1

The goal of this chapter is to provide a brief introduction to one of the most rigorous nonexperimental analytical methods currently employed by education researchers: regression discontinuity.

Applying Regression Discontinuity Design in Institutional Research

Allyson Flaster, Stephen L. DesJardins

Institutional researchers are often tasked with studying the effects of a variety of postsecondary education practices, policies, and processes. Examples include the effectiveness of precollegiate outreach programs such as summer bridge programs; whether first-year experience and developmental classes affect student outcomes; whether students residing in living–learning communities have outcomes that differ from their nonparticipating colleagues; and whether financial aid provision affects student outcomes such as persistence and completion. Given scarce institutional resources and the push for accountability in postsecondary education, decision makers are increasingly interested in whether institutional policies and programs actually achieve their intended goals.

Although a large body of research about the effectiveness of institutional interventions designed to improve student and institutional outcomes exists, there have been calls to improve the rigor of our research (DesJardins & Flaster, 2013; Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). In particular, there has been a push for education researchers to be able to make more rigorous ("causal") claims about our practices, policies, and processes.

Experiments (randomized controlled trials, or RCTs), which are characterized by the random assignment of subjects into treatment and control groups, are considered the "gold standard" for making causal claims (Schneider et al., 2007; Shadish, Cook, & Campbell, 2002). The rationale for conducting experiments is to be able to provide an unbiased estimate of the treatment on an outcome, but RCTs are often impracticable or may even be unethical in some research contexts (see Bielby, House, Flaster, & DesJardins, 2013; DesJardins & Flaster, 2013, for details). There are, however, statistical methods that can be employed when using observational (nonexperimental) data. These quasi-experimental methods attempt to

New Directions for Institutional Research, no. 161 © 2014 Wiley Periodicals, Inc.
Published online in Wiley Online Library (wileyonlinelibrary.com) • DOI: 10.1002/ir.20064

remedy the inferential problems that arise when units of observation, such as students, are not randomly assigned into treatment or control groups. Even though these methods, some of which are discussed in this volume, do not randomize units into treatment/control status, when properly applied they can substantially reduce any estimation bias due to nonrandom assignment.

Here we provide an introduction to one of these methods: the regression discontinuity (RD) design. In the next section, we discuss a framework often used as the conceptual basis in nonexperimental analyses such as RD.

Overview of the Counterfactual Framework

Many social scientists have employed the *counterfactual framework* in support of analysis designed to make rigorous claims about the effectiveness of institutional practices, policies, or processes ("interventions"). Collectively, these interventions are often referred to as *treatments*. The counterfactual framework posits that, hypothetically, each unit (individuals, classrooms, households, and so on) under study has two potential outcomes: one outcome under treatment and another outcome under nontreatment (Holland, 1986; Murnane & Willett, 2011). Ideally, to determine whether a treatment causes an effect, we would compare each unit's outcome in a world where it received the treatment and then compare its outcome in a counterfactual world where it did not receive the treatment.

For example, imagine we want to determine whether the provision of student financial aid (the treatment) improves the retention rate of students to the sophomore year (the outcome; henceforth, first-year retention). One way to study the effects of the provision of aid on retention is to use the students as controls (counterfactuals) for themselves. We could do this by providing some students with financial aid in their first year of college and then measure whether they are retained to the beginning of the sophomore year. Then, if we had a time machine, we would turn the clock back to the beginning of the freshman year, not give these students financial aid, measure their retention rate at the beginning of year two, and then calculate the difference between the two retention rates. The intuition is that comparing students to themselves under both the treatment and control conditions accounts for all the observed and unobserved factors that may affect their retention to the sophomore year. Thus, any difference in the retention rates between these two groups represents financial aid's *causal effect,* because the treatment condition would be the only factor that was different across these two states of the world. If we did this for a large number of students and averaged over each of their outcomes, we could ascertain an unbiased *average treatment effect* (ATE)—an estimate of the treatment's effect on the population of interest that is purged of influence from (possibly) confounding factors.

The fundamental problem in the example provided earlier, and in the application of the counterfactual framework more generally, is that we cannot observe units in these two different states of the world (Holland, 1986). We observe them only under the factual condition (the world we can observe), whereas outcomes under the counterfactual condition remain unknown. This fundamental problem is, essentially, a missing data problem (Murnane & Willett, 2011). For example, assume that the outcome (Y) is retention to the sophomore year and the treatment (T) equals 1 when aid is provided to a student and 0 when it is not. Typically we possess data on the outcome under treatment for those in the treatment group ($Y_1 \mid T = 1$), but not the outcome under treatment for those in the control group ($Y_1 \mid T = 0$), and vice versa.

