Using Regression Discontinuity Design for Program Evaluation

Hyunshik Lee[1] and Tom Munk[1]
[1]Westat, 1650 Research Blvd., Rockville, MD 20850

Abstract
In many education and social intervention programs it is thought efficacious and ethical to offer a treatment to individuals based on a measure such as need or worthiness. This entails assigning subjects to the treatment group if they are at or above some cutoff score on the measure. The outcome of the treatment is then analyzed using the regression discontinuity design technique assuming that the treatment effect can be detected at the assignment cutoff through regression analysis. In this study the power of such a design is analyzed to determine the sample size necessary for implementing the parametric regression model under some common assumptions. Violations of those assumptions threaten the validity of the RDD study. Ways to address such violations are also discussed.

Key Words: Effect size, power, sample size, random control trial, relative efficiency

1. Introduction

Regression discontinuity design (RDD) is a popular quasi-experimental design used to evaluate program effects. It differs from the randomized control trial (RCT) mainly in the assignment of study subjects. In RCT, subjects are assigned to the treatment group(s) and control group randomly. Any detected effects can be attributed to the program because any confounding effects are nullified by the random assignment. However, in RDD, this assignment principle is deliberately violated by assigning subjects to the treatment by a non-random assignment rule.

In many education and social intervention programs it is thought efficacious and ethical to offer the treatment to individuals based on a measure such as need or worthiness. This entails assigning subjects to the treatment group if they are above (or below) some cutoff score. For example, in a study of a remedial writing course for college students, students were assigned to a remedial writing course when their SAT or ACT score was below a certain cutoff point (Aiken et al., 1998). In another example, the Reading First Impact Study, schools from each participating district were selected for a reading intervention if they were above or below a cutoff on a rating of their need for and/or ability to benefit from the intervention (Bloom et al., 2005). Obviously, RCT is not feasible for these examples because assignment of subjects is not random. However, the assignment is based on a selection variable with a cut score. Employing RDD, one tries to use this feature in evaluating the program effect by assuming that the outcome variable is a continuous function of the assignment score before the treatment is applied. If the treatment has an effect on the outcome variable, then there will be a jump or drop (i.e., discontinuity) in the regression line at the cutoff point. An estimated size of this discontinuity is used as an estimate of the program effect.

To describe the basic idea of the design, suppose that the pre-treatment relationship between the assignment score (S) and the outcome variable (Y) is given by the following straight line linear regression:

$$Y = \alpha + \beta S + \varepsilon, \tag{1}$$

where α and β are regression coefficients, and ε is the error term. After treatment, if the treated subjects are affected by a constant treatment effect (β_0) on the outcome variable, then the regression equation can be revised as follows:

$$Y_i = \alpha + \beta_0 T_i + \beta_1 S_i + \varepsilon_i \,, \tag{2}$$

where T_i is the assignment indicator having a value of 1 if subject i is assigned to the treatment, or 0, if it is assigned to the control group. Because of the constant effect assumption, the slope of the regression line does not change but the intercept term changes to $\alpha + \beta_0$ for the treatment group, β_0 being the constant effect. We interpret this constant effect β_0 as the program effect. The basic idea of this simple RDD is depicted graphically in figure 1.

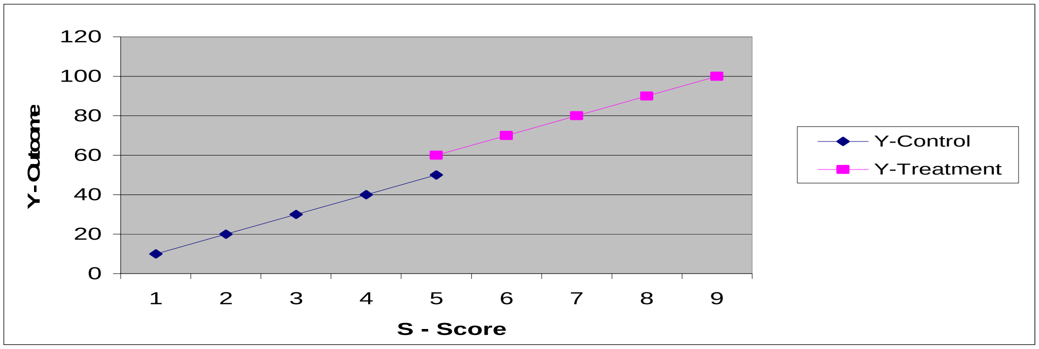

Figure 1: Basic straight line regression model for RDD

In reality, the RDD model is often more complex than the simple straight line regression model used in the illustration. The constant effect over the range of the score variable may not hold either, and real life application of RDD may face various issues that could invalidate the results of RDD analysis. These issues will be discussed in section 4. Nevertheless, the technique has been successfully applied in many areas such as education, sociology, psychology, criminology, health, etc.

