Paper title: Randomized Controlled Trials and Regression Discontinuity Design Extrapolations: An Empirical Comparison
Authors: Felipe Barrera-Osorio, Deon Filmer, Joe McIntyre, Luke Miratrix, Nozomi Nakajima
Contact information: nnakajima@g.harvard.edu

Background/Context:
In regression discontinuity (RD) design, causal inference is drawn by comparing units immediately above and below some cutoff. As a result, inference is limited to the local average treatment effect (LATE). Recent studies propose leveraging multiple cutoffs to extrapolate or generalize the LATE estimated from RD (Cattaneo et al. 2016; Bertanha 2017; Cattaneo et al. 2019). For example, Cattaneo et al. (2016) show that the treatment effect parameter estimated by normalizing the running variable and pooling all observations together is a weighted average across cutoffs of LATEs for all units facing each particular cutoff. Our paper contributes to the RD extrapolation literature by conducting an *empirical comparison* of the average effects estimated from a multi-cutoff RD with a benchmark treatment effect estimated from a randomized controlled trial (RCT) within the same study.

We conduct a within-study comparison (WSC), to assess the empirical correspondence between estimates derived from non-experimental methods and experimental methods (Cook, Shadish & Wong, 2008). Prior research using WSC have examined how much the RD estimate at the cutoff differs from the RCT estimate at this same cutoff (Chaplin et al. 2018). Our paper differs from prior work by focusing on how much the extrapolated RD estimate differs from the *overall* RCT estimate within the same study. We also introduce a bootstrapping approach to formally test for disagreement between the estimators.

Purpose/Objective/Research Question:
The aim of this study is to better understand RD extrapolation by comparing the average effect estimated from multiple RD cutoffs against the average effect estimated from an RCT. Units are clustered in groups and each group faces a different RD cutoff.

We seek to answer two research questions:
1. What is the correspondence between the average effect estimated from a multi-cutoff RD and the average effect estimated from an RCT?
2. To what extent does within- and between-group variation explain the difference between the average effect estimated from a multi-cutoff RD and the average effect estimated from an RCT?

Setting & Population:
We use data on two scholarship programs in Cambodia.[1] These scholarship programs were evaluated in 204 schools, where each treated school had its own local discontinuity in scholarship award.

Intervention/Program/Practice:
The aim of the scholarship programs was to provide financial assistance to improve primary

[1] The RCT results have been published in Barrera-Osorio & Filmer (2016) and Barrera-Osorio, de Barros & Filmer (2018). Our paper uses these previously published RCT estimates as the benchmark for the average effects estimated from the multi-cutoff RD.

school progression.

Research Design:
As shown in Figure 1, the program was designed as an experiment in which schools were randomly assigned into a merit-based scholarship program, a poverty-based scholarship program, or a control group. Students in treatment schools who scored above the median merit or poverty score received the scholarship.

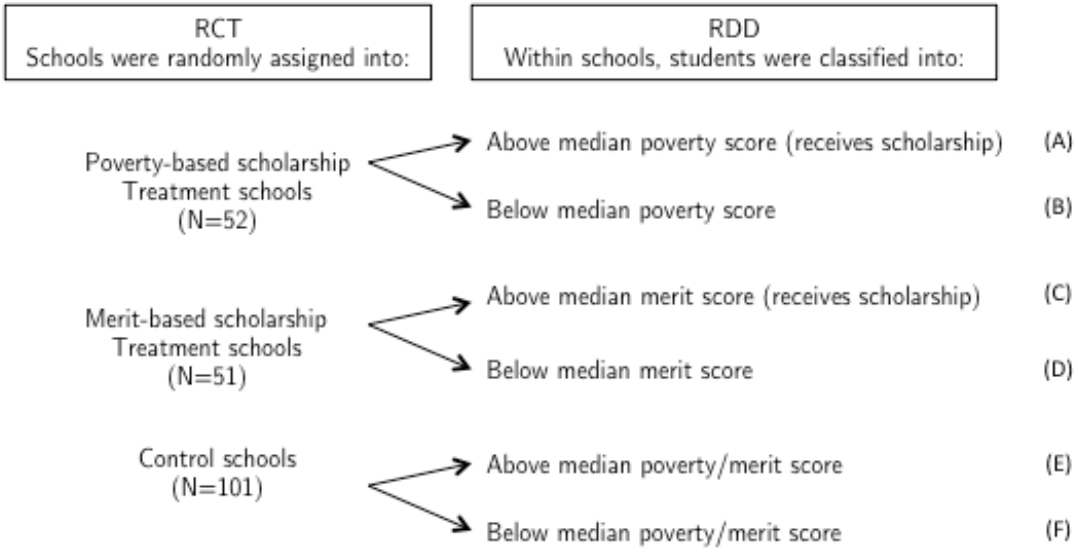

Figure 1. Research design

Analysis:
In the RCT analysis, the treatment effect is estimated by simply contrasting (A) and (E) or (C) and (E) in Figure 1. In the multi-cutoff RD analysis, the treatment effect is estimated by contrasting (A) and (B) or (C) and (D) in Figure 1.