Researchers often attempt to approximate the counterfactual condition by employing experiments where units are randomized into treatment. When correctly implemented, these designs result in the treated and control groups having (on average) identical observable and unobservable characteristics and differing only with regard to their treatment status (Murnane & Willett, 2011; Schneider et al., 2007). When this is the case, the ATE can be obtained by simply comparing the average outcomes, or the means, for the treated and control groups. When treatment assignment is done using randomization, the mechanism by which assignment takes place is exogenous. In the context of causal inference, *exogeneity* refers to variation that occurs because it is determined outside of the model under analysis and is used to assign units to either the treatment or control condition. Its converse, *endogeneity*, occurs when a unit is assigned to treatment status by an "agent" within the system under study (see Murnane & Willett, 2011, for additional details).

In colleges and universities, endogenous treatment assignment is the norm. Students, a common unit of analysis, often choose the classes they take, the types of financial aid they apply for, and the support services they receive. Similarly, faculty members often choose different types of pedagogy, whether to engage in interdisciplinary research, or whether to participate in technology training. Endogenous treatment assignment complicates making causal inferences about these interventions, because units such as students or faculty who elect to choose a treatment may be systematically different in unobserved or unmeasured ways than those who do not choose treatment. Any unobserved factors that are related to both the receipt of treatment and outcomes are called *confounding* factors. For instance, motivation, which is typically unmeasured in observational data, may affect one's treatment status (whether a student receives aid, which requires an application) and also affect the outcome (e.g., that student's retention). Untangling this confoundedness is a major challenge, one that we can attempt to remedy using different designs and statistical methods.

Randomized trials may be the best method for untangling confoundedness, but it is not always possible to use them to study the effectiveness

of a program or policy (Cook, 2002). For example, there may be resistance to randomly assigning students, regardless of their motivation to succeed in their coursework, to participate in a new tutoring program. Furthermore, RCTs necessitate considerable work in the research design stage to ensure that they are properly executed (Murnane & Willett, 2011). This requires institutional researcher involvement with program evaluations *before* treatments are administered. Oftentimes, however, institutional researchers are not asked to evaluate the effectiveness of an intervention until *after* the intervention has been implemented.

Fortunately, over the course of several decades, analytical methods have been developed that can help researchers to make very rigorous inferences when treatment assignment is nonrandom (Cook, Shadish, & Wong, 2008). Of the various analytical approaches used to solve the missing data problem inherent in the counterfactual framework, RD design is generally viewed to be one of the more rigorous nonexperimental methods available for estimating treatment effects (Steiner, Wroblewski, & Cook, 2009; U.S. Department of Education, 2011).

Fundamentals of Regression Discontinuity Design

RD design has a relatively lengthy history within the field of education and program evaluation (Cook, 2008). Thistlethwaite and Campbell employed the technique in the late 1950s to study the effects of "certificates of merit" provided by the National Merit Scholarship (NMS) Program on (a) high school students' ability to obtain funding for college and (b) their plans to pursue advanced degrees. They proposed that legitimate counterfactuals could be produced by capitalizing on a feature of the certificate awarding process: Students were eligible to receive an NMS certificate if they scored at or above a threshold on a standardized test. Thistlethwaite and Campbell (1960) reasoned that—unlike students at opposite ends of the test score distribution—students who narrowly missed earning a certificate (e.g., by one point), and those who received a certificate by scoring at or just above the threshold, shared very similar characteristics. Key to their argument was the assertion that, as is true in a randomized trial, the only substantial observed and unobserved difference between the two groups of students at the threshold margin was that one group was exogenously provided with a treatment by the National Merit Scholarship Corporation and the other was not. Thus, any significant differences in outcomes between the two groups could be attributed to the certificates' treatment effects.