The technique was first introduced by Thistlewaite and Campbell (1960) in educational psychology. Goldberger (1972) independently reinvented the technique in economics as well as Tallmadge and Horst (1976) rediscovered it in education in a somewhat different form, estimating the treatment effect at the mean of the assignment score instead at the cutoff point. Before the paper by Tallmadge and Horst, research in RDD was mostly academic and presented as “illustrative analyses” (Trochim, 1984). However, Tallmadge and Horst presented RDD as an option to evaluate Title I programs of the 1965 Elementary and Secondary Education Act, and about 40 applications were made in the 1979-1980 academic year (Trochim, 1984). Other than that, the technique did not attract much attention except some sporadic applications until late 1990, when economists, disappointed by the instrumental variable method of the Heckman-type, started looking at alternatives (see Imbens and Lemieux, 2008).

In statistics, there are even fewer publications on RDD except for a few methodologists such as Rubin (1977) who contributed to the theory. Rubin provided a formal treatise for estimating intervention effect when assignment of treatment is solely based on a covariate. RDD was a special case in this formulation. He noticed that without a priori information about the relationship between the outcome and the covariate, it would not be feasible to estimate the treatment effect under RDD due to lack of overlap between the treatment and control groups.

In education RDD staged a comeback when the influential Institute of Education Sciences (2004) has specified the use of RDD if RCT is not feasible. Bloom et al. (2005) was a manifestation of this new trend in education. Recently, Schochet (2008) wrote about statistical power of RDD for IES focussing on the clustered RDD.

Following this trend, this paper attempts to alert statisticians to the usefulness of RDD, when RCT is not feasible and the assignment strategy fits in the design of regression-discontinuity. This paper is organized as follows: in Section 2, we will discuss assumptions for parametric RDD and estimation of the treatment effect under these assumptions. In Section 3, we will talk about power and sample size determination. Section 4 is devoted to issues concerning RDD applications and possible options to address those issues. In Section 5, we will give some conclusions

2. Estimation of Treatment Effect

2.1 Assumptions of RDD

As with any evaluation design RDD requires some basic assumptions. The first is about the unique feature of the assignment strategy to the treatment and control groups. It is assumed to be fully known in advance, and solely based on a score variable *S*. Study subjects are assigned to the treatment group if their score is at or above the cutoff and all others to the control group, or vice versa. If there is no misassignment, the design is called "sharp". On the other hand, if there are some misassignments or treatment crossovers, the design is called "fuzzy". We will mainly focus on the sharp design in this paper with only a brief discussion on the fuzzy RDD. A pretest score is often as the assignment score in education.

The second assumption is that a single regression model holds for the pre-treatment relationship between the outcome variable, Y and the score variable, S over the entire range of S.

The third assumption is that there is no other factor that causes discontinuity at the cutoff so that any discontinuity at the cutoff can be attributed to the treatment.

The fourth assumption is the usual assumption for implementation of an experiment, that is, stable unit treatment value assumption (SUTVA) (Rubin, 1977), which is often neglected or tacitly assumed in the RDD literature. This is the "no interference" principle, where treatment of a unit does not interfere in the outcome of treatment of other units.

The fifth assumption is that the conditional average treatment effect (ATE) given the score variable, S is the same for all S. Under this assumption, the regression equations of the treatment and control groups are parallel if they are extended beyond the cutoff boundary. In this sense, the regression equation of the control group plays the role of counterfactual.

These are strong assumptions, which are often violated in reality. If that happens, it could destroy the internal validity of the analysis results. However, we can still obtain some useful results by using different analytical strategies that are suitable for such situations. We will discuss these issues in Section 4.

