NCEE 2008-4026

U. S. DEPARTMENT OF EDUCATION

Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations

August 2008

Peter Z. Schochet
Mathematica Policy Research, Inc.

Abstract

This report examines theoretical and empirical issues related to the statistical power of impact estimates under clustered regression discontinuity (RD) designs. The theory is grounded in the causal inference and HLM modeling literature, and the empirical work focuses on commonly-used designs in education research to test intervention effects on student test scores. The main conclusion is that three to four times larger samples are typically required under RD than experimental clustered designs to produce impacts with the same level of statistical precision. Thus, the viability of using RD designs for new impact evaluations of educational interventions may be limited, and will depend on the point of treatment assignment, the availability of pretests, and key research questions.

NCEE 2008-4026
U.S. DEPARTMENT OF EDUCATION

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences under Contract ED-04-CO-0112/0006.

Disclaimer
The Institute of Education Sciences (IES) at the U.S. Department of Education contracted with Mathematica Policy Research, Inc. to develop methods on how to calculate statistical power for regression discontinuity designs in educational evaluations. The views expressed in this report are those of the author and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

August 2008

Schochet, Peter Z. (2008). *Technical Methods Report*: *Statistical Power for Regression Discontinuity Designs in Education Evaluations* (NCEE 2008-4026). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the IES website at http://ncee.ed.gov.

Alternate Formats
Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of Potential Conflicts of Interest

The author for this report, Dr. Peter Schochet, is an employee of Mathematica Policy Research, Inc. with whom IES contracted to develop the methods that are presented in this report. Dr. Schochet and other MPR staff do not have financial interests that could be affected by the content in this report.

Contents

Chapter 1: Introduction **1**

Chapter 2: Measuring Statistical Power **3**

Chapter 3: Considered Designs **5**

Chapter 4: Aggregated Designs: RD Design Theory and Design Effects **7**
Theoretical Underpinnings 7
Variance Calculations 11
The Design Effect for the RD Design 12
The RD Design Effect for the MDE Calculations 13
Including Additional Baseline Covariates 18
The Fuzzy RD Design 18

Chapter 5: Multilevel RD Designs **21**
Designs II and III 21
Designs IV, V, and VI 24

Chapter 6: Selecting the Score Range for the Sample **25**

Chapter 7: Illustrative Precision Calculations **27**
Presentation and Assumptions 27
Results 31

Chapter 8: Summary and Conclusions **33**

Appendix A **A-1**

Appendix B **B-1**

References **R-1**

List of Tables

Table 4.1: Formulas for ρ_{TS} Under RD Designs, by Score Distribution 14
Table 4.2: Design Effects for RD Designs, by Score Distribution, the Location of the Cutoff Score Value, and Other Key Parameters 16
Table 4.3: Design Effects for RD Designs for a 50-50 Split of the Treatment and Control Group Samples 17
Table 7.1: Required School Sample Sizes to Detect Target Effect Sizes, for Various RD Designs 28
Table 7.2: Required School Sample Sizes to Detect Target Effect Sizes, for Various RD Designs 29
Table 7.3: Required School Sample Sizes to Detect Target Effect Sizes, for Various Random Assignment (RA) Designs 30
Table A.1: Values for *Factor(.)* in Equation (1) of Text, by the Number of Degrees of Freedom, for One- and Two-Tailed Tests, and at 80 and 85 Percent Power A-1

List of Figures

Figure 4.1: The RD Method Visually 9
Figure 4.2: Graphs of Four Score Distributions 15

Chapter 1: Introduction

Regression discontinuity (RD) designs are increasingly being used by researchers to obtain unbiased impact estimates of education-related interventions. These designs are applicable when a continuous "scoring" rule is used to assign the intervention to study units (for example, school districts, schools, or students). Units with scores below a pre-set cutoff value are assigned to the treatment group and units with scores above the cutoff value are assigned to the comparison group, or vice versa. For example, Jacob and Lefgren (2004) examined the effects of attending summer school on the outcomes of New York City students using the rule that only students with standardized test scores below a cutoff value were required to attend summer school. As another example, the design for the National Evaluation of Early Reading First (ERF) (Jackson et al. 2007) was based on an independent reviewer scoring process where grantees with the highest application scores were awarded ERF grants to improve local preschools. As a final example, Ludwig and Miller (2007) exploited the variation in Head Start funding across counties to examine the program's effects on schooling and health. Cook (2008), Imbens and Lemieux (2008), and Shadish et al. (2002) provide reviews of the RD design.

Under well-designed RD designs, the treatment assignment rule is *fully* observed and can be modeled to yield unbiased impact estimates. A regression line (or curve) is fit in the outcome-score plane for the treatment group and similarly for the comparison group, and differences in the intercepts of these lines is the impact estimate. An impact occurs if there is a "discontinuity" in the two regression lines at the cutoff score. Because the selection rule is fully known under the RD design, selection bias issues tend to be less problematic under the RD design than under other non-experimental designs.

The literature suggests that the RD design might be a suitable alternative to a random assignment (RA) design when an experiment is not feasible (Cook 2008). RD designs tend to interfere less with normal program operations than RA designs, because treatment assignments for the study population are determined by rules developed by program staff or policymakers rather than randomly. Thus, treatments can be targeted to those who normally receive them (for evaluations of existing interventions) or to those who are deemed likely to benefit most from them (for evaluations of new interventions). Thus, RD designs may be easier to "sell" to program staff and participants, which could facilitate efforts to recruit study sites.

A major drawback of the RD design relative to the RA design, however, is that much larger sample sizes are typically required to achieve impact estimates with the same level of statistical power. If the score variable is normally distributed and centered on the cutoff, Goldberger (1972) demonstrated that for a *nonclustered* design, the sample under a RD design must be 2.75 times larger than for a corresponding experiment to achieve the same level of statistical precision. Cappelleri et al. (1994) extended this work to allow for a wider range of cutoff values. The reduction in precision in the RD design arises due to the substantial correlation, by construction, between the treatment status and score variables that are included in the regression models; this correlation is not present under the RA design.

This paper extends the work of Goldberger (1972) and Cappelleri et al. (1994) by addressing two main research questions: (1) What is the statistical power of RD designs under *clustered* (group-based) designs that are typically used in impact evaluations of education interventions, and (2) When are RD designs in a school setting feasible from a cost perspective?

The paper examines commonly-used clustered designs where groups (such as districts, schools, or classrooms) are assigned to a research status. Schochet (2008) and Bloom et al. (2005a) demonstrate that relatively large numbers of schools must be sampled under clustered RA designs (for example, about 60 if pretests are available) to yield impact estimates with adequate levels of precision. Because of additional

precision losses under RD designs, statistical power is critical for assessing whether RD designs can be a viable alternative to RA designs in the education field. Although there is a large literature on appropriate methods for analyzing data under RD designs (see, for example, Imbens and Lemieux 2008), much less attention has been paid to examining statistical power under RD designs.

This paper builds on the literature in several other ways. It examines statistical power under RD designs that is anchored in the causal inference and hierarchical linear modeling (HLM) literature. The paper also examines statistical power for a wider range of score distributions than have been explored previously, and for both sharp RD designs (where all units comply with their treatment assignments) and fuzzy RD designs (which allow for noncompliers). In addition, the paper discusses power implications of including additional baseline covariates in the regression models, and criteria for determining the appropriate range of scores for the study sample. Finally, the paper uses the theoretical formulas and empirically-based parameter assumptions to calculate appropriate sample sizes for alternative RD designs. These estimates can serve as a guide for future RD designs in the education field.

The empirical analysis focuses on achievement test scores of elementary school and preschool students in low-performing school districts. The focus is on test scores due to the accountability provisions of the No Child Left Behind Act of 2001, and the ensuing federal emphasis on testing interventions to improve reading and mathematics scores of young students.

The rest of this paper is in seven chapters. Chapter 2 discusses how to measure statistical power, and Chapter 3 discusses the considered clustered designs. In Chapter 4, assuming that student-level data are aggregated to the group level, I discuss the theory underlying the RD and RA designs, variance calculations, and RD design effects. In Chapter 5, the analysis is extended to multilevel models where the data are analyzed at the student level, and in Chapter 6, I briefly discuss the appropriate range of scores for the study sample. Chapter 7 discusses empirical results and Chapter 8 presents conclusions.

Chapter 2: Measuring Statistical Power

An important part of any evaluation design is the statistical power analysis, which demonstrates how well the design of the study will be able to distinguish real impacts from chance differences. To determine appropriate sample sizes for impact evaluations, researchers typically calculate minimum detectable impacts, which represent the smallest program impacts—average treatment and comparison group differences—that can be detected with a high probability. In addition, it is common to standardize minimum detectable impacts into *effect size units*—that is, as a percentage of the standard deviation of the outcome measures (also known as Cohen's *d*)—to facilitate the comparison of findings across outcomes that are measured on different scales (Cohen 1988). Hereafter, minimum detectable impacts in effect size units are denoted as "MDEs."

Mathematically, the MDE formula can be expressed as follows:

$$(1)\quad MDE = Factor(\alpha, \beta, df) * \sqrt{Var(impact)} / \sigma,$$

where *Var(impact)* is the variance of the impact estimate, σ is the standard deviation of the outcome measure, and *Factor*(.) is a constant that is a function of the significance level (α), statistical power (β), and the number of degrees of freedom.[1] *Factor(.)* becomes larger as α and *df* decrease and as β increases (see Table A.1).

As an example, consider an experimental design with a single treatment and control group and α=.05 and β=.80. In this case, for a given sample size and design structure, there is an 80 percent probability that a two-sample *t*-test will yield a statistically significant impact estimate at the 5 percent significance level if the true impact were equal to the MDE value in equation (1).

This approach for measuring statistical power differs slightly from the one used in Cappelleri et al. (1994) who apply Fisher's Z transformation to the partial correlation coefficient between the outcome measure and treatment status. This difference in metric accounts for the small differences between comparable results in this paper and those in Cappelleri et al. (1994).

[1] Specifically, *Factor(.)* can be expressed as $[T^{-1}(\alpha) + T^{-1}(\beta)]$ for a one-tailed test and $[T(\alpha/2) + T^{-1}(\beta)]$ for a two-tailed test, where $T^{-1}(.)$ is the inverse of the student's t distribution function with *df* degrees of freedom (see Murray 1998 and Bloom 2004 for derivations of these formulas). Equation (1) ignores the estimation error in the standard deviation.

Chapter 3: Considered Designs

The analysis presented below applies to commonly-used clustered designs in the education field where one of the following "units" is assigned to a single treatment or control group: school districts, schools, classrooms or students. In these designs, students are nested within higher-level units (groups).

Clustering in multilevel designs comes from two potential sources: (1) the *assignment* of units to a research condition, and (2) the random *sampling* of units from a broader universe of units before or after treatment assignments take place. This paper considers the following designs that combine these two sources of clustering (that are ordered based on design structure):

I. **Students are the unit of assignment and site (school or district) effects are fixed.** In some designs, students in purposively-selected schools or districts are randomly assigned directly to a research group. For example, in the Impact Evaluation of Charter School Strategies (Gleason and Olsen 2004), within each charter school area that volunteered for the study, students interested in attending a charter school were randomly assigned through a lottery to either a treatment group (who were allowed to enroll in a charter school) or a control group (who were not). Under these designs, sites can be treated as fixed strata if the impact results are to be viewed as pertaining to the study sites only. To estimate these models, impacts can be estimated using a pooled model where the covariates include treatment status and site indicators (and perhaps, site-by-treatment interactions); the error structure would include random student-level terms only.