Let T_j denote the school's treatment status, S_{ij} denote the running variable (either the poverty or merit score) for student *i* in cohort *j*, and C_j denote the cutoff that each student *i* in cohort *j* faces, which has support $\mathcal{C} = \{c_1, c_2, \ldots c_J\}$ with $P[C_j = c] = p_c \in [0,1]$. For a student in a treatment school, assignment to a scholarship depends on both the running variable S_{ij} and the cutoff C_j. Students receive a scholarship if the value of S_{ij} exceeds the value of the cutoff and attend a treatment school ($Z_{ij} = 1$ if $S_{ij} > C_j$ and $T_j = 1$). Students do not receive a scholarship if the value of S_{ij} is below the value of the cutoff ($Z_{ij} = 0$ if $S_{ij} \leq C_j$).

As shown in Figure 2, there is substantial variation in the cutoff along the running variable. We normalize the forcing variable for each student *i* in cohort *j* by taking $S_{ij} - C_j$, pool all the observations as if there was only one cutoff at $c = 0$, and use standard RD techniques. Formally, the pooled RD causal estimand is:

$$\tau_{RD} = \sum_j E\left[Y_{ij}(1) - Y_{ij}(0) | S_{ij} = C_j, j\right]\omega(j)$$

where $\omega(c)$ are the weights, which are measured as the number of students in each cohort *j* (i.e., the number of student facing the cutoff *c*).

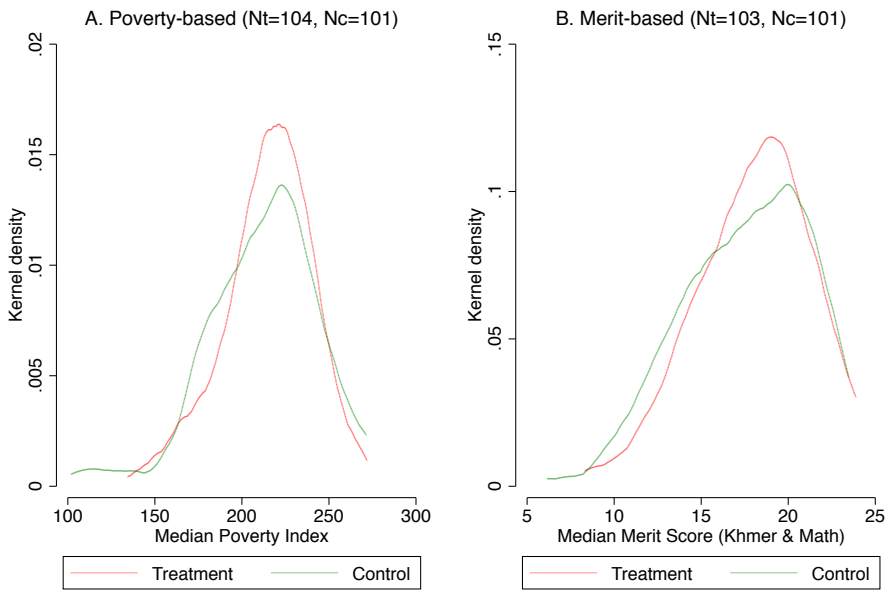

Figure 2. Kernel density of cutoff scores for multi-cutoff RD

Finally, we are interested in empirical comparisons of τ_{RCT} and τ_{RDD}, which we estimate as $\Delta = \tau_{RCT} - \tau_{RD}$. We use a bootstrapping approach to estimate the standard error for Δ.

Findings/Results:

The estimates for τ_{RCT}, τ_{RD} and Δ are shown in Tables 1 and 2 below. Overall, we find that τ_{RD} and τ_{RCT} are not significantly different from one another in the case of the poverty-based scholarship but τ_{RD} is significantly smaller than τ_{RCT} in the case of the merit-based scholarship.

There are two sources of variation in the treatment effects from the scholarship programs. One variation is the school's median score relative to other schools in the data. In other words, schools that have very low (or very high) median scores may have differential impacts from other schools. A second source of variation is the student's score relative to other students in their school. Since we focus on students close to the cutoff for estimating τ_{RD}, the estimator may underestimate the extent of the program's impact if there is considerable treatment effect variation based on distance away from the cutoff.

