Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 19, Number 9, August 2014 ISSN 1531-7714

Estimating unbiased treatment effects in education using a regression discontinuity design

William C. Smith
Pennsylvania State University

The ability of regression discontinuity (RD) designs to provide an unbiased treatment effect while overcoming the ethical concerns plagued by Random Control Trials (RCTs) make it a valuable and useful approach in education evaluation. RD is the only explicitly recognized quasi-experimental approach identified by the Institute of Education Statistics to meet the prerequisites of a causal relationship. Unfortunately, the statistical complexity of the RD design has limited its application in education research. This article provides a less technical introduction to RD for education researchers and practitioners. Using visual analysis to aide conceptual understanding, the article walks readers through the essential steps of a Sharp RD design using hypothetical, but realistic, district intervention data and provides additional resources for further exploration.

The 'gold standard' for evaluating interventions is the well known Random Control Trial (RCT) where individuals are randomly assigned to a treatment group (which receives the intervention) or a control group (which does not receive the intervention). RCTs, however, are often inappropriate in educational settings due to the ethical concerns over random assignment. As one of the goals of education is to reduce the achievement gap between disadvantaged groups and their more privileged peers, excluding the treatment from those that need it most, simply to comply with the requirements of RCT, can be vehemently opposed by parents and challenged on moral grounds. The Regression Discontinuity (RD) design, the focus of this article, is an alternative to RCT by providing unbiased estimates of the treatment effect while overcoming the ethical concerns associated with random assignment. RD is the only explicitly recognized quasi-experimental approach identified by the Institute of Education Statistics to meet the prerequisites of a causal relationship (IES, 2013). It been increasingly used to evaluate social interventions and is becoming common in fields such as economics. However, the statistical complexity inherent in RD means that the available analyses in the field of education tend to be technical in nature.

This article provides a less technical introduction to RD for education researchers and practitioners. Through a hypothetical scenario, laid out in a sub-urban school district in the Northeastern United States, the article walks the reader through the essential steps in conducting RD, identifies limitations and suggests additional resources. Throughout the article visual analysis is highlighted to aide conceptual understanding. The first section of the article defines RD and explains how it estimates unbiased treatment effects. This is followed by responses to two broad but important questions: Why is RD important? And why is RD appropriate for evaluating educational interventions? The fourth section lies out the hypothetical scenario, including the data set and policy intervention. Section five uses the hypothetical data to illustrate the steps necessary when conducting a basic RD analysis. Limitations and other resources for RD conclude the article.

What is RD?

The RD design is named after the discontinuity or displacement of the regression line at a given point in an assignment variable that differentiates those in the treatment group from those in the control group. The assignment variable is a continuous variable, often based on merit or need, which is used to designate individuals above or below a set cut-point for the intervention. Figure 1 displays a linear regression line. Note how the line is without any apparent breaks, suggesting an equivalent effect on the outcome variable (y-axis) for each individual, regardless of their position on an assignment variable (x-axis). Figure 2 illustrates a discontinuity in the regression line at the assignment score Z. The gap between the solid line and the dashed line can be understood as a change in the y-intercept at the assigned score, centered on Z. The difference between the intercept of the dashed line (treatment group) and the solid line (control group) at this cut-point provides an unbiased estimate of the treatment effect (Shadish, Cook & Campbell, 2002). This effect is point-specific and interpreted as the unbiased treatment effect of those that barely received treatment relative to those that barely failed to receive treatment.