II. **Classrooms are the unit of assignment and school effects are fixed.** Classroom-based designs are appropriate for interventions that are administered at the classroom level and where potential spillover effects of the intervention from treatment to control group classrooms are deemed to be small. If schools are purposively selected for the study, school effects could be treated as fixed strata. This design was used in the Evaluation of the Effectiveness of Educational Technology Interventions (Dynarski and Agodini 2003) where teachers in participating volunteer schools were randomly assigned to use a technology or not. In estimating these models, random classroom effects would be included in the model error structure, and school indicators (and perhaps, school-by-treatment interactions) would be included as model covariates.

III. **Schools are the unit of assignment and no random classroom effects.** School-based designs are common in the education field, and are often preferred over classroom-based designs because of concerns over potential spillover effects. These designs are also necessary for testing interventions that can affect the entire school (such as those that aim to change the school climate). The exclusion of random classroom effects can be justified if students are sampled from all targeted classrooms within the study schools. To estimate these models, random school effects would be included in the error structure in the HLM models. A variant of this design is if school districts are the unit of assignment and students are selected within districts without regard to their schools or classrooms.

IV. **Students are the unit of assignment and site effects are random.** This design is similar to Design I, except that sites are considered to be randomly sampled from a broader universe of sites, so that study results are to be viewed as generalizing outside the site sample (that is, as being externally valid). For estimation, random site and site-by-treatment interaction terms would be included in the error structure in the HLM models.

V. **Classrooms are the unit of assignment and school effects are random.** This design is a modification to Design II where school effects are treated as random. For estimation, the model error structure would include random classroom, school, and school-by-treatment interaction terms.

VI. **Schools are the unit of assignment and classroom effects are random**. This design is appropriate if classrooms within study schools are sampled for the study, or if all classrooms are included in the study but are considered to be sampled from a larger classroom population. For estimation, the model error structure would include random school and classroom effects.

These designs are discussed in more detail in Schochet (2008) in the context of RA designs using a unified HLM framework.

To simplify the presentation and fix concepts, I first discuss the theory underlying the RD design for Design I and Designs II and III where the analysis is conducted using data that are *averaged* to the unit of treatment assignment (classrooms, schools, or districts). These designs are referred to as "aggregated" designs. I then discuss "multilevel designs" that include Designs II and III where the analysis is conducted using student-level data and Designs IV to VI.

In what follows, the RD design is discussed in the context of the causal inference theory underlying RA designs (Neyman 1923, Rubin 1974, Holland 1986, Imbens and Rubin 2007, Schochet 2007). This framework is then used to discuss impact and variance estimation methods that are required to calculate MDEs.

Chapter 4: Aggregated Designs: RD Design Theory and Design Effects

Theoretical Underpinnings

This paper considers both RD and RA designs where n study units are assigned to either a single treatment or control condition (for simplicity, the comparison group under the RD design is hereafter referred to as the "control" group). The sample contains np treatment units and $n(1-p)$ control units, where p is the sampling rate to the treatment group ($0<p<1$).

Let Y_{Ti} be the "potential" outcome for unit i in the treatment condition and Y_{Ci} be the potential outcome for unit i in the control condition. Potential outcomes for the n study units are assumed to be random draws from potential treatment and control outcome distributions in the study population. The means of these distributions are denoted by μ_T for potential treatment outcomes and μ_C for potential control outcomes. It is assumed further that $Score_i$— the variable that is used to assign units to a research status under the RD design—is a random draw from the population score distribution with mean μ_S and variance σ_S^2. To consistently compare statistical power under the RD and RA designs, it is assumed that the score variable is also available for the RA design.[2]

The difference between the two potential outcomes, $(Y_{Ti} - Y_{Ci})$, is the unit-level treatment effect, and the average treatment effect parameter (*ATE*) under this "superpopulation" causal inference model is $ATE = E(\overline{Y}_T - \overline{Y}_C) = \mu_T - \mu_C$. The unit-level treatment effects, and hence, the *ATE* parameter, cannot be calculated directly because for each unit, the potential outcome is observed in either the treatment or control condition, but not in both. Formally, if T_i is a treatment status indicator variable that equals 1 for treatments and 0 for controls, then the *observed* outcome for a unit, y_i, can be expressed as follows:

$$(2) \quad y_i = T_i Y_{Ti} + (1 - T_i) Y_{Ci}.$$

The simple relation in (2) forms the basis for the theory underlying both the RA and RD designs.

In what follows, constant treatment effects are assumed within the population, which implies (1) the same variance, σ^2, for the random variables Y_{Ti} and Y_{Ci}, and (2) the same covariance, σ_{SY} (and associated correlation, ρ_{SY}) between $Score_i$ and Y_{Ti} and Y_{Ci}. These assumptions are consistent with ordinary least squares (OLS) methods that are typically used to estimate program impacts in education research, and are required to ensure that variances based on OLS methods are justified by the Neyman model of causal inference (Freedman 2008; Schochet 2007).

The RA and RD designs differ in the treatment assignment process. Under the RA design, treatment status, T_i^{RA}, is assigned randomly to study units, whereas under the RD design, treatment status, T_i^{RD}, is

[2] Neyman (1923) considered a finite population RA model where Y_{Ti} and Y_{Ci} are assumed to be fixed for the study population and where the only source of randomness is treatment status. This paper considers a "superpopulation" version of the model (see, for example, Schochet 2007). Note that a finite population version of the RD model would need to assume that $Score_i$ is random (for example, due to measurement error).

assigned depending on whether $Score_i$ is larger or smaller than a cutoff value K. This paper considers RD designs with the following treatment assignment rule:

$T_i^{RD} = 1$ if $Score_i \geq K$ and
$T_i^{RD} = 0$ otherwise.

All results apply, however, if, instead, the treatment were offered to those with scores less than K. For simplicity, the same cutoff value is assumed within and across study sites.

Next, the RA and RD designs are discussed in more detail. The RA design is discussed first because it provides the foundation for examining statistical power under the RD design.

The RA Design

Under the RA design, the difference in expected observed outcomes between treatments and controls can be calculated using (2) as follows:

$$(3) \quad E(y_i^{RA} \mid T_i^{RA} = 1) - E(y_i^{RA} \mid T_i^{RA} = 0) = E(Y_{Ti} \mid T_i^{RA} = 1) - E(Y_{Ci} \mid T_i^{RA} = 0) = \mu_T - \mu_C,$$

where the last equality holds because of random assignment. Accordingly, $(\overline{y}_T^{RA} - \overline{y}_C^{RA})$ is an unbiased estimator for the *ATE* parameter.

This simple differences-in-means *ATE* estimator can also be obtained by rearranging (2) and applying OLS methods to the following regression equation:

$$(4) \quad y_i^{RA} = \alpha_0 + \alpha_1 T_i^{RA} + u_i,$$

where $\alpha_0 = \mu_C$ and $\alpha_1 = (\mu_T - \mu_C)$. The error term $u_i = T_i^{RA}(Y_{Ti} - \mu_T) + (1 - T_i^{RA})(Y_{Ci} - \mu_C)$ has mean zero and variance σ^2 and is uncorrelated with T_i^{RA}.

Although not needed to produce unbiased estimates, $Score_i$ can be included as an "irrelevant" variable in the regression equation to improve the precision of the impact estimates. The true model is still (4), but the estimation model is now:

$$(5) \quad y_i^{RA} = \alpha_0 + \alpha_1 T_i^{RA} + \alpha_2 Score_i + e_i,$$

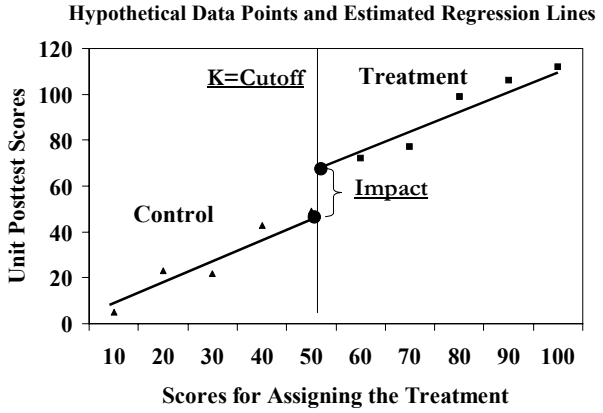
where e_i is an error term (conditional on $Score_i$) with variance σ_e^2. OLS methods yield consistent estimates of α_1 in (5) because T_i^{RA} and $Score_i$ are asymptotically uncorrelated due to random assignment (Schochet 2007; Yang and Tsiatis 2001). As discussed below, the model in (5) is used to compare the RA and RD designs.

The RD Design

Figure 4.1 displays graphically the theory underlying the RD design, where hypothetical posttest data (averaged to the unit level) are plotted against hypothetical treatment assignment scores (for example, pretest scores). The figure also displays fitted regression lines based on the observed data for treatments and controls, assuming constant treatment effects. The estimated impact under the RD design is the vertical difference between the two regression lines at the hypothetical score cutoff value of 50 (that is, at the point of discontinuity). The regression line for *potential* treatment group outcomes can be obtained by extending the regression line for the treatment group over the full score distribution, and similarly for potential control group outcomes. These extended regression lines pertain also to the fitted regression lines under the RA design, where units are randomly assigned across the entire score distribution.

Figure 4.1

The RD Method Visually



Hahn, Todd, and Van der Klaauw (2001) formally prove that if the conditional expectations $E(Y_{Ti} \mid Score_i = S)$ and $E(Y_{Ci} \mid Score_i = S)$ are continuous in S (as in Figure 4.1), the average causal effect of the treatment at the cutoff score K can be identified by comparing average observed outcomes immediately to the right and left of K:

$$\lim_{S \downarrow K} E(y_i^{RD} \mid Score_i = S) - \lim_{S \uparrow K} E(y_i^{RD} \mid Score_i = S).$$

Using (2), this average causal effect, ATE_K, can be expressed in terms of potential outcomes as follows:

$$(6) \quad ATE_K = E(Y_{Ti} \mid Score_i = K) - E(Y_{Ci} \mid Score_i = K).$$

Equation (6) suggests that impact estimates under the RD design generalize to a population that is typically narrower (units with scores right around the cutoff score value) than under the RA design (units with scores that cover the *full* score distribution). In our case, the ATE_K parameter equals the ATE parameter because of the constant treatment effects assumption, but this equality will not necessarily hold in general. The ATE_K parameter can also be interpreted as a marginal average treatment effect (MATE) parameter (Heckman and Vytlacil 2005), which addresses whether a marginal expansion of the program is warranted for units with scores just beyond the cutoff value.

The RD design has sometimes been compared to an experimental design for units with scores right around the cutoff (Campbell and Stanley 1963). Using this analogy, the ATE_K parameter in (6) can be estimated using simple differences-in-means procedures, where the sample includes *only* those units with scores right around *K*. In this case, statistical power considerations are similar for the RD and RA designs.