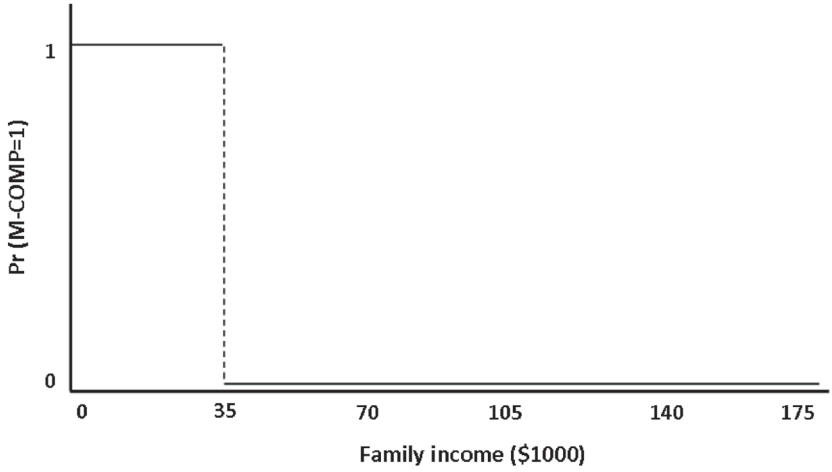
Researchers in education psychology, economics, and statistics have made many contributions to RD's conceptual and practical development since the 1960s (Cook, 2008). We provide an introduction to the method in the following sections, noting when it is appropriate to use RD, some of the assumptions underlying its use, and how to test whether these assumptions are being violated. We do so using a hypothetical example from

the world of institutional research. Because this chapter is geared toward researchers who have little to no experience with RD design, we keep detailed technical and conceptual points to a minimum and point interested readers to additional sources where appropriate.

A Hypothetical Example. Midwest University is a (fictional) large university that is striving to become more socioeconomically diverse. Two years ago, administrators decided to implement a loan replacement grant program for low-income students modeled after similar financial aid programs at Princeton University and the University of North Carolina-Chapel Hill. The grant program, dubbed the "Midwest Compact" (M-Comp), replaces all loans in eligible students' financial aid packages with institutional gift aid that does not need to be repaid. Students are automatically eligible to receive the M-Comp grant if they have a family income lower than a maximum amount set each year by the university. In the first year of the program, only students with a gross family income of $35,000 or less—as reported on their prior year federal income tax form—were eligible to receive the M-Comp grant. In the second year, the income cutoff was increased to $37,000. Approximately 10% of Midwest University undergraduates were eligible to receive the grant each year ($N = 3{,}050$ in year one; $N = 3{,}402$ in year two). The family income cutoffs were set by enrollment managers prior to the administration of grant funds, and the specific income cutoffs were not publicly announced prior to students applying for financial aid from Midwest University in either year. The university spent about $5 million on the M-Comp grant program in its first two years of implementation, so administrators were eager to know if the money was well spent. Next we will introduce the basic concepts of the RD design while discussing how an institutional researcher at Midwest University could apply RD to estimate the causal effect of the loan replacement grant program on the first-year retention of recipients.

Running Variable. A defining feature of RD design is that an individual's probability of receiving the treatment is determined by that person's value on a (typically) continuous variable, referred to as the *running* or *forcing variable* (Imbens & Lemieux, 2008). The point along the running variable at which one's probability of receiving the treatment jumps considerably or discontinuously (hence the name) is called the *cut-point* or *treatment threshold*. In the M-Comp example discussed earlier, the running variable that determines eligibility for treatment is family income, and the cut-points are $35,000 (in year one) and $37,000 (in year two). As a first step in applying RD, the researcher should examine the relationship between the probability of receiving the treatment and the running variable. Figure 1.1 illustrates the case where all students with a value to the left of (or "below") the cut-point (represented by the dashed vertical line at 35) for year one receive the M-Comp grant in their financial aid package, and all students with a value to the right of (or "above") the cut-point do not. In such a case, a student's *probability* of being treated is either 1 or 0,

Figure 1.1. Probability of Receiving M-Comp Grant by Student Family Income in a Sharp Regression Discontinuity Design, Year One

