.2007.00593.x.

Gelman, Andrew, and Guido Imbens. 2019. "Why High-Order Polynomials

Should Not Be Used in Regression Discontinuity Designs." *Journal of Business & Economic Statistics: A Publication of the American Statistical Association* 37 (3): 447–56. https://doi.org/10.1080/07350015.2017.1366909.

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.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75. https://doi.org/10.2307/2951620.

Imbens, Guido, and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies* 79 (3): 933–59. https://doi.org/10.1093/restud/rdr043.

Jalan, Jyotsna, and Elena Glinskaya. 2013. "Small Bang for Big Bucks? An Evaluation of a Primary School Intervention in India." Centre for Training and Research in Public Finance and Policy. https://doi.org/10.13140/RG.2.1.2978.7040.

Khanna, Gaurav. 2023. "Large-Scale Education Reform in General Equilibrium: Regression Discontinuity Evidence from India." *Journal of Political Economy* 131 (2): 541–91 https://doi.org/10.1086/721619.

Kinal, Terrence W. 1980. "The Existence of Moments of k-Class Estimators." *Econometrica* 48 (1): 241–49. https://doi.org/10.2307/1912027.

Kolesár, Michal, and Christoph Rothe. 2018. "Inference in Regression Discontinuity Designs with a Discrete Running Variable." *American Economic Review* 108 (8): 2277–2304. https://doi.org/10.1257/aer.20160945.

Kumar, Hemanshu, and Rohini Somanathan. 2016. "Creating Long Panels Using Census Data: 1961–2001." cdedse.org/pdf/work248.pdf.

MacKinnon, James G., and Halbert White. 1985. "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample

Properties." *Journal of Econometrics* 29 (3): 305–25.
https://doi.org/10.1016/0304-4076(85)90158-7.

Minnesota Population Center. 2020. Integrated Public Use Microdata Series, International: Version 7.3 [dataset]. https://doi.org/10.18128/D020.V7.3

Over, Mead. 2022. "grc1leg2. Add a common legend to combined graphs."
http://digital.cgdev.org/doc/stata/MO/Misc.

Pandey, Raghaw Sharan. 2000. "Going to Scale with Education Reform: India's District Primary Education Program, 1995–99." Country Studies. Education Reform and Management Publication Series I (4). World Bank.
http://web.worldbank.org/archive/website00238I/WEB/PDF/INDIA.PDF.

Pisati, Maurizio. 2004. "Simple Thematic Mapping." *Stata Journal* 4 (4): 361–78. https://doi.org/10.1177/1536867X0400400401.

Planning Commission, Government of India. 1994. *Annual Plan 1994–95.*
https://planningcommission.gov.in/docs/reports/publications/anpl_9495e.pdf. Accessed July 1, 2022.

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

Roodman, David. 2022. "Schooling and Labor Market Consequences of School Construction in Indonesia: Comment." *arXiv [econ.GN].*
http://arxiv.org/abs/2207.09036.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66.
https://doi.org/10.1177/0956797611417632.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50 (2): 301–16.
https://doi.org/10.3368/jhr.50.2.301.

Van de Kerckhove, Wendy, Leyla Mohadjer, and Thomas Krenzke 2014. "A Weight Trimming Approach to Achieve a Comparable Increase to Bias across Countries in the Programme for the International Assessment of Adult Competencies." Proceedings of the Joint Statistical Meetings Survey Research Methods Section. American Statistical Association. http://www.asasrms.org/Proceedings/y2014/files/311170_87007.pdf.