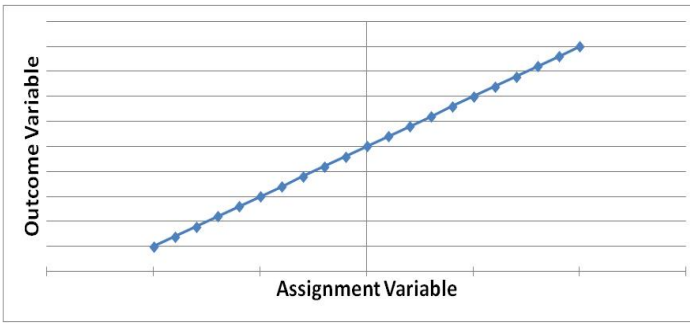

Figure 1: Typical Linear Regression

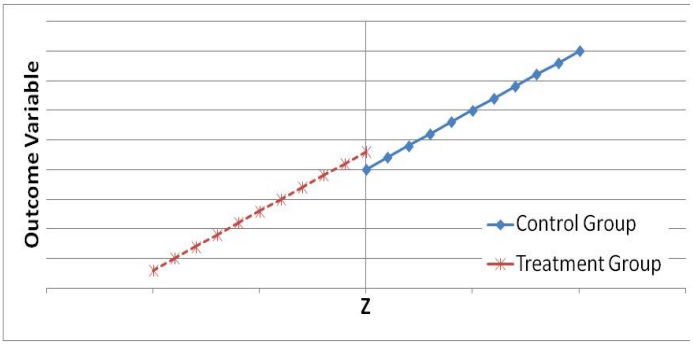

Figure 2: Regression with Discontinuity at Z

RD designs are able to provide an unbiased estimate of the treatment effect because the process in which individuals are selected into the treatment group is "completely known and perfectly measured" (Shadish et al., 2002, p. 224). Measurement error is not present because the score on the assignment variable is used to distinguish groups not inform a construct. For example, a student's family income can be used as an assignment score to identify who has access to a treatment (i.e. free or reduced lunch). The cut-point of this assignment score, however, is only used to separate those in the treatment group from those in the control group, not to measure a broader concept, such as socio-economic status, which it would do so imperfectly with measurement error. Since the selection process is completely known and perfectly measured any difference between the treatment and control group at the cut-point should be "due to either the intervention or to random fluctuation" (Luyten, 2006, p. 399).

The first published article using a RD design dates back to 1960 (Thistelwaite & Campbell, 1960). After a period of inactivity, the approach was reinvented in various fields in the 1970s and 1980s (Shadish et al. 2002). In education, an initial rush to use the design to evaluate the effect of Title I funding on schools after the Elementary and School Education Act of 1965 quickly diminished. Currently, the method has received relatively little attention in the field (Shadish et al. 2002).

Why is RD important?

RD is an important methodological approach because it provides stronger evidence for causal inference than any design outside of random assignment (Shadish et al., 2002). When using the same data set some researchers have found equivalent effect sizes in RCT and RD estimates (Finkelstein, Levin & Robbins, 1996). One of the strengths of the RD design is its ability to provide a counterfactual for the treatment and control group. Counterfactuals attempt to address the question; what would the treatment effect be if the same individual could be in both the control and the treatment group. Counterfactuals, therefore, present an impossibility that researchers attempt to address by having statistically equivalent groups prior to treatment (RCTs) or comparing matched individual post hoc (i.e. propensity score matching). Figure 3 illustrates how a counterfactual is generated through RD. The figure extends the regression lines used in Figure 2 across the cut-point providing both the control group and treatment group with their actual effects (solid line) and

their estimated effects representing the counterfactual (dashed line).

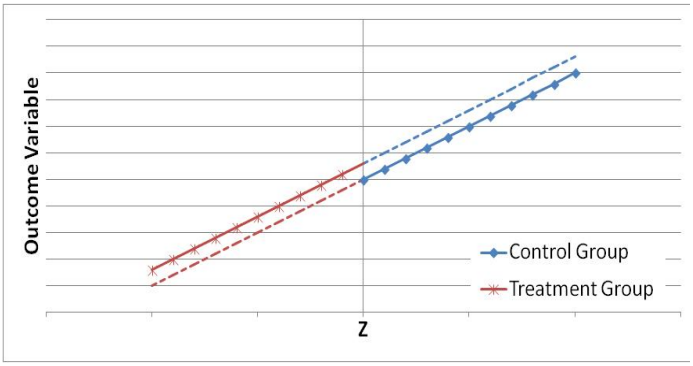

Figure 3: Regression with Discontinuity at Z and Counterfactual

An additional advantage of the RD design is its ability to provide unbiased estimates without the need for additional background information (Luyten, 2006). All that is needed for RD is information on the assignment variable, the classification into the treatment and control group, and the outcome[1]. The ability of RD to function without additional background 'controls', including previous achievement, make it appropriate for use with cross-sectional data. Cross-sectional data is often the preferred data in education because the collection of longitudinal data is time consuming, expensive, and prone to problems of attrition. Although not necessary for estimation, when used the inclusion of relevant background variables in the RD model can increase precision and statistical power.