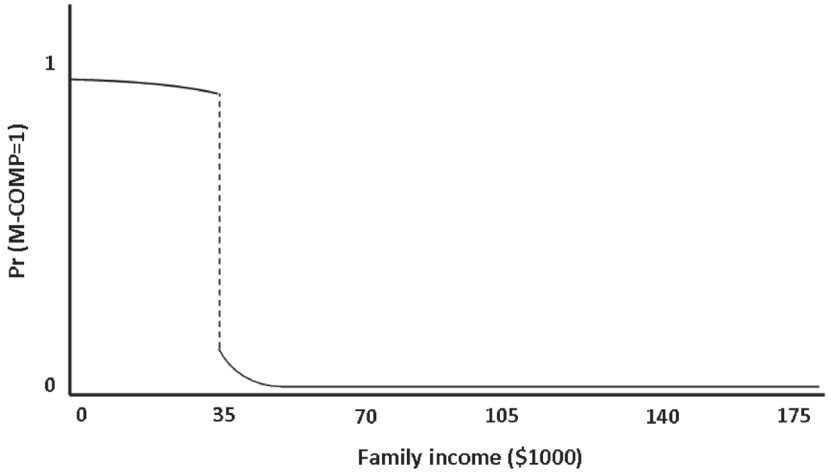


determined wholly by the student's family income. When the probability of treatment is wholly determined by the cut-score, such as in Figure 1.1, the RD design is said to be a "sharp" design (DesJardins, McCall, Ott, & Kim, 2010; McCall & Bielby, 2012). Sharp designs occur when all subjects being studied comply with the threshold-based policy that determines who is in the treatment group and who is in the control group (Lesik, 2008).

However, sometimes there is noncompliance with the mechanism determining treatment assignment. For example, administrators may adjust eligibility criteria, so that individuals who should be placed in the control group based on the established threshold are suddenly eligible for the treatment. In addition, individuals eligible for treatment given the criteria being used to establish the cut-point may opt out of treatment receipt, thereby placing themselves in the control group.

To examine the extent of compliance with the assignment mechanism, an institutional researcher at Midwest University graphed students' probabilities of receiving the M-Comp grant by family income in year one and found that some students whose family income was below the cut-point did not receive the M-Comp grant, and some students whose income was above the cut-point did (see Figure 1.2). After speaking with employees in the financial aid office, the IR staff member learned that the M-Comp grant is a "last dollar" award, meaning that the grant is used to make up the difference between the students' other need- and merit-based scholarships and their total cost of attending Midwest University. Some low-income students received a sufficient amount of federal, state, and departmental scholarships (all of which are applied to need first) to cover their cost of attendance so they did not need the M-Comp grant. The IR staff member also learned

Figure 1.2. Probability of Receiving M-Comp Grant by Student Family Income in a Fuzzy Regression Discontinuity Design, Year One



that, occasionally, staff in the financial aid office exercised professional judgment, which is perfectly legitimate, allowing students with family incomes of up to $45,000 to be eligible for M-Comp if they had extenuating circumstances such as a parental job loss within the past year or a family size greater than four.

Earlier we discussed the case when the assignment mechanism is deterministic. In instances where the cut-score does not strictly determine treatment status, but there is a "jump" in the probability of treatment at the cut-score (as in Figure 1.2), the RD is said to be a "fuzzy" design (McCall & Bielby, 2012; Trochim, 1984). In a fuzzy design, treatment assignment is determined by both the exogenous running variable and other factors that are potentially endogenous (DesJardins et al., 2010). This makes the estimation of causal effects a bit more complex. We will discuss the implications of fuzzy designs for making causal inferences in greater detail in the following sections.

Assumptions of Regression Discontinuity. The hypothetical example from the preceding section highlights the importance of understanding the *mechanism(s)* behind treatment assignment when employing RD design. Once the researcher understands the mechanism behind selection into treatment status, and thus whether the RD design for her study will be sharp or fuzzy, she can begin to examine the relationship between the treatment variable and the outcome. First, it is important that researchers have a firm understanding of the assumptions that underlie RD analysis. An important assumption that needs to be checked when using RD is that the observations are randomly distributed near ("locally" around) the cut-point (Lee & Lemieux, 2010). This concept—known as *local randomization*—was

fundamental to Thistlethwaite and Campbell's (1960) assertion that, on average, students who barely won and those who just missed winning an NMS certificate were essentially identical in observed and unobserved ways. Perhaps the students who missed earning a certificate by one point were just as meritorious as those just above the threshold, but performed slightly less well on the standardized test that determined treatment because they were, for any number of reasons, having an "off day."

The local randomization assumption holds that if, in a counterfactual world, the National Merit Scholarship Corporation administered the same standardized test to the same group of potential certificate winners again, these students' placement around the test score cut-point would be randomly determined. In other words, students locally distributed around the cut-point when the test was administered the first time would have a 50-50 chance of falling above or below the cut-point the next time the test was administered. *Essentially, the probability of placement into the treatment or control group is akin to a coin-flip for students locally distributed around the cut-point*. These students' treatment group assignment could potentially vary across the two counterfactual worlds only if they do not have perfect control over whether they are treated or not (Lee & Lemieux, 2010). Thus, as discussed previously, another underlying assumption of RD design is that there is some degree of *exogenous variation* in treatment group assignment (Lee & Lemieux, 2010).

