David Roodman
Open Philanthropy
September 18, 2023

I deeply appreciate the care and effort that Gaurav Khanna, the anonymous referees, and the editor have invested in reviewing the comment I submitted to *JPE*. The intense engagement, especially with the original author, has deepened my understanding of the evidence. Here, I summarize what I learned from the reports, make nine major points in response to them, and group minor points of agreement and disagreement under an 11th heading. The comments apply almost exclusively to the Khanna report, which I will sometimes refer to simply as "the report."

The Khanna report devotes the most space to generalizing the R23 finding that the main K23 results depend on the unexplained absence of four districts. It demonstrates other ways to delete observations near the threshold in order to obtain results like those in K23. A key message in response, laid out below, is that those modifications are ad hoc and not econometrically justified, and only generate consistently significant results by ignoring major R23 comments on the inference in K23. Nevertheless, Khanna's challenge led me to think more carefully about sensitivity to "inliers," and led to a new section in the revised comment on robustness testing.

1 What I learned and what I changed

As I will discuss in section 11.1, the Khanna report usefully challenges R23's detection and handling of multiparent districts. In my revision, I now exclude three more such districts, retaining just five because the mixing of their parentage is minimal. This change hardly affects regression estimates. It does modestly affect the preliminary discontinuity plots in that switching to the new data set does not shrink the apparent discontinuity in years of schooling as much as in my original submission. As a result, the discussion of the plots no longer emphasizes so exclusively the role of default overrides in generating the impression of discontinuities. It says, rather, that relinquishing those overrides *and* moving to the improved data erases the impression.

In addition, in dropping K23's binning choices, the revision has switched from rdplot's default, "mimicking-variance" algorithm to the alternative CCT algorithm, which seeks to match the first moment of the regression function rather than the second. This algorithm appears to better match the spirit of K23's manual choices.

More important, the Khanna report's demonstration that there are many ways to selectively delete observations near the threshold in order to produce results comparable to those in K23 led me to think more carefully about sources and consequences of misclassification of districts closest to the literacy threshold. The true threshold used to assign treatment may have been higher than the exact national average assumed in K23 and R23. If so, then many districts near the threshold may be effectively misclassified, generating substantial noise. In my revision, a new section generates evidence on the actual placement of the program threshold and tests systematically for robustness to varying the threshold used in the econometrics. It also check robustness to varying the bandwidth, as in the Khanna report. The testing demonstrates the robustness of the R23 finding that, with complete data and proper inference, impact estimates are indistinguishable from zero.

2 The Khanna report does not provide a specific, plausible explanation for why the four districts just above the cutoff are missing.

The Khanna report implies that the omission occurred when merging the main data set with data sets used mostly in the K23 appendix. "K23 merges many more datasets than R23….The analysis in K23 includes various additional outcomes (firm-level outcomes, school-level outcomes, test scores, district domestic product, etc.), many of which do not cover all districts, but are necessary for the comprehensive analysis." But merging more data sets need not and ought not cause deletion of data from the main data set. For example, Stata's merge command performs outer joins by default. Moreover, it is hard to see how merging, say, a data set on school-level outcomes would only cause deletion of a handful districts that were dispersed geographically yet extremely close to each other—and to the cutoff—in their female literacy rates.

3 Khanna remains in violation of two provisions of the *JPE* data policy.

One way to end the mystery of the four missing districts would be for Khanna to share the K23 intermediate data and code. *JPE* policy states that "Authors are invited to submit… intermediate data files and programs as an option; if they are not provided, authors must fully cooperate with investigators seeking to conduct a replication who request them." Pursuant to this policy, I requested the intermediate data and code in an email to Gaurav Khanna on March 29, 2023. I have received no response.

Separately, code for the "unbiased estimates of the model's elasticities" cited in the K23 abstract is missing from the public code and data archive posted by *JPE*. This omission also goes against *JPE* policy.