The RD-RA analogy is not exact, however, because under the RD design, chance alone may not fully determine which units are on either side of the cutoff. Furthermore, in many practical applications, there are not enough observations around the cutoff to obtain precise impact estimates. Thus, observations further from the cutoff are typically included in RD study samples (as in Figure 4.1). In these situations—which are the focus of this paper—treatment effects must be estimated using parametric or nonparametric methods where potential outcomes are modeled as a smooth function of the assignment scores. Unbiased impact estimates will result only if this outcome-score relationship is modeled correctly. Thus, unlike RA designs, RD designs hinge critically on the validity of key modeling assumptions.

For the analysis, it is assumed that the true functional form relationship between potential outcomes and scores is linear:

$$(7a) \quad E(Y_{Ti} \mid Score_i) = \alpha_0 + (\mu_T - \mu_C) + \alpha_2 Score_i$$

$$(7b) \quad E(Y_{Ci} \mid Score_i) = \alpha_0 + \alpha_2 Score_i .$$

The same slope coefficient, α_2, applies to both (7a) and (7b) because of the constant treatment effects assumption, and is the same coefficient as in (5) for the RA design. A linear specification is adopted, because this is a reasonable starting point for an analysis of data from RD designs, and simplifies the variance and power calculations. Furthermore, the linear specification is consistent with the local linear regression approach (Fan and Gijbels 1996) that has become increasingly popular in the literature for analyzing data under RD designs. It is also likely to approximately hold if the score is a pretest. The exact outcome-score relationship will depend on the specific design application, but the linearity (and constant treatment effects) assumptions will likely provide a lower bound on RD design effects.

Using (2), equations (7a) and (7b) yield the following regression model for the RD design:

$$(8) \quad y_i^{RD} = \alpha_0 + \alpha_1 T_i^{RD} + \alpha_2 Score_i + \eta_i ,$$

where η_i is a mean zero error term with variance σ_η^2. [3]

In Appendix B, it is proved (for the more general multilevel models) that the OLS estimator $\hat{\alpha}_1$ in (8) yields a consistent estimator of the ATE_K parameter (and *ATE* parameter in our case) assuming that the model is specified correctly.[4] Importantly, this result holds even if $Score_i$ is correlated with η_i (for

[3] It is often convenient to include ($Score_i$-*K*) in the model rather than $Score_i$ (especially if score-by-treatment interactions are included as covariates) so that α_1 always represents the treatment effect at the cutoff score. This scaling, however, has no effect on the results presented in this paper, and thus, the simpler specification in (8) is used.

[4] Rubin (1977) and Griliches and Ringstad (1971) provide proofs of this result for nonclustered designs in a slightly different context (see also Cappelleri et al. 1991).

example, due to measurement error in *Score*$_i$), because conditional on *Score*$_i$, T_i^{RD} and η_i are independent. Thus, although the estimates of α_0 and α_2 will be asymptotically biased if *Score*$_i$ is correlated with η_i, the estimator for α_1 will be asymptotically unbiased. A similar situation occurs under the RA design in (5).

Variance Calculations

As shown in Appendix B, the asymptotic OLS variance estimator for $\hat{\alpha}_1$ under the RD model in (8) is as follows:

$$(9) \quad AsyVar_{RD}(\hat{\alpha}_1) = \frac{\sigma_\eta^2}{np(1-p)(1-\rho_{TS}^2)} = \frac{\sigma_{y^{RD}}^2(1-R_{RD}^2)}{np(1-p)(1-\rho_{TS}^2)}.$$

In this expression, ρ_{TS} is the correlation between T_i^{RD} and *Score*$_i$, $\sigma_{y^{RD}}^2$ is the variance of the posttest measure, and R_{RD}^2 is the asymptotic regression R^2 value (which will depend on the strength of the outcome-score relationship and the size of the treatment effect).

The asymptotic variance of the impact estimate under the comparable RA design in (5)— with the same sample units and the same value for *p*—is as follows:

$$(10) \quad AsyVar_{RA}(\hat{\alpha}_1) = \frac{\sigma_e^2}{np(1-p)} = \frac{\sigma_{y^{RA}}^2(1-R_{RA}^2)}{np(1-p)} = \frac{\sigma^2(1-\rho_{SY}^2)}{np(1-p)},$$

where R_{RA}^2 is the asymptotic regression R^2 value under the RA design.

There are two key features of these variance formulas:

1. ***The term*** $[1/(1-\rho_{TS}^2)]$ ***enters the variance expression for the RD design but not for the RA design.*** This occurs because, by construction, treatment status and assignment scores are correlated in the RD regression model, but not in the RA model. As discussed further below, this correlation tends to be quite large in absolute value, which substantially increases the variance estimates under the RD design. Intuitively, the treatment effect in (8) is *net* of the score variable. Thus, the substantial collinearity between the treatment status and score variables reduces the information contained in the treatment status variable, which lowers the effective sample size for analysis.

2. ***The error variances*** σ_η^2 ***and*** σ_e^2 ***are identical.*** Specific values for the error terms e_i and η_i may differ depending on treatment assignments under the two designs. However, the variances of the error terms in (5) and (8) are the same, because the spread of the posttest values around the common fitted regression lines are the same for the two designs. This result implies that the variances of the posttests are likely to differ for the two designs because $\sigma_{y^{RD}}^2 = \sigma_{y^{RA}}^2 + 2\alpha_1\alpha_2\sigma_{TS}$ (assuming no correlation between the score variable and the model error terms). Thus, for example, if impacts are positive and the score and posttest variables are positively correlated (as in Figure 4.1), the variance of the posttest values will be larger under the RD than RA design. Thus, differences between R_{RD}^2 and R_{RA}^2 values directly compensate for differences in the outcome variances across the two designs.

The Design Effect for the RD Design

The design effect for the RD design relative to the RA design—as measured as the ratio of the asymptotic variances of the impact estimators in (9) and (10)—is as follows:

$$(11) \quad RD\ Design\ Effect = \frac{AsyVar_{RD}(\hat{\alpha}_1)}{AsyVar_{RA}(\hat{\alpha}_1)} = \frac{1}{(1-\rho_{TS}^2)}.$$

As the size of the squared correlation increases, the design effect increases. The design effect represents the increase in the sample size that is required under the RD design to produce impact estimates with the same level of statistical precision as the RA design.[5]

The design effect depends on (1) the distribution of the assignment scores in the study population, (2) the location of the cutoff score in this distribution, and (3) the treatment-control split in the sample. Importantly, the design effect does *not* depend on the total sample size, the size of the impact estimate (α_1), or the strength of the outcome-score relationship (α_2). During the planning stages of an evaluation, the RD design effect could be approximated if data are available on the likely score distribution.

To provide guidance on likely design effects, Table 4.1 displays formulas for calculating ρ_{TS} for the following four score distributions that are likely to occur in practice:

1. ***Normal distribution,*** which was examined by Goldberger (1972) and Cappelleri et al. (1994).
2. ***Uniform distribution,*** where scores are equally prevalent across the score range.
3. ***Truncated normal distribution,*** which is relevant if the entire score distribution is normally distributed, but if the sample is limited to units with scores within a specified bandwidth around the cutoff score.
4. ***Bimodal distribution,*** which is calculated as a mixture (simple average) of two normal distributions with different means (that are equidistant from the full bimodal distribution mean) but the same variance. If the two means are sufficiently spread out, this symmetric distribution will have two peaks, where each peak is centered around the mean of one component normal distribution. This distribution would arise if there are clusters of high and low scores, with fewer scores in the middle of the distribution.

Figure 4.2 displays graphs of each probability distribution function (pdf) as well as key distribution parameters.

All formulas in Table 4.1 are parameterized by *p*—the percentage of the sample that is assigned to the treatment group—which is also assumed to equal *q*—the percentage of the score distribution that lies to the right of the cutoff score. The truncated normal and bimodal distributions are also functions of several additional parameters (see Table 4.1 and Figure 4.2). Formulas involving normal distributions are

[5] The design effect in (11) is slightly different than that developed in Cappelleri et al. (1994) who, as discussed, used a different metric for measuring statistical power, and compared all RD designs to a RA design with a 50-50 treatment-control split.

functions of inverse standard normal distributions, which can be calculated using standard statistical packages such as SAS.

Table 4.2 displays design effects using the formulas in Table 4.1 for various parameter values. Table 4.3 displays comparable values under a common design where the treatment and control samples are of *equal* size regardless of the cutoff location. The figures in Table 4.3 were obtained using simulations, because the implied subsampling needed to obtain the balanced research samples yields complex score distributions, making it difficult to find closed-form solutions for ρ_{TS}.[6]

The key finding is that RD design effects tend to be large. Design effects vary somewhat depending on the treatment-control sample split (Table 4.3), but for a given sample allocation, they do not vary much across score distributions or cutoff values (Table 4.2). For p=0.50, design effects range from about 2.75 to 5 (Table 4.3). Design effects do not materially decrease unless the treatment-control sample split is highly unbalanced (Table 4.2). It is interesting that design effects tend to be largest for p=0.50, even though for a given sample size, this allocation yields the most precise impact estimates under the RA design.

For the truncated normal distribution, there is a complex relationship between the design effect and score bandwidth. For instance, if p=q=0.50, the design effect increases as the bandwidth becomes narrower, but this does not necessarily hold for other p-q configurations (Tables 4.2 and 4.3). As discussed below, these findings have implications for assessing the appropriate bandwidth for selecting the sample. Finally, for the bimodal distribution, the design effect tends to increase as the means of the two component normal distributions become further apart.

The RD Design Effect for the MDE Calculations

To calculate MDEs, standard errors of the impact estimates must be divided by standard deviations of the outcome measures (see equation [1]). What standard deviations should be used in the MDE calculations for RD designs?

In principle, impact estimates under RD designs pertain to only those units with scores right around the cutoff value. Thus, one option would be to use standard deviations for these units only. These standard deviations, however, are likely to be much smaller than the full-population values that are used for RA designs. Thus, I do not adopt this approach, because it would likely lead to serious (and somewhat artificial) increases in MDEs for RD designs relative to RA designs.

A second option would be to use standard deviations based on the models in (5) and (8). These two standard deviations are likely to differ because, as discussed, $\sigma^2_{y^{RD}} \approx \sigma^2_{y^{RA}} + 2\alpha_1\alpha_2\sigma_{TS}$. However, because these differences are a function of the unknown parameters α_1 and α_2, they would be difficult to compute without further assumptions.

Instead, I assume the *same* standard deviation for both the RD and RA designs that pertain to the study "superpopulation," even if this population is not delineated precisely. Under this approach, the square root of the RD design effect in (11) for the variance calculations applies to the MDE calculations.

[6] The simulations were conducted by (1) obtaining 100,000 random draws from the full score distribution, (2) defining treatments and controls based on the pertinent cutoff value, (3) subsampling from the larger research group to generate a 50-50 sample split; and (4) calculating the empirical correlation coefficient between the treatment status and score variables. I conducted 5,000 simulations and report average simulated correlation coefficients.