Why is RD appropriate for evaluating educational interventions?

RD is an excellent fit for educational evaluations because it overcomes the ethical concerns and other objections which often plague RCTs; concerns about a meritous group being denied a reward or a group in need being denied support (Linden, Adams & Roberts, 2006; Shadish et al., 2002). RD is a quasi-experimental approach, often discussed with other quasi-experimental methods such as propensity score analysis (Adelson, 2013; Stone & Tang, 2013) and instrumental variable estimation (Murnane & Willet, 2011; Shadish et al., 2002). Unlike the above designs, which can also be used post-hoc on cross-sectional data, the RD design provides unbiased estimates of similar strength to RCTs (Finkelstein et al., 1996). Moreover, RD is recognized by the Institute for Educational Statistics as the only quasi-experimental design that meets the necessary prerequisites for establishing a causal relationship (IES, 2013).

In addition to the unbiased effects available with RD, the typical application of education policy makes it an appealing methodological choice. School policy is often uni-laterally mandated by a school or district, creating little variation in application. Without this variation, where some schools/groups being assigned to a treatment group (new school policy enacted) and others assigned to a control group (new school policy not enacted), RCTs are impossible, making RD the best possible approach. RD designs have gained momentum in education research and been used to evaluate a wide range of interventions including: class size reduction (Angrist & Lavy, 1999); compensatory reading programs (Trochim, 1984); developmental math programs (Lesik, 2007); high school exit exams (Ou, 2010); financial aid offers (van der Klaauw, 2002); Head Start (Ludwig & Miller, 2007); and school facility investments (Cellini, Ferreira & Rothstein, 2010).

How to Apply a RD Design

To illustrate a RD design the following section contextualizes the key elements and necessary steps in analysis by introducing a hypothetical but realistic district level educational intervention.

Hypothetical Scenario

A school district in the Northeast of the United States (hereafter Northeastern SD) has seen a sharp influx of English language learners (ELLs) over the past ten years. ELLs are one of the fastest growing student demographics in U.S. schools (Uro & Barrio, 2013). No longer are immigrant ELL students restricted to metro areas that dot the U.S. periphery, increasingly these students and their families are moving to areas of the U.S. that have not been traditional destination states (Terrazas, 2011). Research suggests that this changing demographic will bring new challenges to school districts as ELL students consistently score below their native English speaking peers in content courses such as math and science (Hemphill & Vanneman, 2011; Valle et al., 2013). As the ELL population reaches 1/3rd of the student body, Northeastern SD decided to implement a policy in which the ELL students with the lowest language proficiency received weekly pull out lessons focusing on content specific vocabulary terms. The

[1] Additional details will be provided in the hypothetical scenario section.

focus on content specific vocabulary terms is designed to improve student's academic language, which affects achievement (Calderon, Slavin & Sanchez, 2011; Duran, 2008; Haag et al., 2013).

To decide who qualifies for the intervention Northeastern SD administered the World Class Instructional Design and Assessment (WIDA) to the 200 incoming ELL sixth grade students (mean = 382.77, SD = 71.56). The WIDA includes scaled scores for reading, writing, listening, and speaking (WIDA, 2013). The district decided to use the overall composite score of the weighted subscales to identify those with low language proficiency. The composite score ranges from 100 to 600. Students that scored below a 400 were identified as low language proficient.

Steps in a RD

In this section the above scenario is used to demonstrate the necessary steps in a basic Sharp RD design, as outlined by Jacob et al. (2012).

Step 1: Is the RD design an appropriate approach?