4 The report largely ignores three major comments in R23 on the main K23 RDD results, which together substantially affect inference.

The report's introduction states, "I respond to each point [in R23] in detail." But it barely addresses these comments:

1. The K23 regressions combine mean-squared error (MSE)-optimal bandwidths and conventional standard errors, a practice that the authoritative guidelines of Cattaneo, Idrobo, and Titiniuk (CIT, 2019, p. 59) call "not only invalid but also methodologically incoherent." Reliable frequentist inference requires, for example, the robust, bias-corrected variance estimator of Calonico, Cattaneo, and Titiniuk (CCT, 2014). The report does not engage with this issue and continues to report "methodologically incoherent" results.
2. CIT (p. 40) states that "Since [RDD] empirical results are often sensitive to the choice of bandwidth, it is important to select [the bandwidth] in a data-driven, automatic way to avoid specification searching and ad hoc decisions." The rdrobust command does so by default, applying the CCT algorithm for estimating the MSE-minimizing bandwidth. Yet K23 overrides the defaults in most regressions by transferring the MSE-optimized bandwidth from one regression—for years of schooling among the young—to all other combinations of sample and specification (young and old; schooling, log wages, return-to-schooling). Since the MSE-optimized bandwidths of all the other combinations are smaller (R23, Table 3), the widening of samples should increase endogeneity bias even as it increases apparent precision and significance through smaller (conventional) standard errors.
3. The K23 regressions do not cluster by district even though a) treatment is assigned by 1991 district and b) 2009 districts are the sampling unit in one level of the complex survey design in the follow-up. Abadie et al. (2022) explains why either trait alone motivates clustering. The theory applies to RDD as an instance of weighted ordinary least squares. This is why R23 clusters by the running variable, female literacy: it is exactly equivalent to clustering by 1991 district, since each district's literacy rate is unique.

   In a footnote, the Khanna report discusses clustering, but in a way that is ambiguous and, to the extent opposed, inapt. The report quotes *Causal Inference* in seeming opposition to R23's clustering on female literacy: "whatever you do, don't cluster on the running variable." However, the underlying authority for this strong directive doesn't apply; it is Kolesár and Rothe (2018), "Inference in Regression Discontinuity Designs with a Discrete Running Variable." Female literacy is continuous in theory and practice, for it takes some 200 values within a typical K23 bandwidth of 0.1. Contrast that density of distribution with the sparsity in Kolesár and Rothe's example, in which the running variable is age in years. Moreover, the degeneracy that Kolesár and Rothe identifies manifests as clustered errors being *smaller* than non-clustered, heteroskedasticity-robust ones. In R23, clustering doubles or triples standard errors.

In another sentence, the report seems to endorse clustering, if in a confusing way: "I also recommend nnclustering, rather than clustering at the running variable (as R23 does) given the mass points." That apposition creates a non sequitur. Whether to plug in nearest-neighbor residuals rather than own residuals is separate from what variable to cluster on. The report therefore endorses a form of clustering without specifying what to cluster on. As a practical, matter "nnclustering" does not work in the K23 regressions because the nearest neighbors of each observation are from the same district, with exactly the same value on the running variable. This leads to degeneracy. rdrobust crashes if one accepts its default of 3 for the number of neighbors. As one raises that number, standard errors fall monotonically, so that one can practically achieve any significance level.

5 The report's argument against incorporating survey weights is based on a misunderstanding.

The K23 regressions do not incorporate sampling weights even though the sampling is presumptively endogenous to the outcomes of interest. R23 runs regressions with and without weights because a few observations carry extreme weights, which poses a sharp bias-variance trade-off. K23 argues that the "weights [option in Calonico, Cattaneo, and Titiniuk's 'rdrobust' command] is meant for RD estimation (as when moving from the triangular to the uniform kernel we weight observations differently)" and not for incorporating survey weights. Sebastian Calonico confirmed with me that this characterization of the intention of the rdrobust authors is incorrect.

