Abstract Title Page

Not included in page count.

Title: Assessing the Generalizability of Estimates of Causal Effects From Regression Discontinuity Designs

Authors and Affiliations: Howard S. Bloom, MDRC and Kristin E. Porter, MDRC

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

In recent years, the regression discontinuity design (RDD) has gained widespread recognition as a quasi-experimental method that when used correctly, can produce internally valid estimates of causal effects of a treatment, a program or an intervention (hereafter referred to as treatment effects). In an RDD study, subjects or groups of subjects (e.g. students or schools) are rated according to a numeric index (a performance indicator, poverty measure, etc.) and treatment assignment is determined by whether one's rating falls above or below an exogenously defined cut-point value of the rating. RDDs have been used to estimate causal effects in a variety of contexts (e.g. for a list of more than 75 studies in the contexts of education, labor markets, political economy, health, crime and more see Lee & Lemieux, 2009), and research on their statistical properties has provided theoretical justification and empirical verification of their internal validity (e.g. Hahn, Todd, & Klaauw, 2001; Rubin, 1977; Imbens & Lemieux, 2008; Lee, 2008). This validity derives from the fact that an RDD has the equivalent of a randomized controlled trial (RCT) at its cut-point (e.g. Mosteller, 1990; Goldberger, 1972a, 1972b). However, the localized nature of this RCT has prompted researchers to question its generalizability, or external validity (e.g. Imbens & Lemieux, 2008).

While it is true that RDD estimates of treatment effects are identifiable only at the cut-point value of the rating, this does not necessarily mean that such estimates apply only to a homogeneous sub-population of subjects. In order for this to be true, treatment effects must vary widely across the target population of interest, and they must co-vary with observed ratings. If treatment effects are not highly correlated with observed ratings, the conditional distribution of individual treatment effects at the cut-point will be *similar* to the unconditional distribution of individual treatment effects across the range of rating values for the target population of interest. However little is known about the extent to which treatment effects vary and even less is known about the correlates of their variation. In the absence of such information, researchers may be averse to generalizing RDD findings.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

Our paper explores the conditions that limit the generalizability of RDD estimates and concludes that that in many cases, generalizability is much greater than often believed. The paper also presents an empirical approach for quantifying the generalizability of RDD findings so that more information can be brought to bear on this important issue.

Setting:

Description of the research location.
(May not be applicable for Methods submissions)
Not applicable.

Population / Participants / Subjects:

Description of the participants in the study: who, how many, key features, or characteristics.
(May not be applicable for Methods submissions)
Not applicable.

Intervention / Program / Practice:

Description of the intervention, program, or practice, including details of administration and duration.
(May not be applicable for Methods submissions)
Not applicable.

Significance / Novelty of study:

Description of what is missing in previous work and the contribution the study makes.

When assessing the generalizability of RDD findings, one does not need to consider all of the variation in treatment effects that might exist; only that part which is correlated with observed (i.e. measured) ratings. In other words, the concern for an RDD is that true average treatment effects differ substantially for different values of the observed rating. *This concern is only justified however, if the mix of subjects with a given value of the rating differs substantially from the mix of subjects with other values of the observed rating—in terms of factors that predict or moderate treatment effects.* Figure 1 indicates the relationships that must exist in order for this to be the case.

One relationship is that between true and observed values of an RDD rating (r^* and r). A subject's true rating (i.e. actual measure of ability, merit, disadvantage, etc.) is unobservable and contains no measurement error or "noise". Observed ratings reflect true ratings *plus* noise. It is widely known that an internally valid RDD *requires* noise in its rating – in order to produce an RCT at the cut-point. For subjects with the same *true* rating, it must be a matter of chance (due to noise) whether their *observed* rating falls above or below the RDD cut-point. Less widely known, but perhaps equally important, is Lee and Lemieux's (2009) insightful observation that noise in a rating is also critical for the external validity (generalizability) of RDD findings. As they note, random error in observed ratings produces heterogeneity of subjects at the cut-point. Such heterogeneity implies generalizability of impact findings. If a rating is pure noise, individuals assigned by the rating are assigned randomly, as in an RCT.