Yon, George Vega. 2012 "Introducing PARALLEL: Stata Module for Parallel Computing." http://fmwww.bc.edu/repec/bocode/p/parallel.pdf.
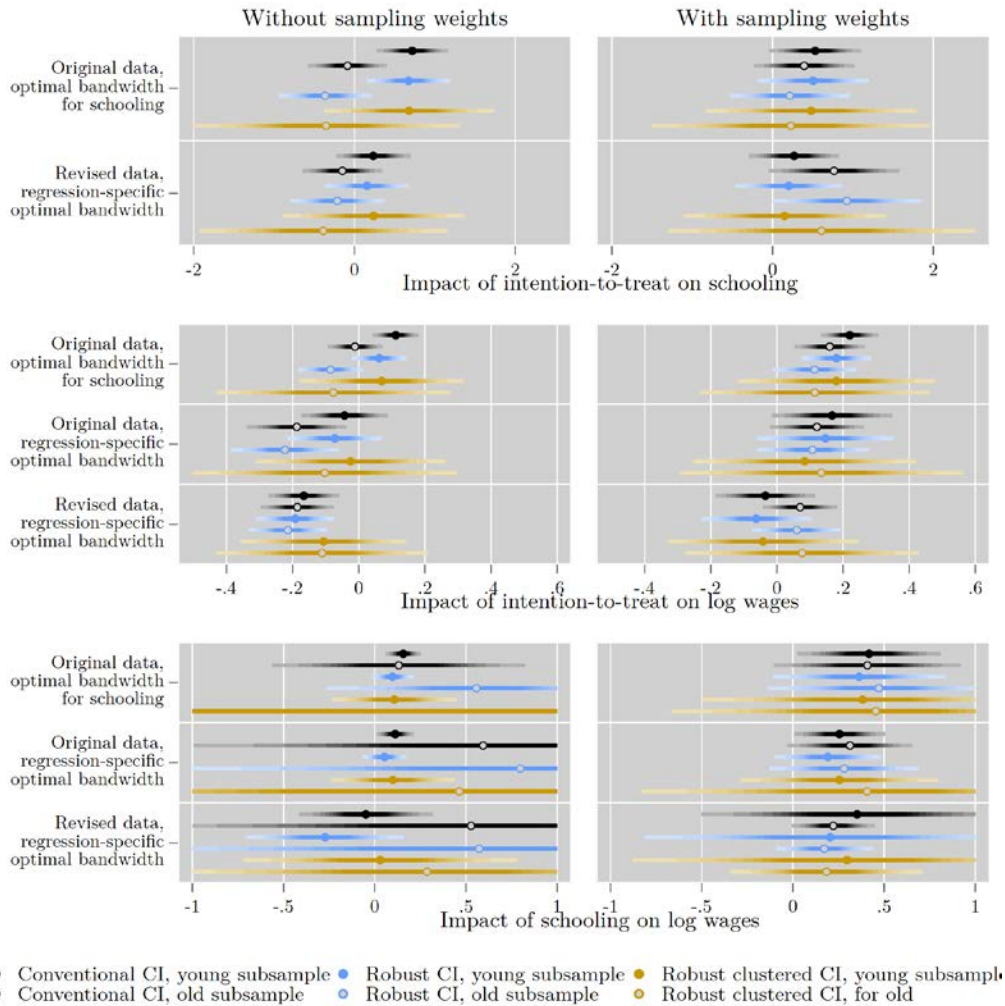
Appendices

A Additional RDD results

Table 2 in section 3 documents the effects of various changes to method and data in key K23 RDD and FRDD regressions. To limit the complexity of the presentation, results are shown from making one change at a time, or making (nearly) all at once. This appendix reports results from more combinations of changes. See Figure A–1 and Table A–1. Much as in Table 2, the three sections of Figure A–1 show RDD estimates of the impact of intention-to-treat on schooling; RDD estimates of the impact of intention-to-treat on log wages; and FRDD estimates of the return to schooling. Estimates from unweighted regressions are on the left and from weighted regressions are on the right. Within each pane, results are computed in up to three ways: using the K23 data and bandwidths; using the K23 data and regression-specific optimal bandwidths; and using the new data and regression-specific bandwidths. Results for the young sample are marked with solid circles and those for the old with hollow circles. The horizontal spikes depict 95% confidence intervals. Confidence intervals in the lower left pane are clipped for the sake of legibility.

Table A–1 arrays the same results in nearly the same format.

Figure A–1. Additional RDD and FRDD estimates of the impact of DPEP

Conventional CI, young subsample ● Robust CI, young subsample ● Robust clustered CI, young subsample
Conventional CI, old subsample ○ Robust CI, old subsample ○ Robust clustered CI, for old

Notes: See appendix text and notes to Table 2 for descriptions of the various specifications.