2.2 Parametric RDD Models

The second assumption provides a parametric framework by which a RDD model can be specified and the treatment effect can be estimated. However, many authors advocate the use of the non-parametric method (e.g., Hahn, Tood, and Van Der Klaauw, 2001; Imbens and Lemieux, 2008). The non-parametric approach is more flexible because it can work under more relaxed assumptions. We will discuss more about this later but for now we will use the parametric set-up since we are particularly interested in the power of the significance test of whether the treatment has a detectable effect or not. We are also interested in the determination of the required sample size to meet certain design criteria to plan an RDD. For these purposes, it is more difficult to use the non-parametric framework.

Model specification is very important because if it is wrong, the estimate of the treatment effect can be severely biased, resulting in an erroneous inference. If there is a priori information about the pre-treatment regression model, one can use it but if such information is lacking (as is often the case in reality), it is essential to examine the plot of the real data. In education, a standard test is often used as the outcome measure, so a good pre-treatment regression model may be known in advance. Even when a priori information is available, it is prudent to check the model using relevant data.

According to the second and fifth assumptions, the parametric RDD model for a single treatment and a single control group is expressed by the following regression equation:

$$Y_i = \alpha + \beta_0 T_i + \beta_1 g\left(S_i\right) + \varepsilon_i \tag{3}$$

where for study subject i, Y_i:outcome variable, T_i:assignment status, S_i : score variable, $g\left(S\right)$: function of S_i, and e_i : error term.

Our approach closely follows Trochim (1984) and Bloom et al. (2005), and it can be analyzed as an ANCOVA model as done by Aiken et al. (1998). From this perspective, Cappelleri, Darlington, and Trochim (1994) studied the power and sample size issue using the Fisher's z-transformation of the partial correlation coefficient between Y and T controlling for S . Schochet (2008) took an approach similar to ours to analyze the power for the clustered RDD.

If the model holds, the coefficient β_0 is the treatment effect (the jump or drop in the regression lines at the cutoff point), and its estimation can be made by fitting a single regression equation (3). Originally, however, a separate regression line was fitted for each of the treatment and control groups (Sween, 1971). Those who use the non-parametric method still advocate fitting the data separately (Hahn, Tood, and Van Der Klaauw, 2001; Imbens and Lemieux, 2008).

Most of functional relationships $\left(g\left(S\right)\right)$ can be described by a polynomial function by including high enough order terms. It is often advisable to include a term one order higher in the model than the plot may suggest. It can be tested whether such terms are needed or not.

A more general model includes interaction terms between T and polynomial terms in S . In this case, the fifth assumption does not hold any longer. It is advantageous to use other covariates, which are highly correlated with the outcome variable.

3. Power Analysis and Sample Size Determination

The primary goal of an evaluation study of a treatment effect is to estimate β_0 in (3), and perform a significance test for the following null and alternative hypotheses:

$$H_0 : \beta_0 = 0 \quad \text{vs.} \quad H_1 : \beta_0 > 0 . \tag{4}$$

Here we assume that a positive effect is of primary interest, so we use one-tailed test. We will use usual notation α for the one-tailed significance level (type I error), which should not be confused with the intercept term in (3), and β for the type II error.

We use the ordinary least square estimator (OLS) to estimate β_0 in (3), which follows asymptotically a normal distribution with the following asymptotic variance (see Bloom et al., 2005):

$$V_{RDD}\left(\hat{\beta}_0\right) = \frac{\sigma^2\left(1-R_M^2\right)}{nP\left(1-P\right)\left(1-R_T^2\right)} , \tag{5}$$

where σ^2 : variance of the outcome variable, Y , P : proportion of subjects assigned to treatment, R_T^2 : squared correlation between T and S , and R_M^2 : R -squared statistic for the regression model.

Assuming that all parameters in (5) are known, we can use the normal test to test the hypotheses in (4), and from the test statistics, we can derive the formulae for determination of the sample size and the power of the significance test.