Prior to the application of RD one must evaluate whether the method is a suitable fit for the research question and the available data. The research query under investigation by the district is "Does the pull out intervention have a significant effect on end of year math achievement scores[2] of ELL students?" As the intervention requires an investment of multiple English as Second Language (ESL) teachers, a non-significant finding may indicate the inefficient use of resources.

To apply a RD design successfully the assignment variable, identification of the cut-point, and classification into the treatment group must meet the following set of qualifications. The assignment variable, also called the forcing variable (Murnane & Willett, 2011), is generally used to identify differences in merit or need. In this scenario the WIDA composite score acts as the assignment variable, separating out those that need more support from those that need less support. Using the assignment variable for equity purposes maximizes the designs "ability to use research-based practice guidelines, survey instruments and other tools to identify those individuals in greatest need and then assign them to the intervention" (Linden et al., 2006, p. 125). Although the WIDA score and math achievement are positively correlated in this scenario, the assignment variable does not need to be related to the outcome variable (Shadish et al., 2002). Finally the assignment variable should be ordinal or continuous in nature (Linden et al., 2006), making the WIDA composite score an appropriate assignment variable.

The cut-point or cut off score is the exogenously set score on the assignment variable that differentiates those that get the intervention (treatment group) from those that do not receive the intervention (control group). For the hypothetical scenario the cut-point is 400 with students scoring below that score placed in the treatment group and those at or above that score placed in the control group. Assignment into groups must be based solely on the cut-point score and be known prior to assignment (Shadish et al., 2002). Furthermore, all participants must have a chance to be in the treatment group; i.e. if the cut-point was adjusted across the range of the assignment variable everyone would have a chance to be included in the treatment group (Shadish et al., 2002).

Once established, treatment should be restricted to those in the treatment group (probability of treatment = 1) and omitted from the control group (probability of treatment = 0). When treatment is administered in the above fashion we have a Sharp RD, which will be the focus of this article. A Fuzzy RD occurs when the above assumptions do not hold (Trochim, 1984). Violations of the treatment assumption may be due to attrition, no shows – where those in the treatment group do not receive the treatment (Jacob et al., 2012), or crossover – where "those assigned to treatment do not take it or those assigned to control end up in treatment" (Shadish et al., 2002, p. 228). Score manipulation is also possible, especially if the cut-point is public knowledge. The placement of individuals in the treatment group based on means other than the cut-point may be more likely in education settings as local administrators and teachers often use discretionary power to ensure 'fair' placement (Trochim, 1984). Applying the Sharp RD design in situations where treatment is not restricted to the treatment group can produce bias in estimates (Shadish et al., 2002). However, if 5% or less of cases are misaligned, the deletion of misaligned cases should not significantly decrease the probability of obtaining reasonable treatment effects (Trochim, 1984). In such instances, once misaligned participants are removed a Sharp RD can be conducted.

[2] Scenario sample math achievement (mean = 79.02, SD = 11.73).

Now that it has been verified that the WIDA variable is an appropriate assignment variable and that classification into treatment and control group was based solely on the cut-point of 400 there are two last checks before proceeding. First, the assignment variable must occur prior to treatment, ensuring that the treatment does not affect assignment (Shadish et al., 2002). Second, the outcome variable (math achievement in the scenario) must be continuous or, if dichotomous, modeled linearly using a logit function translation (Linden et al., 2006).

Step 2: Visual examination of the data.

At least two graphs should be produced to initially examine the data. First, a basic scatter plot should be created to identify whether there is indeed a discontinuity or jump at the cut-point. Figure 4 provides a scatter plot for our hypothetical scenario with half of the 200 ELL students scoring below the cut-point of 400 and half above the cut-point. In examining the scatter plot there should be no discontinuities other than that at the cut-point. The scatter plot can also help identify the appropriate functional form and any potential outliers.