Table 4.1: Formulas for ρ_{TS} Under RD Designs, by Score Distribution

Score Distribution	Formulas for ρ_{TS}	Parameter Definitions
Normal	$\frac{\phi(\Phi^{-1}(1-p))}{\sqrt{p(1-p)}}$	$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$; $\Phi^{-1}(.)$ is the inverse of the standard normal distribution
Uniform	$\sqrt{3p(1-p)}$	
Truncated Normal	$\frac{p}{\sigma_S\sqrt{p(1-p)}}\left[\left\{\frac{\phi(k_2)-\phi(k_1)}{\Phi(k_2)-\Phi(k_1)}\right\}-\left\{\frac{\phi(k_2)-\phi(c)}{\Phi(k_2)-\Phi(c)}\right\}\right]$; $c = \Phi^{-1}[p\Phi(k_1)+(1-p)\Phi(k_2)]$; $\sigma_S^2 = \left[1-\left\{\frac{k_2\phi(k_2)-k_1\phi(k_1)}{\Phi(k_2)-\Phi(k_1)}\right\}-\left\{\frac{\phi(k_2)-\phi(k_1)}{\Phi(k_2)-\Phi(k_1)}\right\}^2\right]$	k_1 and k_2 are the number of standard deviations from the mean of the full normal distribution that the left and right truncation points fall (see Figure 4.2).
Symmetric Bimodal Distribution: Mixture of Two Normal Distributions With Different Means but the Same Variances	$\frac{1}{\sigma_S\sqrt{p(1-p)}}\{w[\phi(d+l)-l(1-\Phi(d+l))] + (1-w)[\phi(d-l)+l(1-\Phi(d-l))]\}$; $\sigma_S^2 = 1+4w(1-w)l^2$; *d is obtained by solving*: $p = 1-w\Phi(d+l)-(1-w)\Phi(d-l)$	w is the weight assigned to the first normal distribution (.5 in our case); l is the number of standard deviations to the right (left) of the mean of the overall bimodal distribution where the first (second) component normal distribution is centered (see Figure 4.2); d is the location of the cutoff score in the bimodal distribution.

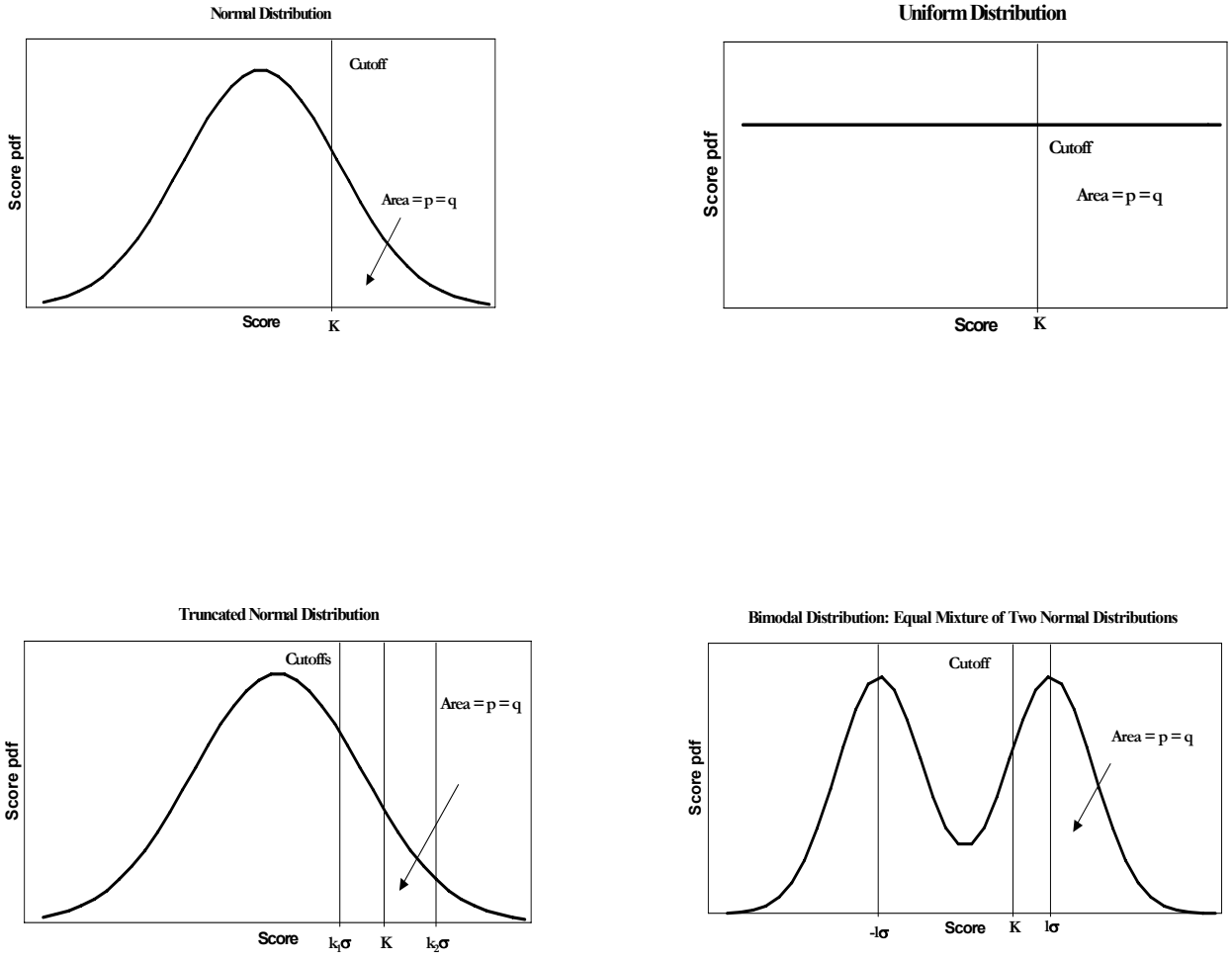
Notes: The parameter p is the treatment group sampling rate and equals q—the percentage of the score distribution that lies to the right of the cutoff score K. The formulas for ρ_{TS} were calculated using the following relations:

$$\rho_{TS} = \frac{\sigma_{TS}}{\sqrt{p(1-p)}\sigma_S} = \frac{E(T_i^{RD} Score_i) - p\mu_S}{\sqrt{p(1-p)}\sigma_S} = \frac{p\left[E(Score_i \mid Score_i \geq K) - \mu_S\right]}{\sqrt{p(1-p)}\sigma_S}.$$

The moments for truncated normal random variables were obtained using results in Maddala (1983).

Figure 4.2

Graphs of Four Score Distributions



Note: A zero mean is assumed for the overall bimodal distribution and for the full normal distribution underlying the truncated normal distribution.

Table 4.2: Design Effects for RD Designs, by Score Distribution, the Location of the Cutoff Score Value, and Other Key Parameters

	The Treatment Group Sampling Rate (p) Which Equals the Percentage of the Score Distribution That Lies Above the Cutoff Score (q)			
Score Distribution	**0.50**	**0.33 or 0.67**	**0.25 or 0.75**	**0.10 or 0.90**
Normal	2.75	2.46	2.17	1.52
Uniform	4.00	2.97	2.29	1.35
Truncated Normal				
Values of k_1 , k_2[a]				
-1.0, 1.0	3.65	2.91	2.30	1.41
-0.4, 0.4	3.94	2.99	2.29	1.38
-0.2, 0.2	3.98	3.00	2.29	1.37
0.0, 2.0	3.14	3.46	3.00	1.72
0.6, 1.4	3.73	3.45	2.70	1.50
0.8, 1.2	3.92	3.26	2.49	1.43
Bimodal: Equal Mixture of Two Normal Distributions				
Value of l[b]				
2.0	5.37	2.84	2.09	1.35
1.4	3.75	2.79	2.20	1.42
0.8	2.94	2.57	2.20	1.50
0.4	2.77	2.48	2.17	1.52

Note: See the text for formulas and other assumptions underlying the calculations.

[a]The parameters k_1 and k_2 are the number of standard deviations from the mean of the full normal distribution that the left and right truncation points fall.

[b]The parameter l is the number of standard deviations to the right (left) of the mean of the overall bimodal distribution where the first (second) component normal distribution is centered. The calculations assume that each normal distribution is weighted equally to create the bimodal mixture distribution (that is, the parameter w=0.5).

Table 4.3: Design Effects for RD Designs for a 50-50 Split of the Treatment and Control Group Samples

	Percentage of the Score Distribution That Lies Above the Cutoff Score Value (q) for p=0.50			
Score Distribution	**0.50**	**0.33 or 0.67**	**0.25 or 0.75**	**0.10 or 0.90**
Normal	2.75	2.79	2.85	3.16
Uniform	4.00	3.70	3.40	2.83
Truncated Normal				
Values of k_1 , k_2[a]				
-1.0, 1.0	3.65	3.50	3.36	2.99
-0.4, 0.4	3.94	3.65	3.40	2.86
-0.2, 0.2	3.98	3.67	3.40	2.84
0.0, 2.0	3.14	3.66	3.90	4.22
0.6, 1.4	3.73	4.01	3.93	3.44
0.8, 1.2	3.92	3.90	3.70	3.11
Bimodal: Equal Mixture of Two Normal Distributions				
Value of l[b]				
2.0	5.37	3.50	3.03	2.65
1.4	3.75	3.33	3.11	2.93
0.8	2.94	2.94	2.97	3.14
0.4	2.77	2.81	2.87	3.16

Note: See the text for formulas and other assumptions underlying the calculations.

[a]The parameters k_1 and k_2 are the number of standard deviations from the mean of the full normal distribution that the left and right truncation points fall.

[b]The parameter l is the number of standard deviations to the right (left) of the mean of the overall bimodal distribution where the first (second) component normal distribution is centered. The calculations assume that each normal distribution is weighted equally to create the bimodal mixture distribution (that is, the parameter w=0.5).

Including Additional Baseline Covariates

The inclusion in the RD models of additional covariates—measured at baseline—can improve the precision of the impact estimates. Similar to experimental designs, covariates can increase power by explaining some of the variance in the outcome measures across units (that is, by increasing regression R^2 values). The use of covariates (such as pretests) is especially important for improving precision in group-based designs where statistical power is often a major concern (Bloom et al. 1999; Schochet 2008).

Conditional on the assignment scores, the covariates will be asymptotically uncorrelated with treatment status if (1) the outcome-score relationship is modeled correctly, and (2) the covariates are a continuous function of the scores (at the cutoff value). Thus, under a well-designed RD study, the inclusion of additional covariates in the RD model should have little effect on the impact estimates (and if they do, model specification error may be present). The situation is analogous to the use of covariates in experimental designs which are asymptotically uncorrelated with treatment status due to random assignment.

When additional covariates are included in the RD model, the asymptotic OLS variance estimator for $\hat{\alpha}_1$ can be expressed as follows:

$$(12) \quad AsyVar_{RD}(\hat{\alpha}_1) = \frac{\sigma^2_{y^{RD}}(1 - R^2_{RD\_X})}{np(1-p)(1-\rho^2_{TS})} ,$$

where $R^2_{RD\_X}$ is the asymptotic regression R^2 value when y_i^{RD} is regressed on T_i^{RD}, $Score_i$, and the vector of covariates X_i (which could include strata indicator variables). The analogous variance expression for the RA design is:

$$(13) \quad AsyVar_{RA}(\hat{\alpha}_1) = \frac{\sigma^2_{y^{RA}}(1 - R^2_{RA\_X})}{np(1-p)} .$$

Importantly, the numerators in (12) and (13) are the same. Thus, the RD design effect in (11) applies also when additional covariates are included in the estimation models.