3.1 Sample Size Determination

When an RDD is planned, it is required to determine the sample size or analyze the power of the significance test for a given sample size. It is customary to use the standardized effect size, β_0/σ for power analysis and sample size determination to avoid the scale effect in different outcome measures (Cohen, 1988). Then the sample size is given in terms of desired minimum detectable (standardized) effect size (MDES). A MDES value between 0.2 and 0.4 is most frequently used in education. Also we need to set the desired significance level $\left(\alpha\right)$ and power level $\left(1-\beta\right)$. The sample size is then determined by

$$n = \left(1 - R_M^2\right)\left(z_{1-\alpha} - z_\beta\right)^2 \Big/ \left\{M^2 P\left(1-P\right)\left(1-R_T^2\right)\right\}, \qquad (6)$$

where M is the desired MDES, and z_α and $z_{1-\beta}$ are normal 100α-th and $100(1-\beta)$-th percentiles, respectively. Table 1 gives the sample sizes for various values of MDES (ranging between 0.2 and 0.5) and R_M^2 (ranging between 0.01 and 0.5) with 5 percent significance level, 80 percent power, $P = 0.5,$ and $R_T^2 = 2/3$.

Table 1: RDD sample size for various MDES and R_M^2 values
$\left(\alpha = 0.05,\ 1-\beta = 0.8,\ P = 0.5,\ \text{and}\ R_T^R = 2/3\right)$

	R_M^2 *- RDD model R-squared*				
MDES	*0.0*	*0.2*	*0.3*	*0.4*	*0.5*
0.200	1,855	1,484	1,298	1,113	927
0.250	1,187	950	831	712	594
0.300	824	660	577	495	412
0.350	606	485	424	363	303
0.400	464	371	325	278	232
0.450	366	293	256	220	183
0.500	297	237	208	178	148

The formula (6) is a large sample formula. For small sample sizes, it would be better to use Student t percentiles with appropriate degrees of freedom rather than normal percentiles in the formula. However, the number of degrees of freedom is unknown when the sample size is unknown. One can start with the sample size determined by using normal percentiles, and then use the determined sample size to decide approximate degrees of freedom. If the new sample size is very different from the first one, then the degrees of freedom is revised based on the new sample size. Only a few iterations will suffice for the sample size to converge. There is an easier way to make this adjustment. As can be seen from Table 2, it is only necessary to increase the normal-based sample size by just three. For example, if the normal-based sample size is 20, then the Student t-based sample size is 23. It is interesting to see that the increment is always 3 even for a very large sample size (e.g., 100,000).

Table 2: Control between normal-based and student t-based sample sizes

Normal-based Sample size	*A:* $(z_{1-\alpha} - z_\beta)^2$	*B:* $(t_{1-\alpha} - t_\beta)^2$	*Ratio (B/A)*	*Student t-based sample size*
20	6.182556	6.814977	0.907201	23
100	6.182556	6.281359	0.984270	103
1,000	6.182556	6.191972	0.998479	1,003
10,000	6.182556	6.183491	0.999849	10,003
100,000	6.182556	6.182651	0.999985	100,003

3.2 Power Calculation

Assuming a large sample size, the power of the significance test is given by

$$1 - \beta = 1 - \Pr\left\{ Z < \left(z_{1-\alpha} - M\sqrt{\frac{nP\left(1-P\right)\left(1-R_T^2\right)}{1-R_M^2}} \right) \right\}, \qquad (7)$$

where Z is the standard normal variate and $z_{1-\alpha}$ is a normal percentile, which can be replaced by a Student t variate and percentile, $t_{1-\alpha}$ with appropriate degrees of freedom (df) if n is small $(df = n -$ number of regression coefficients in the RDD model$)$.

Table 3 provides the power in percent of the significance test for various MDES values, with 5 percent significance level, $P = 0.5,\ R_T^2 = 2/3,\ n = 500.$

Table 3: Power (%) of the Significance Test for the Effect Size $\left(\alpha = 0.05,\ P = 0.05,\ R_T^2 = 2/3,\text{ and } n = 500\right)$

	R_M^2 *- RDD model R-squared*				
MDSE	*0.0*	*0.2*	*0.3*	*0.4*	*0.5*
0.200	36	42	46	51	57
0.250	49	56	61	67	74
0.300	61	70	75	80	86
0.350	73	81	85	90	94
0.400	83	89	93	95	98
0.450	90	95	97	98	99
0.500	94	98	99	99	100