However, noise in observed ratings is only one of several factors that can produce heterogeneity of treatment effects for each observed rating value, which in turn increases the generalizability of RDD findings. Hence, the force of Lee and Lemieux's insight is even stronger than immediately evident. In fact, as illustrated in Figure 1, the relationship between r and E^* is the *product* of three intervening "linkages" – in addition to that between r and r^*, there is the linkage between r^* and a moderator (we focus on a theoretical ideal composite moderator, M^*), and the linkage between M^* and E^*. Weaknesses in each of these linkages can weaken the relationship between observed ratings and true effects. Furthermore, the weaknesses are compounding and accumulate rapidly. Consequently, a modest weakness in each linkage can produce a substantial cumulative weakness in the overall chain. This in turn increases the generalizability of RDD findings substantially.

Statistical, Measurement, or Econometric Model:

Description of the proposed new methods or novel applications of existing methods.

Focusing on RDDs in which individuals are rated and assigned to a treatment, our paper presents a theoretical basis for an approach for quantifying the generalizability of RDDs in terms of the relative heterogeneity of treatment-effect moderators. We note that the generalizability of RDD estimates of treatment effects depends on the heterogeneity of true treatment effects (E^*) at the RDD cut-point, relative to the heterogeneity of true treatment effects for the treatment's target population. This relative heterogeneity can be expressed as a "generalizability coefficient" (G_{E^*}). If we assume homoskedastic errors, (G_{E^*}) equals one minus the R-square for observed ratings and true treatment effects.

In practice however, it is not possible to observe individual treatment effects because a treatment effect for an individual is the difference between his potential outcomes with and without treatment, which cannot both be observed (Rubin, 1974). Thus without strong assumptions, it is not possible to identify a distribution of individual treatment effects. It is possible however, to identify the distribution of individual values for moderators of effects. Thus it is possible to compare the conditional cut-point distribution of a true moderator (M^*) with its unconditional distribution for a target population, which is our estimand of interest, G_{M^*}. Even an ideal composite of all moderators of a treatment's effect can only predict systematic variation in these effects; it cannot predict idiosyncratic variation. Consequently, G_{M^*} *understates* G_{E^*}.

Our paper focuses on counterfactual outcomes as a composite moderator of treatment effects. Counterfactual outcomes are the potential outcomes that sample members would experience in the absence of treatment. Counterfactual outcomes are a composite moderator because they reflect all factors that determine outcome *levels* in the absence of treatment and thus probably reflect most (if not all) factors that determine outcome levels in the presence of treatment.

Usefulness / Applicability of Method:

Demonstration of the usefulness of the proposed methods using hypothetical or real data.

Our approach can only provide approximate results. It is proposed based on the premise that researchers should use informed judgments about the generalizability of RDDs instead of defaulting to a professed state of ignorance. We recommend that researchers investigate generalizability for a planned or actual RDD study by estimating generalizability coefficients with (1) data from outside the RDD and/or (2) data from within the RDD. We present examples with both types of data.

Research Design:

Description of the research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).
(May not be applicable for Methods submissions)

Not applicable.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.
(May not be applicable for Methods submissions)

Using non-RDD data, we estimate generalizability coefficients based on relationships between academic pre-tests (observed ratings) and academic post-tests in the absence of a specific treatment (observed counterfactual outcomes). The analyses are based on data from five large urban school districts, covering several years for each district, for grades 3, 5, 8, and 10, using student scores on standardized tests of reading and math (from Bloom, Richburg-Hayes, & Black, 2007). We used the published R-square values for observed ratings and observed counterfactual outcomes (scores on pre-tests and post-tests taken a year apart, respectively) to approximate generalizability coefficients.

To illustrate our approach with RDD data, we simulate an RDD by building on the existing relationship between pretests and posttests in a dataset from an RCT. Specifically, we use the restricted use file associated with MDRC's Evaluation of Enhanced Academic Instruction in After School Programs, funded by the Institute for Education Sciences (IES) (NCEE 2009-4077). This analysis presents the steps needed to estimate the generalizability coefficient, G_{M^*}.