The Fuzzy RD Design

Thus far, it has been implicitly assumed that all sample units comply with their treatment assignments, that is, that all treatments and no controls receive intervention services. Under this "sharp" RD design, the probability of receiving the treatment changes from zero to one at the cutoff value.

The "fuzzy" RD design (Trochim 1984; Hahn et al. 2001) allows for noncompliers— treatment group nonparticipants and control group crossovers. Under this design, the jump in the probability of receiving the treatment at the cutoff is less than one. As an example, Van der Klaauw (2002) examined the effects of financial aid offers on college attendance, where the "score" variable was based on the applicant's SAT scores and grades, and cutoffs were based on rules used by colleges to award aid. Applicants in higher scoring groups were more likely to receive financial aid offers than applicants in lower scoring groups. However, some higher scoring applicants did not receive financial aid offers (treatment group nonparticipants) and some lower scoring applicants did receive offers (crossovers). Thus, this is a fuzzy RD design, because application information (such as essays and extracurricular activities) that was not measured in the score also played a role in financial aid award decisions.

Under the fuzzy RD design, modifications to the estimation methods discussed above are necessary to obtain impacts that adjust for noncompliers. This situation is analogous to the distinction between intention-to-treat (ITT) and treatment-on-the-treated (TOT) estimators under RA designs (Angrist et al. 1996).

The modifications can be understood by first classifying units at the cutoff score into four mutually exclusive compliance categories: compliers, never-takers, always-takers, and defiers (Angrist et al. 1996). Let R_{Ti} denote the "potential" service receipt indicator variable in the treatment condition, and R_{Ci} denote the potential service receipt indicator variable in the control condition. *Compliers (CL)* are those would receive intervention services only if they were assigned to the treatment group (R_{Ti}=1 and R_{Ci}=0). *Never-takers (N)* are those who would never receive treatment services (R_{Ti}=0 and R_{Ci}=0) and *always-takers* are those would always receive treatment services (R_{Ti}=1 and R_{Ci}=1). Finally, *defiers* are those who would receive the treatment only in the control condition (R_{Ti}=0 and R_{Ci}=1).

The ATE_K parameter for the pooled sample can be expressed as a weighted average of the ATE_K parameters for each of the unobserved compliance groups:

$$ATE_K = p_{CL} ATE_{K\_CL} + p_N ATE_{K\_N} + p_A ATE_{K\_A} + p_D ATE_{K\_D},$$

where p_g is the percentage of the study population in compliance group g ($\sum p_g = 1$), and ATE_{K_g} is the associated impact parameter. The ATE_{K_CL} parameter under the fuzzy RD design can then be identified under two key assumptions (Hahn et al. 2001; Imbens and Lemieux 2008). The first is that there are no defiers—the monotonicity assumption. This implies that p_D=0 and $p_{CL} = (p_T - p_C)$, where p_T is the treatment group participation rate (service receipt rate) and p_C is the control group crossover rate. The second key assumption is that the distributions of potential outcomes are independent of treatment assignments for the never-takers and always-takers—the exclusion restriction. This assumption implies that never-takers and always-takers receive identical services regardless of the treatment condition to which they are assigned. This restriction implies that $ATE_{K\_N} = ATE_{K\_A} = 0$.

Under these two assumptions, the following impact parameter can be identified under the fuzzy RD design:

$$(14) \quad ATE_{K\_CL} = \frac{E(Y_{Ti} \mid Score_i = K) - E(Y_{Ci} \mid Score_i = K)}{P(R_{Ti} = 1 \mid Score_i = K) - P(R_{Ci} = 1 \mid Score_i = K)} = \frac{ATE_K}{(p_T - p_C)}.$$

This parameter represents the average causal effect of the treatment for compliers at the cutoff score.

A consistent estimator for the ATE_{K_CL} parameter can be obtained by dividing consistent estimators for the numerator and denominator in (14), which can both be obtained using RD regression methods. The ATE_K parameter can be estimated using equation (8). An estimator for $(p_T - p_C)$ can be obtained as the parameter estimate on T_i^{RD} from a regression of observed treatment receipt status, r_i, on T_i^{RD} and a smooth function of $Score_i$.

This ATE_{K_CL} ratio estimator can also be obtained using instrumental variables (IV) techniques (Hahn et al. 2001) that are similar to the methods developed by Bloom (1984) and Angrist et al.(1996) to adjust for noncompliers in RA evaluations. For example, consider the following two-stage-least-squares procedure:

(1) Calculate predicted values, $\hat{r}_i$, from the RD regression model discussed above for estimating $(p_T - p_C)$; (2) Estimate equation (8) using $\hat{r}_i$ in place of T_i^{RD}. The second-stage coefficient estimate on $\hat{r}_i$ yields the ATE_{K_CL} estimator.

In principle, the variance of this ATE_{K_CL} estimator must account for the estimation error in both the ATE_K and $(p_T - p_C)$ parameters.[7] As an approximation, however, I treat $(p_T - p_C)$ as fixed and use equation (12) to obtain the following asymptotic variance expression for the ATE_{K_CL} estimator:

$$(15) \quad AsyVar_{RD}(ATE_{K\_CL}) = \frac{\sigma^2_{y^{RD}}(1 - R^2_{RD\_X})}{np(1-p)(1-\rho^2_{TS})(p_T - p_C)^2}.$$

The corresponding variance expression for the TOT estimator under the RA design is:

$$(16) \quad AsyVar_{RA}(TOT) = \frac{\sigma^2(1 - R^2_{RA\_X})}{np(1-p)(q_T - q_C)^2},$$

where (q_T - q_C) represents the treatment-control difference in service receipt rates for the *full* study population (not just those at the cutoff).

Consequently, the design effect for the fuzzy RD design is:

$$(17) \quad RD\ Fuzzy\ Design\ Effect = \frac{(q_T - q_C)^2}{(1-\rho^2_{TS})(p_T - p_C)^2}.$$

This design effect reduces to (11) if service receipt rates for units right around the cutoff mimic service receipt rates for units across the full score distribution.

[7] Using a Taylor series expansion, the variance of the ratio estimator is:

$\frac{Var(\hat{\alpha}_1)}{d^2} + \frac{\alpha_1^2 Var(\hat{d})}{d^4} - \frac{2\alpha_1 Cov(\hat{\alpha}_1, \hat{d})}{d^3}$, where $d = (p_T\text{-}p_C)$ and α_1 is the ATE_K parameter.

Chapter 5: Multilevel RD Designs

In this section, results from above are generalized to multilevel designs where data are analyzed at the student rather than group level. Designs II and III are discussed first, followed by a discussion of Designs IV to VI.

Designs II and III

The causal inference theory discussed above can be extended to the two-level design where students are nested within units that are assigned to a research status. As before, let Y_{Ti} and Y_{Ci} be unit-level potential outcomes and $Score_i$ be unit-level assignment scores, whose joint distributions are defined as in Chapter 4.[8] The sample contains np treatment units and $n(1-p)$ control units.

Suppose that m students are sampled from the student superpopulation within each study unit. Let W_{Tij} be the potential outcome for student j in unit i in the treatment condition and W_{Cij} be the corresponding potential outcome for the student in the control condition. It is assumed that W_{Tij} and W_{Cij} are random draws from student-level potential treatment and control outcome distributions (that are conditional on school-level potential outcomes) with means Y_{Ti} and Y_{Ci}, respectively, and common variance σ_θ^2.

In what follows, the two-level RA and RD designs are discussed using this causal inference framework.

The RA Design

Under the RA design, the observed outcome for a student, w_{ij}^{RA}, can be expressed as follows:

$$(18) \quad w_{ij}^{RA} = T_i^{RA} W_{Tij} + (1 - T_i^{RA}) W_{Cij} .$$

As before, terms in (18) can be rearranged to create the following regression model:

$$(19) \quad w_{ij}^{RA} = \alpha_0 + \alpha_1 T_i^{RA} + (\lambda_i + \theta_{ij}) ,$$

where:

1. α_0 and α_1 (the ATE parameter) are defined as above
2. $\lambda_i = T_i^{RA}(Y_{Ti} - \mu_T) + (1 - T_i^{RA})(Y_{Ci} - \mu_C)$ is a unit-level error term with mean zero and between-unit variance σ_λ^2 that is uncorrelated with T_i^{RA}
3. $\theta_{ij} = T_i^{RA}(W_{Tij} - Y_{Ti}) + (1 - T_i^{RA})(W_{Cij} - Y_{Ci})$ is a student-level error term with mean zero and within-unit variance σ_θ^2 that is uncorrelated with λ_i and T_i^{RA}

[8] For illustration simplicity, common symbols and subscripts are used for each two-level design. This convention is followed for the remainder of this section.

Importantly, (19) can also be derived using the following two-level hierarchical linear (HLM) model (Bryk and Raudenbush 1992):

$$Level\,1:\ w_{ij}^{RA} = y_i^{RA} + \theta_{ij}$$
$$Level\,2: y_i^{RA} = \alpha_0 + \alpha_1 T_i^{RA} + \lambda_i,$$

where Level 1 corresponds to students and Level 2 corresponds to units. Inserting the Level 2 equation into the Level 1 equation yields (19). Thus, the HLM approach is consistent with the causal inference theory presented above.

Suppose that $Score_i$ and other unit- and student-level baseline covariates are included in (19) as covariates. In this case, the asymptotic variance of the two-level (*TL*) OLS estimator for the *ATE* parameter is as follows:

$$(20)\quad AsyVar_{RA}(\hat{\alpha}_1^{TL}) = \frac{1}{p(1-p)}\left[\frac{\sigma_\lambda^2(1-R_{RA\_X\_B}^2)}{n} + \frac{\sigma_\theta^2(1-R_{RA\_X\_W}^2)}{nm}\right],$$

where $R_{RA\_X\_B}^2$ is the asymptotic regression R^2 value for the between-variance component and $R_{RA\_X\_W}^2$ is the asymptotic regression R^2 value for the within-variance component. These two R^2 values could differ depending on the nature of the covariates.

The within-school variance term in (20) is the conventional variance expression for an impact estimate under a nonclustered design. Design effects in a clustered design arise because of the first variance term, which represents the correlation of the outcomes of students within the same units (Murray 1998; Donner and Klar 2000; Raudenbush 1997). Design effects can be large because the divisor in the between-unit term is the number of units rather than the number of students.