Table A–1. Additional RDD and FRDD estimates of the impact of DPEP

		Young, un-weighted	Old, un-weighted	Young, weighted	Old, weighted
Impact of intention-to-treat on schooling					
Original data, optimal bandwidth for school-ing	Conventional	.72	−.086	.531	.391
		(.199)	(.218)	(.256)	(.28)
	Robust	.674	−.367	.504	.217
		(.235)	(.258)	(.307)	(.335)
	Bandwidth	.103	.103	.134	.134
	Observations	10,175	11,293	14,277	16,007
	Robust, clustered	.678	−.352	.479	.226
		(.476)	(.743)	(.587)	(.775)
	Bandwidth	.124	.124	.141	.141
	Observations	13,511	15,083	14,802	16,607
New data, regression-specific optimal band-width	Conventional	.235	−.15	.268	.766
		(.206)	(.222)	(.25)	(.364)
	Robust	.156	−.213	.202	.926
		(.234)	(.26)	(.296)	(.414)
	Bandwidth	.084	.079	.128	.064
	Observations	8,869	10,425	14,235	8,985
	Robust, clustered	.24	−.387	.15	.609
		(.505)	(.689)	(.563)	(.854)
	Bandwidth	.112	.119	.141	.102
	Observations	12,332	15,971	15,245	12,962
Impact of intention-to-treat on log wages					
Original data, optimal bandwidth for school-ing	Conventional	.112	−.011	.22	.16
		(.031)	(.037)	(.039)	(.048)
	Robust	.062	−.085	.18	.114
		(.037)	(.044)	(.047)	(.057)
	Bandwidth	.103	.103	.134	.134
	Observations	10,175	11,290	14,277	16,004
	Robust, clustered	.069	−.077	.18	.114
		(.111)	(.159)	(.133)	(.155)
	Bandwidth	.124	.124	.141	.141
	Observations	13,511	15,080	14,801	16,604
Original data, regres-sion-specific optimal bandwidth	Conventional	−.043	−.188	.167	.121
		(.058)	(.067)	(.081)	(.063)
	Robust	−.072	−.223	.147	.107
		(.065)	(.073)	(.093)	(.076)
	Bandwidth	.033	.037	.047	.082
	Observations	3,963	4,907	5,134	9,136
	Robust, clustered	−.025	−.103	.084	.134
		(.128)	(.179)	(.15)	(.191)
	Bandwidth	.069	.082	.074	.097
	Observations	7,354	9,136	7,575	10,763

New data, regression-specific optimal bandwidth	Conventional	−.166 (.048)	−.186 (.05)	−.035 (.067)	.07 (.05)
	Robust	−.193 (.053)	−.215 (.053)	−.062 (.074)	.06 (.06)
	Bandwidth	.029	.034	.036	.082
	Observations	3,934	5,775	4,637	10,707
	Robust, clustered	−.106 (.112)	−.111 (.143)	−.042 (.129)	.076 (.158)
	Bandwidth	.067	.081	.069	.095
	Observations	7,360	10,597	7,614	11,965

Impact of intention-to-treat on log wages

Original data, optimal bandwidth for schooling	Conventional	.156 (.043)	.132 (.31)	.417 (.176)	.408 (.229)
	Robust	.097 (.051)	.558 (.366)	.363 (.211)	.471 (.273)
	Bandwidth	.103	.103	.134	.134
	Observations	10,175	11,290	14,277	16,004
	Robust, clustered	.107 (.152)	10.2 (21.5)	.382 (.393)	.455 (.5)
	Bandwidth	.124	.124	.141	.141
	Observations	13,511	15,080	14,801	16,604
Original data, regression-specific optimal bandwidth	Conventional	.112 (.046)	.594 (.708)	.256 (.111)	.313 (.153)
	Robust	.054 (.054)	.798 (.86)	.192 (.131)	.281 (.183)
	Bandwidth	.087	.063	.087	.099
	Observations	8,846	7,601	8,846	11,018
	Robust, clustered	.099 (.151)	.463 (1.37)	.255 (.242)	.406 (.554)
	Bandwidth	.107	.096	.115	.12
	Observations	10,983	10,396	12,249	13,944
New data, regression-specific optimal bandwidth	Conventional	−.049 (.164)	.528 (.782)	.352 (.382)	.222 (.101)
	Robust	−.272 (.192)	.573 (.93)	.205 (.454)	.172 (.119)
	Bandwidth	.081	.084	.093	.129
	Observations	8,460	10,964	9,605	17,814
	Robust, clustered	.03 (.336)	.286 (.651)	.298 (.524)	.185 (.236)
	Bandwidth	.099	.112	.117	.144
	Observations	10,344	15,172	12,760	19,130

Notes: Standard errors in parentheses. See notes to Table 2 for descriptions of the specifications.

B Revised method for estimating general equilibrium effects and elasticities

This appendix reviews and comments on the K23 methods for estimating GE effects and elasticities of substitution. It proposes the revisions to the method that are incorporated into the final column of Table 4.