3.3 Comparison of RDD with RCT

It is well known that RDD is less efficient than RCT if everything is the same except the assignment strategy. Goldberger (1972) calculated the relative efficiency of RCT over RDD to be 2.75 if the assignment score is normal. This means that an RDD needs a 2.75 times larger sample size than RCT to be able to detect the same MDES. This factor is 4 for the uniform distribution, and between 2.75 and 4 for many real life examples. This difference owes to the absence of collinearity between T and S for the RCT design because of random assignment. The regression model for RCT is also given by (3) but $R_T^2 = 0$ for RCT while $R_T^2 > 0$ for RDD. The variance of $\hat{\beta}_0$ for RCT is then given by

$$V_{RCT}\left(\hat{\beta}_0\right) = \frac{\sigma^2\left(1 - R_M^2\right)}{nP\left(1-P\right)}, \tag{8}$$

which does not have $1 - R_T^2$ in the denominator, contrary to the variance formula in (5). So the relative efficiency (RE) is given by the ratio of (8) to (5) and written as

$$\text{RE} = \frac{1}{1 - R_T^2}. \tag{9}$$

In Tables 1 and 2, $R_T^2 = 2/3$ and thus, the RE of 3 was used.

4. Issues in RDD Applications and Options

In RDD, the shape of the functional relation between the outcome variable and the assignment score is very important. Misspecification of the model can lead to biased estimation of the treatment effect. So when a polynomial model is fitted, a term of higher order than the plot may suggest should be included in the starting model. However, it becomes less efficient. In this case, a balanced design $(\text{i.e., } P = 0.5)$, in which the treatment and control groups are of equal size, is less affected by the inclusion of higher order terms than the unbalanced design. To achieve balance, some cases may need to be thrown out, but then the sample size and efficiency are reduced.

To reduce the bias due to model misspecification, over-fitting of the model may be necessary but it will require a larger sample size or result in lower efficiency. On the other hand, under-fitting of the parametric model causes a larger bias. So it is necessary to strike a balance between bias and efficiency (variance-bias trade-off).

To avoid model misspecification, many recent authors use the nonparametric method (e.g., Hahn, Tood, and Van Der Klaauw, 2001; Imbens and Lemieux, 2008). One notable drawback of this approach is that it requires a larger sample size than the parametric approach to meet the same efficiency requirement.

The fifth assumption given in Subsection 2.1 is needed to use the estimated β_0 as the overall average treatment effect. The assumption is based on what is known to be the principle of extrapolation, where in our case, the regression slope for the treatment group can be extrapolated over the range of the assignment score for the control group and vice versa, and they become parallel. However, it is well recognized that it is hard to meet this assumption. One simple example is an RDD with an interaction term, under which the treatment effect is no longer constant. If it is difficult to assume one regression model for the entire range of data, it would be better to fit the treatment group data separately from the control group data. The fitted regression lines could be differently shaped and non-parallel when extended beyond the cutoff. However, the extrapolation assumption which is required in order to estimate the overall average treatment effect may not be tenable.

Because of this recognition, people resort to the original concept of RDD that RDD behaves like RCT in a narrow range of the cutoff. Hahn, Tood, and Van Der Klaauw (2001) proved that under a couple of weak conditions that many RDDs can meet, the treatment effect can be estimated at the cutoff. However, this approach has two problems. The first is that only so-called the marginal average treatment effect (MATE) can be estimated and one cannot say much about the treatment effect outside of the narrow range. Secondly, the sample size will be severely restricted and consequently the efficiency will suffer greatly, so only large studies can support such analysis. In this case, it would be advantageous to use the parametric approach because it would be much easier to meet the basic assumptions within the narrow range of the cutoff.

The first assumption does not allow misassignment of study subjects to the treatment but in reality this assumption is sometimes violated during assignment or implementation of the treatment. It is sometimes due to assignment errors but some other times it can be part of design or due to intentional misassignment. The latter happens especially when treatment is perceived as beneficial, so the assigner intentionally allows subjects just below the cutoff to receive treatment. This causes so called “fuzziness” in RDD, which results in reduction in efficiency. Therefore, if it is not part of design it is important to monitor the implementation of assignment. If it is not part of design fuzziness should be taken into account in analysis (see Bloom et al., 2005 and Schochet, 2008).

For a study with multiple treatments with multiple cutoffs, the sample can be pooled by combining multiple RDDs into one RDD model. However, if multiple RDD models are too different each other, pooling may be harmful because the shape of the pooled model may become very different from the shapes of multiple RDD models. Pooling basically assumes the similarity of multiple RDDs in terms of their shape.