6 The report's new demonstrations of K23's sensitivity to sample are ad hoc and lack econometric justification.

The Khanna report double's down on R23's finding that matching the K23 results requires deleting four districts. It shows that one can generate similar estimates by deleting or deemphasizing districts near the cutoff according to various combinations of several rules:

1. Dropping only the closest district to the cutoff.
2. Dropping all districts within 0.4 percentage points of the cutoff ("donut RDD").

3. Dropping single-parent split districts, i.e., 2009 districts whose territory lay within a single 1991 district.
4. Dropping Cuddalore, the one single-parent split district among the missing four.
5. Switching to the uniform kernel while retaining a bandwidth optimized for the triangular.

All of these tweaks lack econometric justification other than ad hoc specification exploration.

1. *Donut RDD is not warranted since the running variable was not manipulated.* As a general matter, one might drop *outliers* from a regression on the idea that extreme observations are most likely to flow from a data generating process distinct from that of the main mass of data. That rationale does not transfer to RDD "inliers" for they *are* the main mass of data, the very observations given the most prominence because of the local exogeneity of intention to treat. Now, if the running variable were manipulated near the cutoff, that would make it locally endogenous and could justify deleting data in the center of a donut (Almond and Doyle 2011; Barreca, Lindo, and Waddell 2016; Cattaneo, Idrobo, and Titiniuk 2019). But in K23, the running variable is a census statistic collected without foreknowledge of DPEP, nor even of what the national average—the cutoff—would be. It could not have been manipulated. The McCrary test in K23 (Figure 1A) reassures as to lack of manipulation. Meanwhile, donut RDD is more biased than is commonly understood and less precise than is conveyed by the usual standard error estimates (Noack and Rothe 2023).

   As an aside, if donut RDD is to be performed, it would be best, in order to ward off *p*-hacking concerns, to explain the choice of inner radius (0.4 points in the Khanna report) and check for robustness to this choice. The Khanna report does neither.

2. *Single-parent split districts do not threaten the consistency of the K23 specifications.* Fully 255 of the 570 districts in the K23 data set are single-parent, split districts. R23 includes a similar number. The Khanna report raises doubts about this practice: "One concern with the R23 sample is that it may be introducing measurement error in treatment assignment right at the cutoff….This noisy assignment could result from various factors, including district splits/merges over time." Referee #3 amplifies this concern by developing a scenario in which districts split during the interval between the fielding of the 1991 census, which set the running variable, and the selection of DPEP districts in the mid- and late-1990s. A district might have split into low- and high-literacy offspring, and the high-literacy descendant might not have received treatment, despite being classed by the low parental literacy as intended-to-treat in

K23. This misclassification would not threaten consistency, for it would not undermine the local exogeneity of intention to treat. But it would add noise. In fact, the scenario is academic: there are *no* cases of two children of a shared, single parent receiving different treatment.

The Khanna report quotes me accurately: "I know there's a lot of complexity because districts split and merge over time. But it looks to me like these four weren't affected by that." I stand by that informal statement in the precise sense that adding the missing four districts, including that one that is a split district, does not undermine consistency.

3. *Though described in the report as "more conventional," the uniform kernel is less preferred than the triangular used in K23 and R23.* The triangular kernel is the default in the dominant RDD package, rdrobust, "because when used in conjunction with a bandwidth that optimizes the mean squared error (MSE), it leads to a point estimator with optimal properties." (CIT, p. 37). An older authority, Imbens and Lemieux (2007, p. 625), supports the uniform kernel on practical grounds *if* backed by checks for robustness to this choice.

   What appears less justified is setting the bandwidth for the uniform kernel to the MSE-optimized bandwidth for the *triangular* kernel.

All that said, as noted in item 1 above, by showing that there are many ways to deemphasize "inliers" in order to generate K23-like results, the Khanna report has usefully demonstrated that the influence of the missing four is part of a larger pattern. My revised comment devotes a section to exploring sensitivity in way that is better motivated and less ad hoc.