B.1 Skill premia

K23's starting point for estimating GE effects is this identity, as presented in equation A.20 of the paper's appendix B.5:

$$\log w_{y,D=1} = \ell_{sy,D=1} \log w_{sy,D=1} + \ell_{uy,D=1} \log w_{uy,D=1}$$

where D is a dummy indicator for treatment assignment; s and u index the skilled and unskilled worker subgroups; $w_{y,D=1}$ is the average wage earned by the young in treatment districts; $w_{sy,D=1}$ and $w_{uy,D=1}$ are the same for the skilled and unskilled subgroups; and $\ell_{sy,D=1}$ and $\ell_{uy,D=1}$ are evidently the skilled and unskilled shares in the labor force of treated districts. The equation is meant to state that a sample average is a weighted average of two subsample averages. As written, it contains two problems. First, earlier in K23 (equation 2), ℓ symbols represent absolute quantities of effective labor in a skill-age cell, not fractional shares; they appear to have been redefined here. Second, the equation is formally incorrect because the logarithm of an average is not the average of logarithms. What appears meant is:

$$\overline{\log w_{y,D=1}} = \ell_{sy,D=1} \overline{\log w_{sy,D=1}} + \ell_{uy,D=1} \overline{\log w_{uy,D=1}}$$

where a bar indicates taking the mean over a group. Stipulating this rewriting, K23's key identities for estimating GE effects (K23 equations 17 and 18) may then be developed precisely, as

$$\Delta_D \overline{\log w_y} = \ell_{sy,D=1}\Delta_D \overline{\log w_{sy}} + \ell_{uy,D=1}\Delta_D \overline{\log w_{uy}} \tag{1}$$
$$+ \Delta_D\ell_{sy} \underbrace{(\overline{\log w_{sy,D=0}} - \overline{\log w_{uy,D=0}})}_{\beta_{y,D=0}}$$

$$= \ell_{sy,D=0}\Delta_D \overline{\log w_{sy}} + \ell_{uy,D=0}\Delta_D \overline{\log w_{uy}} \tag{2}$$
$$+ \Delta_D\ell_{sy} \underbrace{(\overline{\log w_{sy,D=1}} - \overline{\log w_{uy,D=1}})}_{\beta_{y,D=1}}$$

Here, the Δ_D operator takes differences across the treatment/non-treatment split; e.g., $\Delta_D \overline{\log w_y} = \overline{\log w_{y,D=1}} - \overline{\log w_{y,D=0}}$. As in K23, the expressions labeled $\beta_{y,D=0}$ and $\beta_{y,D=1}$ represent returns to skill within treated or untreated districts.[21] It is readily verified that expanding the definitions on the right of both equations and cancelling terms produces the left side. In words, the difference in log wages among the young across the treatment/non-treatment split can be decomposed in two similar ways. Both decompose it into a weighted average of differences for the two skills group along with the mathematical effect of movement between the groups.

As explained in section 4, one could estimate $\beta_{y,D=0}$ and $\beta_{y,D=1}$ directly, as implied by the labeled expressions in (1) and (2). Instead, K23 proceeds by first solving for them in (1) and (2):

$$\beta_{y,D=0} = \frac{\Delta_D \overline{\log w_y} - \ell_{sy,D=1}\Delta_D \overline{\log w_{sy}} - \ell_{uy,D=1}\Delta_D \overline{\log w_{uy}}}{\Delta_D\ell_{sy}} \tag{3}$$

$$\beta_{y,D=1} = \frac{\Delta_D \overline{\log w_y} - \ell_{sy,D=0}\Delta_D \overline{\log w_{sy}} - \ell_{uy,D=0}\Delta_D \overline{\log w_{uy}}}{\Delta_D\ell_{sy}} \tag{4}$$

K23 interprets the difference between these two, $\Delta_D\beta_y$, as the general equilibrium effect of DPEP on the returns to skill among the young. Though not obvious from the above, it can be computed simply as

$$\Delta_D\beta_y = \Delta_D \overline{\log w_{sy}} - \Delta_D \overline{\log w_{uy}} \tag{5}$$

[21] K23 subscripts these symbols with s. I drop that for clarity. The quantities are differences across the two skill groups, so an index referencing one group, the skilled, is not necessary or meaningful.

That is: the difference in the skill premium across the treatment/non-treatment divide is the corresponding difference in skilled wages minus the difference in unskilled wages.