Findings / Results:
Description of the main findings with specific details.
(May not be applicable for Methods submissions)

Our findings from the non-RDD data provide information about the likely generalizability of RDDs that assign students individually to an educational intervention based on their prior test scores and measure the intervention's effects on their future test scores. For reading and math, respectively, we find that about 66 percent and about 80 percent of the estimated generalizability coefficients (across 5 school districts, 4 grades and 3 subgroups of schools) are at or above 0.5; many are well above 0.5. Thus in a large majority of cases, half or more of the unconditional heterogeneity in observed counterfactual outcomes exists for each value of the observed rating and thus would exist at an RDD cut-point value. In interpreting these results, we make the case that estimated generalizability coefficients based on R-square values for pre-existing outcomes and untreated future outcomes probably represent lower bounds on the generalizability of RDD estimates of treatment effects. We also find that narrowing a target population in terms of school-level characteristics (student income and achievement level) that are correlated with an RDD rating (pre-test scores) can increase one's ability to generalize to the target population.

Conclusions:
Description of conclusions, recommendations, and limitations based on findings.

The ability to generalize RDD effect estimates is likely much greater than is commonly thought. Limitations on the generalizability of RDD findings are only as strong as the relationship that exists between observed ratings and true treatment effects. The relationship depends on the strength of multiple linkages, which are likely to be weak in many cases. We reach this conclusion based on strong theoretical arguments and empirical results. We recommend that researchers apply our framework to RDD studies– using multiple sources of data to estimate generalizability coefficients and using substantive knowledge to interpret their estimates.

Future research in this area will focus on other types of ratings and outcomes, on the RDD setting in which groups of individuals are rated and assigned to a treatments, and on scenarios in which homoscedasticity cannot be assumed.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using Covariates to Improve Precision for Studies that Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59.

Goldberger, A. S. (1972a). *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations*. University of Wisconsin, Institute for Research on Poverty, June. Discussion Paper 126-72, Madison, WI.

Goldberger, A. S. (1972b). *Selection Bias in Evaluating Treatment Effects: The Case of Interaction*. Institute for Research on Poverty. Madison, WI.

Hahn, J., Todd, P., & Klaauw, W. v. d. (2001). Identification and Estimation of Treatment Effects with a Regression Discontinuity Design. *Econometrica, 69*, 201-209.

Imbens, G., & Lemieux, T. (2008). Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics, 142(2)*, 615-635.

Lee, D. S. (2008). Randomized Experiments from Non-Random Selection in U.S. Senate House Elections. *Journal of Econometrics*(142), 807-828.

Lee, D. S., & Lemieux, T. (2009). Regression Discontinuity Designs in Economics.

Mosteller, F. (1990). Improving research methodology: An overview. In L. Sechrest, Perrin, E. and Bunker, J. (Ed.), *Research Methodology: Strengthening causal interpretations of nonexperimental data* (pp. 221-230). Rockville, MD: U.S. Public Health Service, Agency for Health Care Policy and Research.

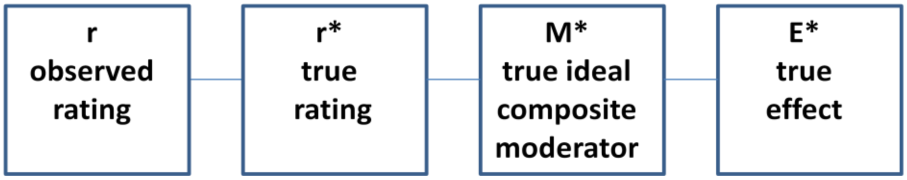
Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.

Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics, 2*(1), 1-26.

Appendix B. Tables and Figures

Not included in page count.

Figure 1
A Model of the Relationship between Observed RDD Ratings and True Treatment Effects


r
observed
rating
r*
true
rating
M*
true ideal
composite
moderator
E*
true
effect