7 The report's new demonstrations of K23's sensitivity to sample lose most statistical significance with proper inference.

As noted, the report does not challenge three important comments in R23 pertaining to inference. Nor does it take those comments on board. The new results seeming to show the robustness of the K23 result generally borrow bandwidths optimized for other specifications and report non-clustered, "methodologically incoherent" conventional standard errors. Table 1 below shows the results from incorporating the three comments into the report's Tables 1 and 2. The new table also extends the testing to reduced-form log wage regressions and to fuzzy RDD return-to-schooling estimation. One of the 18 estimates differs from 0 at $p < .1$. None differs from 0 at $p < .05$.

Table 1. Khanna report robustness tests with proper inferential procedures

			Drop districts	Drop only split	Drop only split	
		Drop district	in donut hole,	districts in	districts in donut hole	Full sample,
	Full sample	nearest cutoff	radius = .004	donut hole	+ Cuddalore	uniform kernel
Schooling	0.240	0.334	1.077*	0.352	0.352	0.322
	(0.505)	(0.535)	(0.550)	(0.498)	(0.498)	(0.548)
Bandwidth	.112	.103	.089	.115	.115	.081
Observations	12,332	10,410	8,979	12,478	12,478	8,460
Log wages	–0.0123	0.0314	0.143	0.0127	0.0127	0.102
	(0.125)	(0.134)	(0.147)	(0.124)	(0.124)	(0.119)
Bandwidth	.074	.08	.086	.075	.075	.084
Observations	7,981	8,382	8,730	8,015	8,015	8,869
Return to	0.316	0.240	0.118	0.217	0.217	0.291
schooling	(0.589)	(0.432)	(0.147)	(0.297)	(0.297)	(0.415)
Bandwidth	.099	.097	.083	.116	.116	.095
Observations	10,344	9,736	8,446	12,610	12,610	9,713

Notes: Robust, bias-corrected, district-clustered point estimates and standard errors are reported. Bandwidths are MSE-optimized (CCT 2014). Samples restricted to wage earners aged 17–34. *$p < .1$. **$p < .05$. ***$p < .01$.

8 The report does not engage with the R23 comments on the econometrics of estimating general equilibrium effects.

R23 demonstrates that the K23 estimates of elasticities and GE effects, contrary to the "unbiased" billing in the abstract, presumptively contain endogeneity bias. They are also effectively unidentified because of infinite first moments. The report responds by emphasizing that in complex systems such as labor markets, nearly all variables are endogenous in the system dynamics sense: each effectively influences all the rest. But K23 claims to measure parameters governing the Indian labor market free of endogeneity in the *econometric* sense. That is the locus of the challenge in R23. For example, while differences in outcomes such as wages are computed across the locally exogenous treatment/non-treatment divide, the groups whose averages are being compared are endogenously defined. The report does not address these comments.

9 Though renamed as "difference-in-discontinuities" the Khanna difference-in-differences regressions still exploit no discontinuities in the sense of locally exogenous jumps in treatment.