K23 takes the roundabout path represented by (3), (4), and (5) because this allows one to plug in for the Δ_D terms with FRDD estimates, which are seen as unbiased. Unlike in the FRDD estimates of the return to schooling, reviewed here in section 3, the treatment variable is now DPEP program assignment rather than schooling attainment. The remaining terms in (3), (4), and (5), such as $\ell_{sy,D=1}$, are computed as subsample means of the relevant variables. The resulting estimate of $\beta_{y,D=0}$ and $\beta_{y,D=1}$ are labeled $\tilde{\beta}_0$ and $\tilde{\beta}_1$.

This strategy for estimating returns to skill and GE effects is useful. It also carries several limitations, in principle and in the K23 implementation. One is that most estimates of terms on the right of (3), (4), and (5) contain presumptively endogenous variation. A direct computation of the skill share $\ell_{sy,D=1}$ (the fraction of young workers in program district who are skilled) depends on DPEP program status (D) and on schooling via the skill dummy s. Both are taken in K23 as endogenous to the wage outcome. Nor, in the present framework, can a difference such as $\Delta_D \overline{\log w_{sy}}$ be estimated consistently with FRDD, for the sample in such regressions is restricted to endogenously defined subgroups such as skilled workers.[22]

Moreover, the results from the roundabout path to $\beta_{y,D=0}$ and $\beta_{y,D=1}$ are intrinsically less stable than the direct paths implied by the labeling in (1) and (2). The right sides of (3) and (4) are ratios whose denominators are estimated with uncertainty. By the same token, each Δ_D term, as estimated by FRDD, is also prone to large variance: FRDD, as an instance of exactly identified

---

[22] Technically the same argument applies to the splitting of the sample by age in K23 (section V.B) and section 3 above. Districts' population age structure may have been systematically related to their propensity to self-select into DPEP. But age is, if not strictly exogenous, at least predetermined. So this concern seems second-order.

instrumental variables estimation has no first or second moments (Kinal 1980).

B.2 Return to schooling

In order to express the skill premia per year of schooling, K23 divides them by an estimate of the average schooling gap between skilled and unskilled workers:

$$
\begin{aligned}
\tilde{\beta}_{y,D=0} &= \frac{\beta_{y,D=0}}{\bar{\bar{S}}_{ys,D=0} - \bar{\bar{S}}_{yu,D=0}} \\
\tilde{\beta}_{y,D=1} &= \frac{\beta_{y,D=1}}{\bar{\bar{S}}_{ys,D=0} - \bar{\bar{S}}_{yu,D=0}} \\
\Delta_D \tilde{\beta}_y &= \frac{\Delta_D \beta_y}{\bar{\bar{S}}_{ys,D=0} - \bar{\bar{S}}_{yu,D=0}}
\end{aligned}
\tag{6}
$$

The denominators are the same, a schooling gap in the untreated sample. They too are presumptively endogenous, because of their dependence on s and u.

B.3 Elasticities of substitution

K23 links the empirics to parameters in a hierarchical model of district-level production. The top level of the model is Cobb-Douglas in a district's labor supply (L_d) and capital (K_d), with constant returns to scale. In turn, L_d is a constant-elasticity-of-substitution (CES) function of the labor supply in various skill cells (L_{sd}), which are themselves CES functions of the labor supply in age-skill cells (ℓ_{asd}).

If each input is paid its marginal product, and if there are two skill groups, indexed by s and u, then the skill premium between the two groups within age cohort a and district d is

$$
\log \frac{w_{asd}}{w_{aud}} = \log \frac{\theta_{sd}}{\theta_{ud}} + \left(\frac{1}{\sigma_A} - \frac{1}{\sigma_E} \right) \log \frac{L_{sd}}{L_{ud}} - \frac{1}{\sigma_A} \log \frac{\ell_{asd}}{\ell_{aud}}
$$

where σ_E and σ_A are the elasticities of substitution in the skill and age levels of the hierarchical model and the θ_{sd} are the productivity coefficient for L_{sd} in the skill level. If we replace the d indexes with $D = 0$ and $D = 1$, the difference in the skill premium across the treatment/no-treatment split is

$$\Delta_D \beta_a := \log \frac{w_{as,D=1}}{w_{au,D=1}} - \log \frac{w_{as,D=0}}{w_{au,D=0}}$$

$$= \underbrace{\Delta_D \log \frac{\theta_{sd}}{\theta_{ud}} + \left(\frac{1}{\sigma_A} - \frac{1}{\sigma_E}\right) \Delta_D \log \frac{L_{sd}}{L_{ud}}}_{\text{independent of age}} - \underbrace{\frac{1}{\sigma_A} \Delta_D \log \frac{\ell_{asd}}{\ell_{aud}}}_{\text{age-specific}} \qquad (7)$$