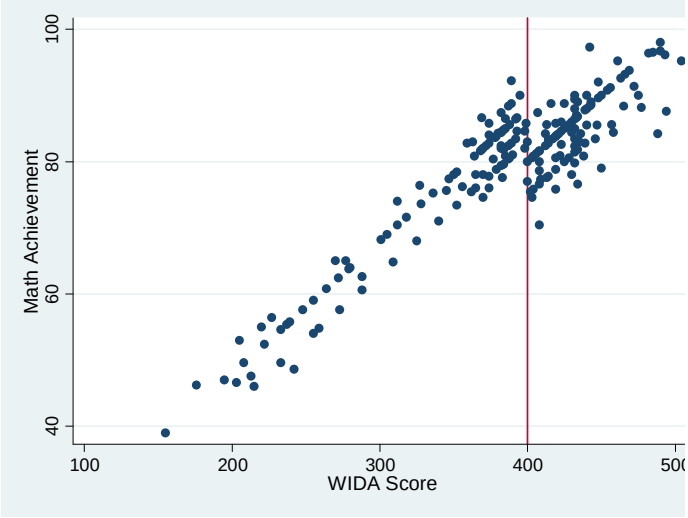

Figure 4: Scatter Plot of WIDA Score by Math Achievement

The second graph should be a frequency histogram of scores on the assignment variable (see Figure 5). Stark differences in frequency just before or after the cut-point can indicate manipulation of the assignment score around the cut-point which threatens the validity of results. Figure 4 and 5 suggest that a discontinuity is present at the established cut-point of 400 and that manipulation of the assignment variable around the cut-point is unlikely.

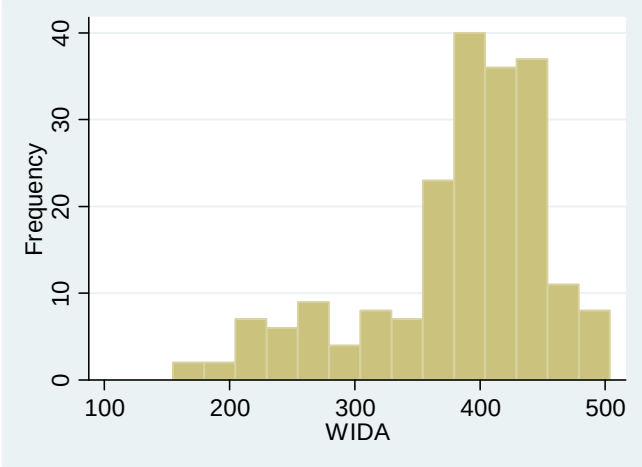

Figure 5: Distribution of WIDA Scores

Step 3: Precision and sample size.

To detect unbiased estimates, a larger sample size is needed when conducting RD, relative to RCTs (Shadish et al., 2002). Cappelleri, Darlington and Trochim (1994) suggest that 2.5 to 3 times as many participants are needed, depending on the predicted effect size. An additional factor to consider is the position of the cut-point relative to the distribution of the assignment variable. Choosing a cut-point on the extreme end of the scale can leave too few points on one side to accurately model the regression (Shadish et al., 2002). This limitation makes it challenging to use RD to evaluate interventions targeting only those most at need or those with the greatest merit. For the scenario, half of the 200 students scored above and below the cut-point to ensure a sufficient sample size.

Step 4: Decide on the bandwidth.

RD designs provide an unbiased estimate of the treatment effect at the cut-point that separates the treatment group from the control group. Although RD can be completed with the full sample, often choosing a smaller sample closer to the cut-point is more appropriate. This may be due to the presence of outliers in the scatter plot or changes in the functional form at the tails of the sample. The bandwidth identifies the range of assignment scores to be included in the analyzed sample. For example, Figure 6 fits a regression line for the treatment group and control group for a reduced sample, setting the bandwidth at 50 (assignment scores = 350-450). The bandwidth is chosen after

visually examining the data and restricts the generalizability of the results. Changing bandwidth is one way to check the robustness of results (see Step 7: Sensitivity Analyses).

Figure 6: Regression Discontinuity with Bandwidth set to 50

Step 5: Estimate the simple linear model.

Equation 1 illustrates an OLS model predicting the outcome variable for individual i from the intercept (β_0), the assignment variable, the treatment, and a random error (ε_i), with β_1 representing the coefficient of the assignment variable and β_2 representing the coefficient for the treatment variable.