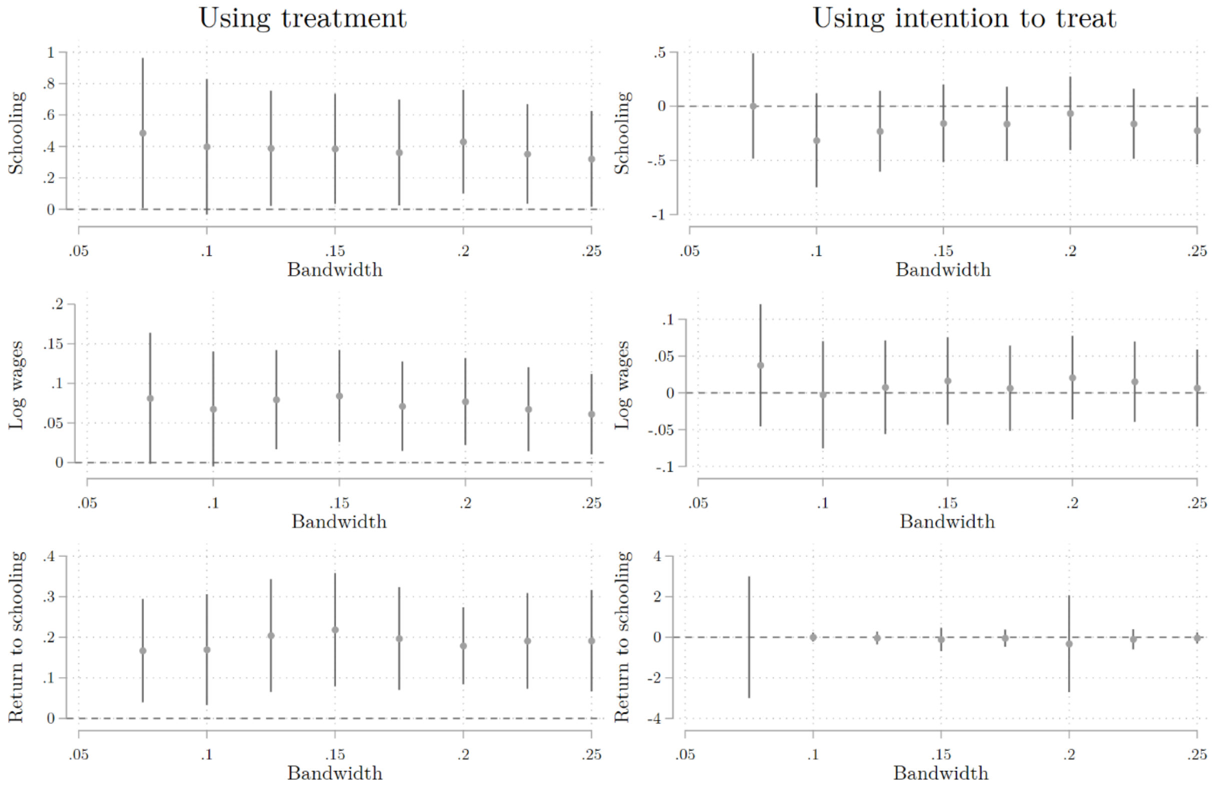
The Khanna report demonstrates the robustness of a "difference-in-discontinuities" estimate of the impact on schooling to changing the bandwidth. "Difference-in-discontinuities" appears to be a misnomer, however, because the estimator exploits no discontinuities as that term is meant in this context. The sample is split by age, at 35 which is not associated with a discontinuity in treatment.[1] And, like K23, it does not exploit the discontinuity in intention to treat, instead comparing actually treated to actually untreated districts. One might view treatment as discontinuous since it is binary; but what matters is that the discontinuity does not generate locally exogenous identifying variation. In contrast, Grembi, Nannicini, and Troiano (2016)'s originating example of difference-in-discontinuities exploits discontinuities in fiscal policy with respect to both time and the population of a municipality. K23's label, "difference-in-differences" (DID) remains more appropriate.

To make this point concrete, I copy the Khanna report in plotting the sensitivity of DID results to bandwidth. The upper left pane of Figure 1, below, approximately replicates Figure 3 in the Khanna report. It confirms that when taking DID with respect to actual treatment (and when clustering by district, as in K23), the association of treatment with schooling is robustly positive. In the remainder of the left column, the same is done for the association with wages, and for the return to schooling, by performing Wald DID as in Duflo (2001).

The computations for the right column are the same except in substituting intention to treat for treatment. This modification quite changes the results: impacts of intention to treat on schooling and wages are essentially indistinguishable from zero at all bandwidths. The disagreement between the treatment and intention-to-treat results should reduce confidence in any causal interpretation of the former, no matter how robust. That is why R23 sets DID aside.