As noted beneath the equation, only the last term depends on age. K23 appears to assume that it is zero for older workers, i.e., when $a = o$. This amounts to assuming that DPEP did not indirectly affect the schooling levels among the workers too old to have attended DPEP-supported schools (say, via migration or entry into or exit from the workforce):

$$\Delta_D \log \frac{\ell_{osd}}{\ell_{oud}} = 0 \Rightarrow \frac{\ell_{os,D=1}}{\ell_{ou,D=1}} = \frac{\ell_{os,D=0}}{\ell_{ou,D=0}} \qquad (8)$$

Plugging in $a = y$ and $a = o$ in (7) then gives

$$\Delta_D \beta_y - \Delta_D \beta_o = -\frac{1}{\sigma_A} \Delta_D \log \frac{\ell_{ysd}}{\ell_{yud}} \qquad (9)$$

$$\Delta_D \beta_o = \Delta_D \log \frac{\theta_{sd}}{\theta_{ud}} + \left(\frac{1}{\sigma_A} - \frac{1}{\sigma_E}\right) \Delta_D \log \frac{L_{sd}}{L_{ud}} \qquad (10)$$

All quantities in these equations other than the elasticities σ_A and σ_E are estimated as described after (5). The equations are then solved for the elasticities. K23 estimates $\sigma_A = 5$ and $\sigma_E = 4.24$, the latter using the rough assumption that $\log(\theta_{sd}/\theta_{ud}) = 0$. K23's public data and archive does not precisely document these calculations.

B.4 A revised method for estimating GE effects and elasticities

In light of the conceptual issues described above, and in response to certain implementation details embedded in the K23 code, I propose a revised method for estimating the GE effects and elasticities. The revisions:

1. Those over 75 are excluded from the sample, as in the rest of K23.

2. Those with more than 14 years of schooling are retained, as in the rest of K23.

3. Non-wage-earners are excluded when computing schooling aggregates, as in

the rest of K23.

4. Where K23 computes some aggregates as medians, others as means, the revision only takes means, which seems most consistent with the theory.

5. In light of the disadvantages of the indirect route to estimating the skill premia, discussed in section B.1, $\beta_{y,D=0}$ and $\beta_{y,D=1}$ are estimated directly, via the labeled expressions in (1) and (2).

6. K23 estimates $\bar{S}_{ys}$ and $\bar{S}_{yu}$, average schooling among the young skilled and young unskilled, in a surprising and probably erroneous subsample: those living in districts that did not receive DPEP treatment despite scoring more than 10 percentage points below the qualifying threshold of 39.29%.[23] This sample is unrepresentative of districts close to the threshold, which are the locus of the FRDD estimates. The revision estimates such aggregates by the following procedure, taking $\bar{S}_{ys}$ as an example: First $\Delta\bar{S}_{ys}$ is estimated by FRDD. Then $\bar{S}_{ys}$ is computed as the average schooling attainment within the FRDD estimation sample, weighting by the FRDD kernel. Then the averages in DPEP and non-DPEP districts are estimated as $\bar{S}_{ys,D=0/1} = \bar{S}_{ys} \mp \Delta\bar{S}_{ys}/2$.

7. As shown in equation (6), in order to estimate returns for schooling, K23 divides the treatment skill premium in treatment districts by a schooling difference in non-treatment districts. The revision instead computes $\tilde{\beta}_1 = \beta_1/(\bar{S}_{ys,D=1} - \bar{S}_{yu,D=1})$.

8. The revision drops assumption (8). In particular, successively plugging $a = y$ and $a = o$ into (7) and subtracting the second version from the first gives

$$\Delta_D \beta_y - \Delta_D \beta_o = -\frac{1}{\sigma_A} \left(\Delta_D \log \frac{\ell_{ysd}}{\ell_{yud}} - \Delta_D \log \frac{\ell_{osd}}{\ell_{oud}} \right)$$

The revision estimates σ_A by solving for it in this equation rather than in (9).

[23] Lines 80–92 of "GE_welfare_bootstrap_part1.do" contain the restriction "female_literacy>0.1 & dpep==0". At this point in the execution, the female literacy variable has been subtracted from 39.268%.