It is common to express the variance expression in (20) in terms of the *intraclass correlation* (*ICC*) (Cochran 1963; Kish 1965), which is defined as the between-unit variance (σ_λ^2) as a proportion of the total variance of the outcome measure ($\sigma^2 = \sigma_\lambda^2 + \sigma_\theta^2$):

$$(21)\quad AsyVar_{RA}(\hat{\alpha}_1^{TL}) = \frac{1}{p(1-p)}\left[\frac{\sigma^2 ICC(1-R_{RA\_B}^2)}{n} + \frac{\sigma^2(1-ICC)(1-R_{RA\_W}^2)}{nm}\right].$$

In this formulation, design effects from clustering are small if the mean of the outcome measure does not vary much across units (that is, if *ICC* is small). In this case, the approach discussed above where student-level data are averaged to the unit level will provide consistent, but inefficient impact estimates. On the other hand, if the *ICC* is large (that is, close to 1), then using unit averages or individual student-level data will produce impact estimates with similar levels of precision. Specific *ICC* (and R^2) values will depend on the design.

The RD Design

Results from Chapter 4 for the RD design can also be extended to the two-level model. Let the observed outcome for a student, w_{ij}^{RD}, be expressed as follows:

$$(22) \quad w_{ij}^{RD} = T_i^{RD} W_{Tij} + (1 - T_i^{RD}) W_{Cij}$$
$$= T_i^{RD} Y_{Ti} + (1 - T_i^{RD}) Y_{Ci} + \left[T_i^{RD} (W_{Tij} - Y_{Ti}) + (1 - T_i^{RD})(W_{Cij} - Y_{Ci}) \right],$$

where the term inside the brackets is a mean zero residual term. If Y_{Ti} and Y_{Ci} are modeled as a linear function of the assignment scores (as in [7a] and [7b]), (22) yields the following two-level RD regression model:

$$(23) \quad w_{ij}^{RD} = \alpha_0 + \alpha_1 T_i^{RD} + \alpha_2 Score_i + (\tau_i + \delta_{ij}),$$

where τ_i is a mean zero unit-level error term with variance σ_τ^2, and δ_{ij} is a mean zero student-level error term with variance σ_δ^2 that is uncorrelated with τ_i. The parameter α_1 is the same ATE_K parameter as in (8) above.

As with the RA design, (23) can be obtained using a two-level HLM model:

$$Level\ 1:\ w_{ij}^{RD} = y_i^{RD} + \delta_{ij}$$
$$Level\ 2:\ y_i^{RD} = \alpha_0 + \alpha_1 T_i^{RD} + \alpha_2 Score_i + \tau_i.$$

Inserting the Level 2 equation into the Level 1 equation yields (23).

In Appendix B, it is proved that the two-level OLS estimator $\hat{\alpha}_1$ in (23) yields a consistent estimator of the ATE_K parameter assuming that the model is specified correctly. This result holds even if $Score_i$ is correlated with τ_i. Assuming that additional baseline covariates are included in the model, the asymptotic variance of this estimator is as follows (see Appendix B):

$$(24) \quad AsyVar_{RD}(\hat{\alpha}_1^{TL}) = \frac{1}{p(1-p)(1-\rho_{TS}^2)} \left[\frac{\sigma_\tau^2 (1 - R_{RD\_X\_B}^2)}{n} + \frac{\sigma_\delta^2 (1 - R_{RD\_X\_W}^2)}{nm} \right],$$

where $R_{RD\_X\_B}^2$ and $R_{RD\_X\_W}^2$ are between- and within-unit asymptotic regression R^2 values, respectively.

The RD Design Effect

A key finding is that the RD design effect *remains* at $1/(1-\rho_{TS}^2)$ under the two-level design. This is because the variances inside the brackets in (24) for the RD design equal the variances inside the brackets in (20) or (21) for the RA design. Thus, relative to the aggregated model presented above, the use of the two-level model for Designs II and III will typically improve the precision of the impact estimates for both the RD and RA designs. However, the proportional improvement in precision is the *same* for each

design, so that the RD design effect does not change. Similar results about the RD design effect apply also for the fuzzy RD design and for calculating MDEs.

Designs IV, V, and VI

Theoretical results from the models discussed above carry over directly to Design VI, where schools are the unit of assignment and classroom effects are treated as random. Ignoring R^2 terms for simplicity, the variance expression for the RD impact estimator for this design is:

$$AsyVar(\hat{\alpha}_1)\ for\ Design\ VI = \frac{1}{p(1-p)(1-\rho_{TS}^2)}\left[\frac{\sigma^2 ICC_1}{n} + \frac{\sigma^2 ICC_2}{nc} + \frac{\sigma^2(1-ICC_1-ICC_2)}{ncm}\right],$$

where ICC_1 is the intraclass correlation at the school level, ICC_2 is the intraclass correlation at the classroom level, n is the number of schools, and c is the number of study classrooms per school. This expression is a product of the variance expression for the RA design and the RD design effect $1/(1-\rho_{TS}^2)$.

The situation is somewhat different for Design IV where the unit of assignment is at the student level and site effects are treated as random. For this design, it is assumed that treatment effects are constant within sites, but *not* across sites. Instead, site-level treatment effects are assumed to be drawn from a population distribution with variance σ_I^2.

In this case, the design effect $1/(1-\rho_{TS}^2)$ affects the student-level variance term, but *not* the site-level variance term. Ignoring R^2 terms, the variance expression for the RD impact estimate under Design IV is:

$$AsyVar(\hat{\alpha}_1)\ for\ Design\ IV = \frac{\sigma_I^2}{n} + \frac{\sigma_\delta^2}{nmp(1-p)(1-\rho_{TS}^2)}.$$

Thus, the RD design effect is smaller for Design IV than for the other designs considered above.

A similar situation occurs under Design V, where the variance expression for the RD impact estimate is:

$$AsyVar(\hat{\alpha}_1)\ for\ Design\ V = \frac{\sigma_I^2}{n} + \frac{1}{p(1-p)(1-\rho_{TS}^2)}\left[\frac{\sigma^2 ICC_2}{nc} + \frac{\sigma^2(1-ICC_1-ICC_2)}{ncm}\right].$$

Chapter 6: Selecting the Score Range for the Sample

A central issue for designing RD studies is assessing the appropriate range of scores for selecting the sample (that is, the score bandwidth around the cutoff score). In some evaluations where the sample universe is small, a large proportion of study-eligible units (with a wide range of scores) must be included in the sample for power reasons, and the estimation of impacts must rely heavily on modeling assumptions. This was the case, for example, in the Early Reading First (Jackson et al. 2007) and Reading First (Bloom et al. 2005b) evaluations, where the sample universe consisted of a relatively small number of grantees who applied for program funds. In other designs, however, there may be many available potential study units, but only a subsample can be included in the study for cost reasons. In these cases, how should study units be sampled?

The key advantage of selecting a narrow bandwidth around the cutoff score is that this approach will likely yield impact estimates with little bias, because the correct posttest-score relationship can usually be specified (and is likely to be approximately linear). Increasing the bandwidth could increase bias if the posttest-score relationship varies across different regions of the score distribution, thereby making the modeling more difficult.

There are, however, three main disadvantages of using a narrower versus wider bandwidth. First, for a given sample size, a narrower bandwidth could yield less precise impact estimates if the outcome-score relationship can be correctly modeled using a wider range of scores. For instance, as discussed above, if scores have a truncated normal distribution and p=0.50, the RD design effect tends to decrease as the bandwidth increases (although this pattern does not generally hold). Second, in instances where there is a limited sample around the cutoff score, widening the bandwidth could yield larger samples, thereby increasing statistical precision.

A third disadvantage of using a narrow bandwidth is that the study will have less basis for extrapolating impact findings to units with scores further away from the cutoff. Theoretically, impact findings from the RD design generalize only to units with scores near the cutoff value. However, the estimated parametric regression lines for the treatment and control groups could be extended to obtain impact estimates for units over a wider score range (see Figure 4.1 above). These extrapolations are likely to be more defensible if the bandwidth is wider rather than narrower (that is, if the sample contains units that cover a broad range of scores).

The choice of the appropriate bandwidth could involve a variance versus bias tradeoff. Methods have been developed for assessing the optimal score bandwidth *after* data have been collected. For example, Ludwig and Miller (2007) propose a cross-validation criterion which selects the bandwidth to minimize the average squared distance between actual outcome values and predicted values from the fitted regression lines. A variant of this approach is to estimate weighted regressions where kernel functions are used to assign larger weights to data points closer to the cutoff than to those further from the cutoff (Porter 2003).

These same approaches could be used to select the appropriate bandwidth in the design phase of RD studies if pertinent secondary data are available for analysis. In these cases, criteria for selecting the bandwidth should include the goodness-of-fit statistics based on the cross-validation models, available bandwidth sample sizes, and external validity considerations.

Chapter 7: Illustrative Precision Calculations

In this section, I collate formulas from above and use key design parameter values from the literature to obtain illustrative MDE calculations for RD designs in the education field. The focus is on standardized test scores of elementary school and preschool students in low-performing schools. MDEs are calculated for each design considered above (where I use the multilevel versions of Designs II and III).

Presentation and Assumptions

Tables 7.1 and 7.2 display, under various assumptions and for each of the considered RD designs, the *total number of schools* that are required to achieve precision targets of 0.20, 0.25, and 0.33 of a standard deviation, respectively. These are benchmarks that are typically used in impact evaluations of educational interventions that balance statistical rigor and study costs (Schochet 2008; Hill et al. 2007). In Table 7.1, it is assumed that the score cutoff is at the center of the score distribution and that the treatment and control group samples are balanced. In Table 7.2, it is assumed that the cutoff is at a tertile of the score distribution and that there is a 2:1 split of the research samples. Table 7.3 presents comparable figures to those in Table 7.1 for the RA design.

Because the amount and quality of baseline data vary across evaluations, the power calculations are conducted assuming R^2 values of 0, 0.20, 0.50, and 0.70 at each group level. The R^2 value of 0.50 is conservative if pretests are available for analysis; the more optimistic 0.70 figure has sometimes been found in the literature (Schochet 2008; Bloom et al. 2005a).

To keep the presentation manageable, RD design effects are presented assuming that scores are normally distributed. As discussed, for a given treatment-control sample split, the RD design effect does not vary much according to the score distribution or location of the cutoff score. Thus, the results that are presented are broadly applicable, but could easily be revised using the alternative score distributions or parameter values that were discussed above.