[1] After a big primary schooling expansion in Indonesia, Duflo (2001) looks (informally) for a kink in outcomes with respect to age, but not a discontinuity.

Figure 1. Difference-in-differences estimates



10 The report provides no precise explanation for K23's overriding of the default binning algorithm in the discontinuity plots but rightly points out the methodological complexity in the choice of polynomial order.

The introduction of the report promises to "show that K23's modifications follow clear and sensible guidelines in the RD literature" in generating the discontinuity plots—i.e., in choosing the order of the global polynomial fits on each side of the discontinuity and the number of bins into which to average the data. As for the polynomial order, the report cites Gelman and Imbens (2018), which is emphatic that "high-order global polynomial approximations should not be used, and that instead, inference based on local low-order polynomials (local linear or local quadratic) is to be preferred." In fact the guidelines in the RD literature are not so clear. Andrew Gelman has excoriated specific discontinuity plots of many orders (linear, quadratic, cubic, quartic). CCT and CIT (p. 18) recommend 4th- or 5th-order polynomials for global fits.

A sensible synthesis is that:

1. Formal inference should proceed from low-order, local polynomial fits, as in K23 and R23 (both use linear fits, the rdrobust default).

2. When graphing, whatever the polynomial order used for global fits, researchers should a) check whether any appearance of a discontinuity is robust to the choice of polynomial order and b) check whether the appearance of a discontinuity is supported by the overlaid, binned scatter plot.

In particular, the primary use of the global fits should be to indicate the overall shape of the regression function, *not* to check for a discontinuity. The scatter plots should carry more of the informal inferential weight with regard to the discontinuity. Unfortunately, it is hard for a reader not to focus on the break at the threshold between the global fits. And I believe my submission overemphasized this aspect of the plots. Nevertheless, R23's point that the appearance of a discontinuity depends in part on overriding a default retains some force, under item 2 above.

Separately, the report provides no precise description or justification for how K23 picks the numbers of bins for each side of the binned scatter plots. The report only quotes the guidance of CIT: "Bins can be chosen in many different ways. Which method of implementation is most appropriate depends on the researcher's particular goal." The report omits the next sentence from CIT, which goes directly against the choice in K23 and corroborates that in R23: "We recommend to start with [mimicking-variance] bins to better illustrate the variability of the outcome as a function of the score." R23 adopts that method simply by accepting rdplot defaults.

11 Minor points

11.1 Minor points of agreement

Formation of multi-parent districts during 2001–09

R23 states that "no new multi-parent districts were formed in 2001–09." That is wrong, as the Khanna report (note 6) points out. Of the report's seven counterexamples, Samba and Ramban are irrelevant because they were not covered by the 1991 census and so lack the running variable; and Tiruppur and Pratapgarh (in Rajasthan) are irrelevant because they were evidently formed late enough that they do not appear in the follow-up surveys data structure. But the report is right, and relevant, about Mohali, Baksa, and Chirang (and Wikipedia is wrong about Chirang!). In my revision I exclude all three under the criterion discussed next.

An "arbitrary" threshold for inclusion of multiple-parent districts should be lowered

Formation of new districts from fragments of multiple old ones poses a measurement problem. How do we impute intent-to-treat and treatment status to the resident of a 2009 district when we don't know which former district the person lived in?

One reasonable response is to drop all observations from multiple-parent districts. But this extreme arguably makes the perfect the enemy of the good. Better, I think, to include districts with *de minimus* mixed parentage in order to increase statistical power at miniscule cost in bias and noise. Consider this example: Between 1991 and 2001, Sultanpur, Uttar Pradesh, received a sliver of territory from Rae Bareli, which technically made Sultanpur multi-parent but increased its population by only 0.5% (Kumar and Somanathan 2016, p. 61). The two parents had similar female literacy rates: 20.8% in Sultanpur, 23.0% in Rae Bareli. Both received treatment (DPEP funding). There is thus almost no ambiguity about the value of the running variable for the average Sultanpur resident.