Equation 1: OLS Regression Equation

$$Outcome_i = \beta_0 + \beta_1 (Assignment_i) + \beta_2 (Treatment_i) + \varepsilon_i$$

Equation 2 centers the assignment variable at the cut-point and is the simple linear model estimated with our hypothetical data (see Table 1 below).

Equation 2: Centered at the Cut-point

$$Outcome = \beta_0 + \beta_1 (Assignment_i - Cut\text{-}point) + \beta_2 (Treatment_i) + \varepsilon_i$$

As RD estimates the treatment effect at the cut-point, equations 3 and 4 substitute the appropriate terms for the treatment and control group respectively. Subtracting equation 3 (treatment group) from equation 4 (control group) yields the treatment effect (β_2), as illustrated in equation 5. The simple linear regression using a Sharp RD design with a bandwidth of 50 yields an unstandardized treatment effect of 8.28 in the hypothetical scenario, suggesting that those ELL students that participated in the pull out for content vocabulary scored over eight points higher on the end of year mathematics test than ELL students who did not participate in pull out instruction.

Equation 3: At the Cut-point, If Treatment=1

$$Outcome = \beta_0 + \beta_1 (Cut\text{-}point - Cut\text{-}point) + \beta_2 (1) + \varepsilon_i$$

Equation 4: At the Cut-point, If Treatment=0

$$Outcome = \beta_0 + \beta_1 (Cut\text{-}point - Cut\text{-}point) + \beta_2 (0) + \varepsilon_i$$

Equation 5: Difference between Equation 3 and Equation 4

$$[\beta_0 + \beta_1 (0) + \beta_2 (1) + \varepsilon_i] - [\beta_0 + \beta_1 (0) + \beta_2 (0) + \varepsilon_i] = \beta_2 = Treatment\ Effect$$

Step 6: Evaluate functional form.

To provide unbiased estimates of the effect size the functional form must be correctly specified (Shadish et al., 2002). Although the visual examination of the data can suggest a particular functional form, curvilinear and interaction terms should be added at this step to ensure that the model is correctly specified. Curvilinear results may result from the presence of outliers, floor, or ceiling effects. Additionally, interaction terms capture changes in the relative slope of the regression line before and after the cut-point. Jacob et al. (2012) suggest at least six models be ran to evaluate functional form and test the robustness of the treatment estimate. The first model is the simple linear model detailed in Step 5. Model two adds an interaction term between the treatment and assignment variable (assignment*treatment) to equation 2. Model three adds a quadratic term ($assignment^2$) to the simple linear model and model four ads the interaction term to model three. Finally, model five adds a cubic term ($assignment^3$) to the quadratic model and model six ads the interaction term to model five. Table 1 provides the treatment estimates of the pull out intervention for all six models with the bandwidth set at 50.

Table 1: Effect of Pull out Instruction on Math Achievement across Functional Forms

	Coeffi-cient	Stand. Error	95% c.i.
M 1: Linear	**8.280**	1.216	5.874-10.685
M 2: Linear + Interaction	**8.228**	1.249	5.758-10.698
M3: Quadratic	**8.239**	1.240	5.785-10.692
M4: Quadratic + Interaction	**8.227**	1.259	5.737-10.718
M5: Cubic	**7.711**	1.653	4.442-10.980
M6: Cubic + Interaction	**7.623**	1.725	4.210-11.036

Note: Bandwidth set to 50. Unstandardized treatment effect in bold.

Step 7: Sensitivity tests.

Multiple sensitivity tests are suggested to check the robustness of the treatment estimate. As discussed in step 4, the first sensitivity test is expanding the bandwidth to include a greater portion of the overall sample in the analysis. As generalizability is limited by bandwidth, results that remain robust with a larger bandwidth can be generalized to a larger percentage of the sample. Robustness across functional forms also tests the sensitivity of the result to different modeling techniques. Table 1 suggests that the treatment results in the hypothetical scenario are robust. Checking if a linear or curvilinear form best fits the data can also be accomplished by "trimming" or dropping 1%, 5%, and 10% of the outermost data and comparing the results to the original functional form model (Jacob et al., 2012). Finally, Imbens and Lemieux (2008) suggest moving the cut-point as a form of sensitivity analysis. In RD, the treatment effect should only be present at the cut-point separating the treatment group from the control group. The presence of significant treatment effects at pseudo cut-points calls into question the validity of the treatment effect. The non-significant effect of the treatment at pseudo cut-points of 300, 350, 450, and 500 (see Table 2) in the hypothetical scenario provides support that the effect at the cut-point results from differences in the applied treatment.