It is advantageous to have good covariates, which have strong correlations with the outcome variables because they will increase the *R*-squared of the RDD regression model, which results in enhanced efficiency. This is more necessary for RDD than RCT because RDD requires much larger sample size.

In both RCT and RDD, sometimes treatment assignment is done at cluster level. For example, students are study subjects but the treatment group is formed by assigning schools to the treatment. In this case, the regression model should be a multi-level model or hierarchical linear model. For more discussion of the power and sample size determination for such designs, see Raudenbush, S. W. (1997); Bloom, Bos, and Lee (1999), and Schochet (2008).

To compensate for much lower efficiency for RDD compared to RCT, sometimes so called tie braking experiment is used within a narrow range of the cutoff, where subjects falling in the range are randomly assigned to the treatment or to control, and outside of the range the usual RDD is used. Cappelleri, Darlington, and Trochim (1994) studied the power of a significance test and the sample size for such designs.

Even with all these possible caveats, after reviewing three studies having both RCT and RDD components within, Cook and Wong (In press/under review) concluded that RDD results may be generally robust.

5. Conclusions

RDD is a viable option when treatment assignment is done by a score system, but it requires a much larger sample size than RCT. For this reason, gathering good covariates for RDD is more necessary than for RCT. Furthermore, RDD requires strong model assumptions but they are often violated in reality so that internal validity can be threatened. This requires careful checking of the assumptions and appropriate remedies to address such issues. Among those violations, model misspecification is more serious, so nonparametric techniques are favored by recent authors (Hahn et al, 1999; Imbens and Lemieux, 2008). However, for smaller studies, which cannot afford a large sample size, the parametric framework is still useful in the planning stage to do the power analysis and to determine the sample size. When using the parametric approach in analyzing RDD, slight over-fitting of the model may be prudent to avoid bias with some loss of efficiency in a manner of striking a balance between bias reduction and increase in variance (bias-variance trade-off).

Acknowledgements

We thank Frank Jenkins for his helpful comments.

References

Aiken, L.S., West, S.G., Schwalm, D.E., Caroll, J., and Hsuing, S. (1998). Comparison of a randomized and two=qausi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review*, 22, 207-244.

Bloom, H., Bos, J., and Lee, S. (1999). Using cluster random assignment to measure program impacts: statistical implications for evaluation of education programs. *Evaluation Review*, 23, 445-469.

Bloom, H.S., Kemple, J., Gamse, B., and Jacob, R. (2005). Using Regression Discontinuity Analysis to Measure the Impacts of Reading First. Paper presented at the annual conference of the American Educational Research Association held in Montreal, Canada.

Cappelleri, J., Darlington, R., and Trochim, W. (1994). Power analysis of cutoff-based randomized clinical trials. *Evaluation Review*, 18, 141-152.

Cohen, J. (1988). *Statistical Power Analysis for Behavioral Sciences*. Hillside, NJ: Lawrence Erlbaum.

Cook, T.D., and Wong, V.C. (In press/under review). Empirical tests of the validity of the regression discontinuity design. *Annales d'Economie et de Statistique*.

Goldberger, A.S. (1972). *Selection bias in evaluating treatment effects: Some formal illustrations* (Discussion Paper No. 123). Madison: University of Wisconsin, Institute for Research on Poverty.

Hahn, J., Todd, P., Van Der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression discontinuity design. *Econometric*, 69, 201-209.

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

Institute of Education Sciences (2004). Redaing comprehension and reading scale-up research grants request for applications. Washington, DC: Department of Education.

Raudenbush, S.W. (1997). Statistical Analysis and Optimal Design for Cluster Randomized Trials. *Psychological Methods*, 2, 173-185.

Rubin, D.B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2, 1-26.

Schochet, P.Z. (2008). Statistical power for regression discontinuity designs in educational evaluations. Report submitted to Institute of Education Sciences by Mathematica Policy Research, Inc.. Princeton, NJ.: Mathematica Policy Research, Inc.

Tallmadge, G.K., and Horst, D.P. (1976). *A procedural guide for validating achievement gains in educational projects*. Monograph series No. 2 on education evaluation. Washington, DC: U.S. Department of Health, Education, and Welfare.

Thistlewaite, D.L., and Campbell, D.T. (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 Design. Beverly Hills, CA: Sage Publications.