The estimates also assume:

- A two-tailed test at 80 percent power and a 5 percent significance level
- The intervention is being tested in a single grade with an average of 3 classrooms per school per grade and an average of 23 students per classroom. Thus, the sample contains 69 students per school.
- 80 percent of students in the sample will provide follow-up (posttest) data, so that posttest data are available for about 55 students per school.
- ICC values of 0.15 at the school and classroom levels (which are consistent with the empirical findings in Schochet 2008, Hedges and Hedberg 2007, and Bloom et al. 2005a)
- An ICC value of 0.15 pertaining to the variance of treatment effects across schools in Designs IV and V (Schochet 2008)
- A sharp RD design rather than a fuzzy RD design (that is, that all units comply with their treatment assignments)

Table 7.1: Required School Sample Sizes to Detect Target Effect Sizes, for Various RD Designs

Assumes a Two-Tailed Test, a Value of .15 for the Intraclass Correlations, the Cutoff Score Is at the Center of the Normal Score Distribution and a Balanced Allocation of the Research Groups

	Number of Schools Required to Detect an Impact in Standard Deviation Units of:		
Unit of Treatment Assignment: Other Fixed or Random Effects	**.20**	**.25**	**.33**
I: Students Within Sites (Schools or Districts): Site Effects Fixed			
$R^2 = 0$	39	25	14
$R^2 = .2$	31	20	12
$R^2 = .5$	20	13	7
$R^2 = .7$	12	8	4
II: Classrooms Within Schools: Fixed School Effects			
$R^2 = 0$	141	90	52
$R^2 = .2$	113	72	41
$R^2 = .5$	71	45	26
$R^2 = .7$	42	27	16
III: Schools: Within Districts: No Random Classroom Effects			
$R^2 = 0$	357	229	131
$R^2 = .2$	286	183	105
$R^2 = .5$	179	114	66
$R^2 = .7$	107	69	39
IV: Students Within Schools: Random Site Effects			
$R^2 = 0$	63	40	23
$R^2 = .2$	50	32	18
$R^2 = .5$	31	20	12
$R^2 = .7$	19	12	7
V: Classrooms Within Schools: Random School Effects			
$R^2 = 0$	165	105	61
$R^2 = .2$	132	84	48
$R^2 = .5$	82	53	30
$R^2 = .7$	49	32	18
VI: Schools Within Districts: Random Classroom Effects			
$R^2 = 0$	459	294	169
$R^2 = .2$	367	235	135
$R^2 = .5$	230	147	84
$R^2 = .7$	138	88	51

Note: See the text for formulas and other assumptions underlying the calculations. The figures assume that the assignment scores are normally distributed.

Table 7.2: Required School Sample Sizes to Detect Target Effect Sizes, for Various RD Designs

Assumes a Two-Tailed Test, a Value of .15 for the Intraclass Correlations, the Cutoff Score Is at a Tertile of the Normal Score Distribution, and a 2:1 Split of the Research Groups

	Number of Schools Required to Detect an Impact in Standard Deviation Units of:		
Unit of Treatment Assignment: Other Fixed or Random Effects	**.20**	**.25**	**.33**
I: Students Within Sites (Schools or Districts): Site Effects Fixed			
$R^2 = 0$	39	25	14
$R^2 = .2$	31	20	12
$R^2 = .5$	20	13	7
$R^2 = .7$	12	8	4
II: Classrooms Within Schools: Fixed School Effects			
$R^2 = 0$	142	91	52
$R^2 = .2$	114	73	42
$R^2 = .5$	71	45	26
$R^2 = .7$	43	27	16
III: Schools: Within Districts: No Random Classroom Effects			
$R^2 = 0$	359	230	132
$R^2 = .2$	288	184	106
$R^2 = .5$	180	115	66
$R^2 = .7$	108	69	40
IV: Students Within Schools: Random Site Effects			
$R^2 = 0$	63	40	23
$R^2 = .2$	50	32	18
$R^2 = .5$	31	20	12
$R^2 = .7$	19	12	7
V: Classrooms Within Schools: Random School Effects			
$R^2 = 0$	166	106	61
$R^2 = .2$	133	85	49
$R^2 = .5$	83	53	30
$R^2 = .7$	50	32	18
VI: Schools Within Districts: Random Classroom Effects			
$R^2 = 0$	462	296	170
$R^2 = .2$	370	237	136
$R^2 = .5$	231	148	85
$R^2 = .7$	139	89	51

Note: See the text for formulas and other assumptions underlying the calculations. The figures assume that the assignment scores are normally distributed.

Table 7.3: Required School Sample Sizes to Detect Target Effect Sizes, for Various Random Assignment (RA) Designs

Assumes a Two-Tailed Test, a Value of .15 for the Intraclass Correlations, and a Balanced Allocation of the Research Groups

Unit of Treatment Assignment: Other Fixed or Random Effects	Number of Schools Required to Detect an Impact in Standard Deviation Units of: .20	.25	.33
I: Students Within Sites (Schools or Districts): Site Effects Fixed			
$R^2 = 0$	14	9	5
$R^2 = .2$	11	7	4
$R^2 = .5$	7	5	3
$R^2 = .7$	4	3	2
II: Classrooms Within Schools: Fixed School Effects			
$R^2 = 0$	51	33	19
$R^2 = .2$	41	26	15
$R^2 = .5$	26	16	9
$R^2 = .7$	15	10	6
III: Schools: Within Districts: No Random Classroom Effects			
$R^2 = 0$	130	83	48
$R^2 = .2$	104	66	38
$R^2 = .5$	65	42	24
$R^2 = .7$	39	25	14
IV: Students Within Schools: Random Site Effects			
$R^2 = 0$	42	27	15
$R^2 = .2$	33	21	12
$R^2 = .5$	21	13	8
$R^2 = .7$	12	8	5
V: Classrooms Within Schools: Random School Effects			
$R^2 = 0$	79	50	29
$R^2 = .2$	63	40	23
$R^2 = .5$	39	25	14
$R^2 = .7$	24	15	9
VI: Schools Within Districts: Random Classroom Effects			
$R^2 = 0$	167	107	61
$R^2 = .2$	134	85	49
$R^2 = .5$	83	53	31
$R^2 = .7$	50	32	18

Note: See the text for formulas and other assumptions underlying the calculations. The figures assume that the assignment scores are normally distributed.

Results

The key results can be summarized as follows:

- ***Much larger sample sizes are typically required under RD than RA designs.*** Consider the most commonly-used design in education-related impact studies where equal numbers of schools are assigned to treatment or control status. Under this design, about 114 total schools (57 treatment and 57 control) are required to yield an MDE of 0.25 standard deviations, assuming a regression R^2 value of 0.5 (Design III; Table 7.1). The corresponding figure for the RA design is only 42 total schools (Table 7.3). Similarly, for the classroom-based Design II, the required number of schools is 45 for the RD design (Table 7.1), compared to only 16 for the RA design (Table 7.3).

- ***Because of resource constraints, school-based RD designs may only be feasible for interventions that are likely to have relatively large effects (about 0.33 standard deviations or more).*** Under Design III, 66 schools (33 treatment and 33 control) are required to achieve an MDE of 0.33 standard deviations (assuming an R^2 value of 0.50; Table 7.1). This number is comparable to the number of schools that are typically included in large-scale experimental impact evaluations funded by the U.S. Department of Education.

- ***A 2:1 split of the sample has a small effect on statistical power.*** The required school sample sizes are similar in Tables 7.1 and 7.2. This occurs because as discussed, a balanced sample allocation yields larger RD design effects than an unbalanced allocation, but also yields smaller variances under the RA design; these two effects are largely offsetting.

- ***R^2 values matter.*** The viability of RD designs in education research hinges critically on the availability of detailed baseline data at the aggregate school or individual student level—and in particular, pretest data—that can be used as covariates in the regression models to improve R^2 values. For instance, for the school-based Design III, the number of schools required to achieve an MDE of 0.33 standard deviations is 39 if the R^2 value is 0.70, 66 if the R^2 value is 0.50, and 131 for a zero R^2 value (Table 7.1).

- ***RD designs may be most viable for less-clustered designs where classrooms or students are the unit of treatment assignment.*** For example, under the classroom-based Design II, 45 schools are required to achieve an MDE of 0.25 standard deviations, assuming an R^2 value of 0.50 (Table 7.1). The comparable figure for the classroom-based Design V (with random school effects) increases to only 53, because, as discussed, RD design effects are smaller for this design than for Design II. For the student-level Design I, the comparable number of required schools is 13 schools (Table 7.1).

Chapter 8: Summary and Conclusions

This paper has examined theoretical and empirical issues related to the statistical power of impact estimates under clustered RD designs that could be conducted in a school setting. The theoretical framework is grounded in the causal inference and HLM modeling literature, and the empirical work focused on group-based designs that are commonly used to test the effects of education interventions on student's standardized test scores.

The main conclusion is that much larger samples are required under RD than RA designs to produce rigorous impact estimates. This occurs because the large RD design effects that have been found previously for nonclustered designs carry over to most multilevel clustered designs that are typically used in education research. This pattern holds for a wide range of score distributions and score cutoff values.

These findings have important implications for the viability of using RD designs for new evaluations in the education field, due to the high cost of recruiting study schools, implementing interventions, and collecting data. Based on resources that are typically devoted to large-scale impact studies by the U.S. Department of Education and other funders, the results suggest that RD designs where schools are assigned to treatment or control status are likely to be feasible only for interventions that can have relatively large effects—0.33 standard deviations or more. RD designs appear to be more viable for less-clustered designs where classrooms or students are assigned directly to a research condition.

A key finding is that clustered RD designs can yield impact findings with sufficient levels of precision only if detailed baseline data—and in particular, pre-intervention measures of the outcomes—are collected and used in the regression models to increase R^2 values. Furthermore, RD designs will typically have sufficient power for detecting impacts at the pooled level only, but not for population subgroups; this problem is more severe for RD than RA designs.

In conclusion, although well-designed RD designs can yield unbiased impact estimates, they cannot necessarily be viewed as a substitute for experimental designs in the education field. School sample sizes typically need to be about three to four times larger under RD than RA designs to achieve impact estimates with the same levels of precision. Furthermore, RD designs yield impact findings that typically pertain to a narrower population (those with scores near the cutoff) than those from experiments (those with all scores), and rely on the validity of critical modeling assumptions that are not required under the RA design. The desirability of using RD designs will depend on the point of treatment assignment, the availability of pretest data, and key research questions.

Appendix A

Table A.1: Values for *Factor(.)* in Equation (1) of Text, by the Number of Degrees of Freedom, for One- and Two-Tailed Tests, and at 80 and 85 Percent Power

	One-Tailed Test		Two-Tailed Test	
Number of Degrees of Freedom	**80 Percent Power**	**85 Percent Power**	**80 Percent Power**	**85 Percent Power**
2	3.98	4.31	5.36	5.69
3	3.33	3.61	4.16	4.43
4	3.07	3.32	3.72	3.97
5	2.94	3.17	3.49	3.73
6	2.85	3.08	3.35	3.58
7	2.79	3.02	3.26	3.49
8	2.75	2.97	3.20	3.42
9	2.72	2.93	3.15	3.36
10	2.69	2.91	3.11	3.32
11	2.67	2.88	3.08	3.29
12	2.66	2.87	3.05	3.26
13	2.64	2.85	3.03	3.24
14	2.63	2.84	3.01	3.22
15	2.62	2.83	3.00	3.21
20	2.59	2.79	2.95	3.15
30	2.55	2.75	2.90	3.10
40	2.54	2.74	2.87	3.07
50	2.53	2.72	2.86	3.06
60	2.52	2.72	2.85	3.05
70	2.51	2.71	2.84	3.04
80	2.51	2.71	2.84	3.04
90	2.51	2.71	2.83	3.03
100	2.51	2.70	2.83	3.03
Infinity	2.49	2.68	2.80	3.00

Note: All figures assume a 5 percent significance level.