From the results above we can conclude that pull out instruction has a significant, positive impact on the math achievement on the ELL students that took the WIDA, with the treatment effect ranging from 7.62 to 8.28 points or 0.65 to 0.71 standard deviations. This effect is robust across functional forms and bandwidths[3] with the simple linear model preferred due to its greater precision. Finally, changes in math achievement appear to be solely attributed to the treatment as no other discontinuities are found in the data (see Figure 4 and Table 2).

Table 2: Testing for Additional Discontinuities through Pseudo Cut-points

Cut-point	Coefficient	Standard Error	95% c.i.
400 (Original)	**8.280**	**1.216**	**5.874-10.685**
300	1.575	2.554	-3.721-6.872
350	-0.109	1.883	-3.867-3.649
450	-1.938	1.690	-5.291-1.420
500	0.410	4.861	-9.846-10.666

Note: Bandwidth set to 50. True cut-point in bold.

Limitations

Similar to RCTs, the largest weakness of RD is limited generalizability[4]. Conclusions can only be drawn relative to the sample and are often limited to a specific bandwidth within the sample. As mentioned in step 3 a larger sample size is needed in RD to accurately estimate effect size. Depending on the predicted effect size of the treatment, sufficient sample size is 2.5 to 3 times larger than that needed for RCTs (Cappelleri et al., 1994). Due to sample size concerns, RD may not be appropriate for interventions targeting the top or bottom decile on an assignment score. When the cut-point is placed at the extreme end of the assignment range, too few individuals are placed in the treatment or control group to accurately measure the effect (Shadish et al., 2002).

The primary treat to validity in an RD design is the improper modeling of functional form. Other threats to validity are relatively miniscule. To threaten the internal validity of the results, an omitted variable would have to produce a discontinuity at the exact cut-point, which is highly unlikely (Shadish et al., 2002). The specificity required of the threat also reduces or eliminates concerns with regression to the mean (Linden et al., 2006). Maturation is a potential threat that must be captured by the appropriate functional form and a selection-instrumentation threat is possible if ceiling or floor effects are present (Shadish et al., 2002).

Other Resources for RD

The above example applies a Sharp RD design to a hypothetical scenario that meets all the requirements laid

[3] Treatment effects for bandwidth at 100, 200 and the entire sample remained positive and significant.

[4] For more on generalizability see Jacob et al. (2012), p. 58-60.

out in step 1. For more information on a Fuzzy RD design, where the probability of the treatment being applied solely to the treatment group is less than one, see Bloom (2012) or van der Klaauw's (2002) example applying a Fuzzy RD design to investigate the effect of financial aid offers on college enrollment. RD can also be incorporated into other methodological approaches. For an example of how RD can be used within a multi-level framework see Luyten (2006). This can be incredibly useful in education where students are nested within classrooms nested within schools. Additionally, for an example of RD using a time-series approach see Lesik's (2007) estimation of developmental mathematics programs on student retention.

For those individuals more familiar with difference in differences approaches see chapter nine of Murnane and Willett (2011) which introduces RD as an extension of difference in differences. Finally, for an extended, less technical breakdown of all topics discussed in this article see Jacob et al. (2012) which provides an excellent guide complete with checklists for researchers to apply Sharp and Fuzzy RD designs.

References

Adelson, J. (2013). Educational research with real-world data: Reducing selection bias with propensity scores. *Practical Assessment, Research and Evaluation*, 18(15). Available at http://pareonline.net/pdf/v18n15.pdf

Angrist, J. & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on student achievement. *Quarterly Journal of Economics*, 114 (May), 535-575.

Bloom, H. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*, 5(1), 43-82.