To define *de minimus*, I computed, for each child district, the standard deviation of the female literacy rate of its parents, weighting by the parents' population contributions to the child. This statistic is low where parents have similar female literacy rates or when one parent contributes nearly all the population—the two extremes in which including a multi-parent district introduces the least measurement error. After examining cases, a 3% ceiling looked like a reasonable operationalization of *de minimus*. R23, note 4, documents this criterion but does not explain it.

The report usefully challenges the 3% threshold. It led me to see that there exist two multi-parent districts whose parental female literacy rates are nearly identical—yet which land on opposite sides of the identifying threshold. That does seem problematically ambiguous. One of the two is Champawat, Uttarakhand, which the report mentions. The other is Moga, Punjab.

I have lowered the multi-parent *de minimus* threshold from 3% to 1%. This excludes those two and moves to a less arbitrary-seeming number. All results reported here and in the revision reflect this change.

11.2 Minor points of disagreement

1. The Khanna report characterizes R23 as "finding no errors in the original analysis code." In fact, R23 points out that four influential districts are missing; that the standard errors are not "valid" in a sense defined by the creators of robust RDD; that the wage variable includes wages only from one activity even for people reporting several; and that a step in the GE calculations uses "a surprising and probably erroneous subsample." It also documents a math error in the appendix: the average of logs is not the log of the average.
2. "R23 then recreates part of the estimation sample." R23 creates all of the estimation sample for the main results, and indeed adds four districts.
3. "R23 uses the revised 2001 list of districts for program assignment rather than the original (and more exogenous) list." This assertion is not backed by evidence, but should be.

Despite a request for intermediate data and code sent on March 29, 2023, Khanna has yet to document the source for program assignment in K23.

4. CCT “make clear that when clustering, we must re-estimate the bandwidths.” One can argue with equal force that when changing the dependent variable, or sample, or estimator (from RDD to FRDD), one “must” re-estimate the bandwidths. But K23 and the Khanna report do not do so, so I do not see a way to define “must” in the quoted sentence that is compatible with the choices in K23 and the Khanna report.
5. In the upper-right of Figure 1, “R23 plots the unbinned raw data at the individual level such that there is wide variation.” Actually, like K23, R23 uses the rdplot command to bin the data for the scatter plots. It differs only in accepting the command’s default.
6. “Researchers likely use different assumptions to handle district splits and merges. R23’s sample construction was not coded in statistical software, so I could not evaluate it in the timeframe provided.” I sent Khanna the district crosswalks in an Excel file by email in April 2022, almost a year before submitting to *JPE*. The same file is in the Github repository linked to from R23.
7. “R23’s data had deleted certain age groups, so I could not re-evaluate K23’s original diff-in-diff.” The supplied “K23 individual-level.dta” file contains cases of every age between 17 and 100.

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): 268–93. https://doi.org/10.1111/ecin.12225.

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. http://jstor.org/stable/25098683.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. “Star Wars: The

Empirics Strike Back." *American Economic Journal. Applied Economics* 8 (1): 1–32. http://10.1257/app.20150044.

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

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.

Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920–80. https://doi.org/10.1257/jel.20171350.

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

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.

Gerber, Alan S., and Neil Malhotra. 2008. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods & Research* 37 (1): 3–30. https://doi.org/10.1177/0049124108318973.

Grembi, Veronica, Tommaso Nannicini, and Ugo Troiano. 2016. "Do Fiscal Rules Matter?" *American Economic Journal. Applied Economics* 8 (3): 1–30. https://doi.org/10.1257/app.20150076.

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." https://cdedse.org/pdf/work248.pdf.

Noack, Cladia, and Chistoph Rothe. 2023. "Donut Regression Discontinuity Designs." *arXiv [econ.EM]*. arXiv. http://arxiv.org/abs/2308.14464.

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.