Appendix B

Lemma 1. Let $\hat{\alpha}_1$ be the OLS estimator for α_1 in the two-level model in (23). If the true functional form relationship between potential outcomes and the treatment assignment score is correctly specified in the model, then, $\hat{\alpha}_1$ is a consistent estimator for α_1. Furthermore, as the number of units, *n*, increases to infinity in (23) and for fixed *m*, $\hat{\alpha}_1$ converges to a normal distribution with variance:

$$(B.2.1) \quad AsyVar_{RD}(\hat{\alpha}_1^{TL}) = \frac{1}{p(1-p)(1-\rho_{TS}^2)}\left[\frac{\sigma_\tau^2}{n} + \frac{\sigma_\delta^2}{nm}\right],$$

where ρ_{TS} is the correlation between T_i^{RD} and $Score_i$. A comparable expression can be obtained for the aggregated model in (8) by setting $\sigma_\delta^2 = 0$ and replacing σ_τ^2 with σ_η^2.

Proof. Write (23) in terms of centered random variables as follows:

$$(B.2.2) \quad w_{ij}^* = \alpha_1 T_i^* + \alpha_2 S_i^* + (\tau_i^* + \delta_{ij}^*),$$

where $w_{ij}^* = w_{ij}^{RD} - E(w_{ij}^{RD})$, $T_i^* = T_i^{RD} - p$, $S_i^* = Score_i - E(Score_i)$, $\tau_i^* = \tau_i - E(\tau_i)$ and $\delta_{ij}^* = \delta_{ij} - E(\delta_{ij})$. Let $\tilde{w}_{ij}$, $\tilde{T}_i$, $\tilde{S}_i$, $\tilde{\tau}_i$ and $\tilde{\delta}_{ij}$ be respective *empirically* centered variables. If $Z_i^* = (T_i^* \; S_i^*)$ and $\tilde{Z}_i = (\tilde{T}_i \; \tilde{S}_i)$, then the OLS estimator for the parameters in (B.2.2) is as follows:

$$\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} = \left[\sum_{i=1}^{n}\sum_{j=1}^{m} \tilde{Z}_i' \tilde{Z}_i\right]^{-1}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} \tilde{Z}_i' \tilde{w}_{ij}\right].$$

Standard asymptotic arguments can be used to prove that as *n* approaches infinity,

$$(B.2.3) \quad \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} \xrightarrow{p} \left[mE(Z_i^{*\prime} Z_i^*)\right]^{-1} E(mZ_i^{*\prime} w_{ij}^*) = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \left[mE(Z_i^{*\prime} Z_i^*)\right]^{-1} E\left[mZ_i^{*\prime}(\tau_i^* + \delta_{ij}^*)\right].$$

In this expression,

$$(B.2.4) \quad \left[mE(Z_i^{*\prime} Z_i^*)\right]^{-1} = \begin{bmatrix} mp(1-p) & m\sigma_{TS} \\ m\sigma_{TS} & m\sigma_S^2 \end{bmatrix}^{-1} = \begin{bmatrix} \sigma_S^2 & -\sigma_{TS} \\ -\sigma_{TS} & p(1-p) \end{bmatrix}\left(\frac{1}{m(\sigma_S^2 p(1-p) - \sigma_{TS}^2)}\right)$$

and

$$E(mZ_i^{*\prime}[\tau_i^* + \delta_{ij}^*]) = \begin{pmatrix} m\sigma_{T\tau} \\ m\sigma_{S\tau} \end{pmatrix},$$

where $\sigma_{T\tau}$ is the covariance between T_i^{RD} and τ_i, and $\sigma_{S\tau}$ is the covariance between *Score*$_i$ and τ_i. Note that the covariance between Z_i^* and δ_{ij}^* is zero because T_i^{RD} and *Score*$_i$ do not vary within schools. Thus, after some algebra, it can be seen that as *n* approaches infinity,

$$(B.2.5) \quad \hat{\alpha}_1 \xrightarrow{p} \alpha_1 + \left(\frac{\sigma_S^2 \sigma_{T\tau} - \sigma_{TS}\sigma_{S\tau}}{\sigma_S^2 p(1-p) - \sigma_{TS}^2} \right).$$

The second term on the right-hand-side of (B.2.5) is zero because it is the coefficient estimate on T_i^{RD} when τ_i is regressed on T_i^{RD} and *Score*$_i$. This conditional expectation is zero, because controlling for *Score*$_i$, there is no variation in treatment status. (Note that this result does not hold if the model is specified incorrectly and τ_i contains omitted score variables.) Thus, $\hat{\alpha}_1$ is asymptotically unbiased.

To obtain the asymptotic distribution of the two-level OLS estimator, we can rewrite (B.2.3) as follows:

$$\sqrt{n}\begin{pmatrix} \hat{\alpha}_1 - \alpha_1 \\ \hat{\alpha}_2 - \alpha_2 \end{pmatrix} = n^{-1/2}\left[mE(Z_i^{*\prime} Z_i^*) \right]^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} Z_i^{*\prime}(\tau_i^* + \delta_{ij}^*) + o_p(1),$$

where $o_p(1)$ denotes a term that converges in probability to zero. Thus, using (B.2.4), we find after some algebra that

$$(B.2.6) \quad \sqrt{n}(\hat{\alpha}_1 - \alpha_1) = n^{-1/2} \frac{1}{m\left(\sigma_S^2 p(1-p) - \sigma_{TS}^2\right)} \sum_{i=1}^{n} (m\tau_i^* + \sum_{j}^{m} \delta_{ij}^*)(\sigma_S^2 T_i^* - \sigma_{TS} S_i^*) + o_p(1).$$

Because $E\left[(m\tau_i^* + \sum_{j=1}^{m} \delta_{ij}^*)(\sigma_S^2 T_i^* - \sigma_{TS} S_i^*) \right] = 0$, a simple application of the central limit theorem (see, for example, Rao 1973) can be used to show that $\hat{\alpha}_1$ is asymptotically normally distributed with mean zero and the following variance:

$$(B.2.7) \; AsyVar(\hat{\alpha}_1) = \frac{E(m\tau_i^* + \sum_{j=1}^{m} \delta_{ij}^*)^2 E(\sigma_S^2 T_i^* - \sigma_{TS} S_i^*)^2}{nm^2[\sigma_S^2 p(1-p) - \sigma_{TS}^2]^2} = \frac{\sigma_S^2(\sigma_\tau^2 + [\sigma_\delta^2 / m])}{n[\sigma_S^2 p(1-p) - \sigma_{TS}^2]}.$$

The expressions in (B.2.7) and (B.2.1) are equivalent because $\sigma_{TS}^2 = \sigma_S^2 p(1-p)\rho_{TS}^2$.

References

Angrist, J., G. Imbens, and D. Rubin (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 91, 444-472.

Bloom, H., J. Bos, and S. Lee (1999). Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for Evaluation of Education Programs. *Evaluation Review,* 23(4), 445-469.

Bloom, H. (1984). Accounting for No-Shows in Experimental Evaluation Designs. *Evaluation Review* 8(2), 225-246.

Bloom, H. (2004). Randomizing Groups to Evaluate Place-Based Programs. New York, NY: MDRC.

Bloom, H., J. Kemple, and B. Gamse (2005a). Memo on the Evaluation Design of the Reading First National Impact Study. New York, NY: MDRC.

Bloom, H., L. Hayes, and A. Black (2005b). Using Covariates to Improve Precision. New York, NY: MDRC.

Byrk, A. and S. Raudenbush (1992). *Hierarchical Linear Models for Social and Behavioral Research. Applications and Data Analysis Methods.* Newbury Park, CA: Sage.

Campbell, D. and J. Stanley (1963). *Experimental and Quasi-experimental Designs for Research on Teaching.* In N.L. Gage (ed.), Handbook of Research on Teaching. Chicago: Rand McNally.

Cappelleri, J., R. Darlington, and W. Trochim (1994). Power Analysis of Cutoff-Based Randomized Clinical Trials. *Evaluation Review,* 18, 141-152.

Cappelleri, J., W. Trochim, T. Stanley, and C. Reichardt (1991). Random Measurement Error Doesn't Bias the Treatment Effect Estimate in the Regression-discontinuity Design: The Case of No Interaction. *Evaluation Review,* 15(4), 395-419.

Cochran, W. (1963). *Sampling Techniques*. New York: John Wiley and Sons.

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

Cook, T. (2008). Waiting for Life to Arrive: A History of the Regression-Discontinuity Design in Psychology, Statistics, and Economics. *Journal of Econometrics*, 142(2), 636-654.

Donner, A. and N. Klar (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.

Dynarski, M. and R. Agodini (2003). The Effectiveness of Educational Technology: Issues and Recommendations for the National Study. Princeton, NJ: Mathematica Policy Research.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall: London.

Freedman, D. (2008). On Regression Adjustments to Experimental Data. *Advances in Applied Mathematics*, 40, 180-193.

Gleason, P. and R. Olsen (2004). Impact Evaluation of Charter School Strategies. Design Documents. Princeton, NJ: Mathematica Policy Research, Inc.

Goldberger, A. (1972). Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations. Working Paper, Economics Department, University of Wisconsin.

Griliches, Z. and V. Ringstad (1971). *Economies of Scale and the Form of the Production Function.* Amsterdan: North Holland.

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

Heckman, J. and E. Vytlacil (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73(3), 669-738.

Hedges, L. (2004). Correcting Significance Tests for Clustering. Chicago, IL: University of Chicago Working Paper.

Hedges, L. and E. Hedberg (2007). Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis,* 29(1), 60-87.

Hill, C., H. Bloom, A. Black, and M. Lipsey (2007). Empirical Benchmarks for Interpreting Effect Sizes in Research. New York, NY: MDRC.

Holland, P. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396), 945-960.

Imbens, G. and D. Rubin (2007). *Causal Inference: Statistical Methods for Estimating Causal Effects in Biomedical, Social, and Behavioral Sciences,* Cambridge University Press, forthcoming.

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

Jacob, B. and L. Lefgren (2004). Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. *Review of Economics and Statistics,* 86(1), 226-244.

Jackson, R. et al. (2007). National Evaluation of Early Reading First. Final Report to Congress. U.S. Department of Education, Institute of Education Sciences: Washington DC.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

Ludwig, J. and D. Miller (2007). Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design, *Quarterly Journal of Economics*, 122(1), 159-208.

Maddala, G. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Murray, D. (1998). *Design and Analysis of Group-Randomized Trials*. Oxford: Oxford University Press.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments: Essay on Principles. Section 9, Translated in *Statistical Science*, 1990: 5(4), 465-472.

Porter, J. (2003). Estimation in the Regression Discontinuity Model. Working Paper, Department of Economics, University of Wisconsin.

Rao, C. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley and Sons.

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

Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Education Psychology*, 66, 688-701.

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

Schochet, P. (2007). Is Regression Adjustment Supported by the Neyman Model for Causal Inference?. Working Paper: Mathematica Policy Research, Inc.: Princeton NJ.

Schochet, P. (2008). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics,* 33(1), 62-87.

Shadish, W., T. Cook, and D. Campbell (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Boston, MA: Houghton Mifflin Company.

Trochim, W. (1984). *Research Design for Program Evaluation: The Regression-discontinuity Design.* Beverly Hills, CA: Sage Publications.

Van Der Klaauw, W. (2002). Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-discontinuity Approach. *International Economic Review 43*, 12940-12945.

Yang, L. and Tsiatis, A. (2001). Efficiency Study of Estimators for a Treatment Effects in a Pretest-Posttest Trial. *American Statistician,* 55(4), 314-321.