Calderón, M., Slavin, R., & Sánchez, M. (2011). Effective instruction for English learners. *The Future of Children*, 21(1), 103-127.

Cappelleri, J., Darlington, R. & Trochim, W.M.K. (1994). Power analysis of cutoff-based randomized clinical trials. *Behavioral Assessment*, 9, 169-177.

Cellini, S.R., Ferreira, F. & Rothstein, J. (2010). The value of school facility improvements: Evidence from a dynamic regression discontinuity design. *The Quarterly Journal of Economics*, 125(1), 215-261.

Durán, R. P. (2008). Assessing English-language learners' achievement. *Review of Research in Education*, 32(1), 292-327.

Finkelstein, M.O., Levin, B. & Robbins, H. (1996). Clinical and prophylactic trials with assured new treatment for those at greater risk: II. Examples. *American Journal of Public Health*, 86(5), 696-705.

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction*, 28, 24-34.

Hemphill, F. C., & Vanneman, A. (2011). Achievement Gaps: How Hispanic and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress. Statistical Analysis Report. NCES 2011-459. *National Center for Education Statistics*.

IES (2013). Postsecondary Education Evidence Review Protocol. IES: What Work's Clearinghouse. Available at http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=242

Imbens, G.W. & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142, 615-635.

Jacob, R., Zhu, P., Somers, M.A. & Bloom, H. (2012). *A Practical Guide to Regression Discontinuity*. New York, NY: MDRC. Available at http://www.mdrc.org/sites/default/files/regression_discontinuity_full.pdf

Lesik, S.A. (2007). Do developmental mathematics programs have a causal impact on student retention? An application of discrete-time survival and regression discontinuity analysis. *Research in Higher Education*, 48(5), 583-608.

Linden, A., Adams, J. & Roberts, N. (2006). Evaluating disease management programme effectiveness: An introduction to the regression discontinuity design. *Journal of Evaluation in Clinical Practice,* 12(2), 124-131.

Ludwig, J. & Miller, D. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, 122(1), 159-208.

Luyten, H. (2006). An empirical assessment of the absolute effect of schooling: Regression discontinuity applied to TIMSS-95. *Oxford Review of Education*, 32(3), 397-429.

Murnane, R. & Willett, J. (2011). *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. New York, NY: Oxford University Press.

Ou, D. (2010). To leave or not to leave? A regression discontinuity analysis of the impact of failing the high school exit exam. *Economics of Education Review*, 29, 171-186.

Shadish, W., Cook, T. & Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York, NY: Houghton Mifflin Company.

Stone, C.A. & Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations.

Practical Assessment, Research and Evaluation, 18(13). Available at http://pareonline.net/getvn.asp?v=18&n=13

Terrazas, A. (2011). Immigrants in new destination states. *Migration Policy Institute*. Available at http://www.migrationpolicy.org/article/immigrants-new-destination-states

Thistelwaite, D. & Campbell, D. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51, 309-317.

Trochim, W.M.K. (1984). *Research Design for Program Evaluation: The Regression Discontinuity Approach*. Newbury Park, CA: Sage.

Uro, G., & Barrio, A. (2013). *English Language Learners in America's Great City Schools: Demographics, Achievement and Staffing*. Washington, D.C.: Council of the Great City Schools. Available at http://files.eric.ed.gov/fulltext/ED543305.pdf

Valle, M. S., Waxman, H. C., Diaz, Z., & Padrón, Y. N. (2013). Classroom instruction and the mathematics achievement of non-English learners and English learners. *The Journal of Educational Research*, 106(3), 173-182.

Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review*, 43(4), 1249-1287.

WIDA (2013). Access for ELLs: Interpretive Guide for Score Reporting. Available at http://www.wida.us/assessment/access/

Citation:

Smith, William C. (2014). Estimating unbiased treatment effects in education using a regression discontinuity design. *Practical Assessment, Research & Evaluation*, 19(9). Available online: http://pareonline.net/getvn.asp?v=19&n=9

Author:

William C. Smith
Education Theory and Policy
Comparative International Education
Pennsylvania State University
State College, PA 16801

Email: wcs152 [at] psu.edu