Preliminary results from analyzing the between-school and within-school variation suggest that these two sources of variation explain the divergence between τ_{RD} and τ_{RCT} for the merit-based scholarship. τ_{RCT} is larger among schools with relatively higher cutoffs and larger among students who have higher merit scores within their schools. In contrast, τ_{RD} is larger among schools with relatively lower cutoffs and larger among students with lower merit scores within their schools.

Conclusions:

In education, researchers and policymakers are increasingly interested in how to generalize treatment effects estimated in one context to another. This raises the question of whether the

treatment effect estimated from RD can be extrapolated to estimate average effects. The findings from this paper suggest that empirical researchers may want to caution against extrapolating from RD estimators to recover average treatment effects. In particular, while some variation (across schools) would be evaluable in a multi-RD design, other variation (different impacts for students within school) would not be so detectable.

Table 1. Results for merit-based scholarship

Outcomes	τ_{RCT}	τ_{RDD}	$\Delta = \tau_{RCT} - \tau_{RDD}$
Math (S.D.)	0.173	0.069	0.104
	(0.095)	(0.132)	(0.154)
Observations	940	985	
Control Group Mean	0.165	-0.138	
Digit Span (S.D.)	0.154*	0.034	0.120
	(0.076)	(0.146)	(0.153)
Observations	940	985	
Control Group Mean	0.0812	-0.0277	
Finished grade 6 (1=Yes)	0.124**	0.045	0.079
	(0.044)	(0.066)	(0.079)
Observations	940	985	
Control Group Mean	0.635	0.567	
Years of education completed	0.196	0.102	0.094
	(0.104)	(0.196)	(0.188)
Observations	897	928	
Control Mean	5.448	5.194	

*** p<0.001, ** p<0.01, * p<0.05
Robust standard errors clustered by schools in parentheses. Standard errors for Δ are bootstrap standard errors. MSE-optimal bandwidth used for RDD specification.

Table 2. Results for poverty-based scholarship

Outcomes	τ_{RCT}	τ_{RDD}	$\Delta = \tau_{RCT} - \tau_{RDD}$
Math (S.D.)	-0.034	-0.099	0.065
	(0.081)	(0.123)	(0.106)
Observations	883	940	
Control Group Mean	0.0180	-0.189	
Digit Span (S.D.)	-0.044	0.056	-0.100
	(0.075)	(0.114)	(0.120)
Observations	883	940	
Control Group Mean	0.0153	-0.0281	
Finished grade 6 (1=Yes)	0.192***	-0.097	0.289***
	(0.040)	(0.067)	(0.069)
Observations	883	940	
Control Group Mean	0.613	0.626	
Years of education completed	0.375**	-0.335	0.710***
	(0.118)	(0.183)	(0.250)
Observations	831	874	
Control Mean	5.377	5.374	

*** p<0.001, ** p<0.01, * p<0.05
Robust standard errors clustered by schools in parentheses. Standard errors for Δ are bootstrap standard errors. MSE-optimal bandwidth used for RDD specification.

References

Barrera-Osorio, F. and Filmer, D. (2016), "Incentivizing Schooling for Learning: Evidence on the Impact of Alternative Targeting Approaches", *Journal of Human Resources*, 51, 461-499.

Barrera-Osorio, F., De Barros, A., & Filmer, D. (2018). "Long-term Impacts of Alternative Approaches to Increase Schooling: Evidence from a Scholarship Program in Cambodia". *Working Paper*.

Bertanha, M. (2017). "Regression Discontinuity Design with Many Thresholds". *Working Paper*.

Cattaneo, M. D., Keele, L., Titiunik, R., & Vazquez-Bare, G. (2016), "Interpreting Regression Discontinuity Designs with Multiple Cutoffs", *Journal of Politics*, 78, 1229-1248.

Cattaneo, M. D., Keele, L., Titiunik, R., & Vazquez-Bare, G. (2019). "Extrapolating Treatment Effects in Multi-cutoff Regression Discontinuity Designs". *arXiv preprint arXiv:1808.04416*.

Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N., & Morris, R. E. (2018). "The Internal And External Validity Of The Regression Discontinuity Design: A Meta-Analysis Of 15 Within-Study Comparisons." *Journal of Policy Analysis and Management*, *37*(2), 403-429.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New findings from Within-study Comparisons." *Journal of Policy Analysis and Management*, *27*(4), 724-750.