Applying local randomization to the M-Comp grant example, an institutional researcher at Midwest University might assume that whether a student's family income is immediately above or below the income cutoff in a given year is determined by random factors, and this assumption can be tested using several methods. One approach is to plot pretreatment characteristics against the running variable (see Imbens & Lemieux, 2008; McCall & Bielby, 2012, for more information). This strategy was used in a study of the effects of the Gates Millennium Scholars (GMS) program on a number of outcomes and explained in detail in McCall and Bielby (2012). To test the assumption of no differences near the cut-point threshold, the researchers regressed the amount of loan aid students received (one of the outcomes of interest) on baseline variables such as the student's gender and SAT score. They then calculated the average predicted values of the outcome for each of the values of the running variable, a noncognitive test score used for treatment assignment. They then plotted the predicted values against the values of the noncognitive test score. This graph allowed them to visually ascertain whether there was a discontinuity in any of the student pretreatment characteristics that corresponds with the running variable (see Figures 5.6 and 5.7 in McCall & Bielby, 2012, for examples of such plots).

Researchers can also check the validity of the local randomization assumption by comparing the average values of pretreatment characteristics immediately around the cut-point (see Calcagno & Long, 2008; DesJardins et al., 2010; Lee, 2008, for more information). For instance, Calcagno and

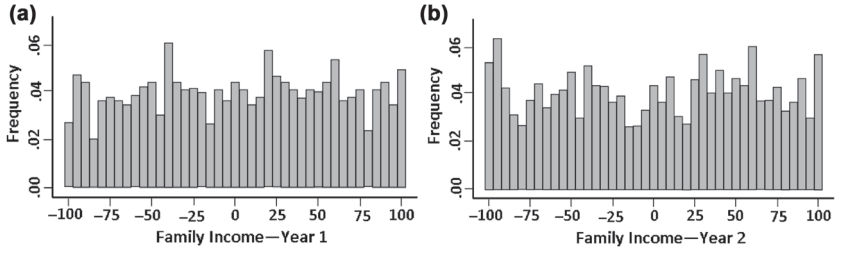
Long (2008) found that, in their full sample, there were differences in average characteristics such as age and gender between students who were assigned to remedial education (using a standardized test) and those who were not. However, these differences disappeared when only students who scored 10 points above and below the cut-point were compared, implying similarities between the two groups based on (at least) these observable characteristics.

It is also important to examine the running variable's distribution to check for violations of the exogenous treatment assignment assumption (Murnane & Willett, 2011). McCrary (2008) notes that researchers should consider whether individuals in the sample had an opportunity to completely manipulate their values on the running variable. Complete manipulation is likely when individuals (a) know the specific cutoff value prior to treatment assignment, (b) are incentivized to seek the treatment (or not), and (c) have the capacity through time and effort to modify their value on the running variable (see McCrary, 2008, for more information).

For example, suppose Midwest University had announced the income cutoffs for the M-Comp grant in a press release before financial aid applications were due. If students viewed the M-Comp grant as desirable and were able to reduce their work hours so that their family income was just below the maximum to qualify for the grant, then they could perfectly manipulate their treatment status. This would be, however, a violation of the assumption of exogeneity needed to make causal inferences when using the RD method. In such a case, an examination of the data would indicate that more students than expected have family incomes directly below the income cutoff of $35,000, and fewer students than expected have family incomes directly above the cutoff. Luckily, however, administrators at Midwest University did not announce the income cutoffs prior to distributing the grant funds.

Examining the distribution of the treated/nontreated within $100 of each side of the cut-point, an institutional researcher at Midwest University found reassuring visual evidence that students were not manipulating their treatment status (see Figure 1.3). Histograms of family income in years one and two (panels a and b, respectively) do not exhibit large jumps in density

Figure 1.3. Family Income Histograms



near the M-Comp grant income cutoffs (the points at 0), suggesting that students did not have the ability to completely determine whether they were eligible for the aid. In situations where complete manipulation of treatment status is suspected, researchers may want greater assurance than a simple visual inspection can provide. For examples of a more rigorous test of treatment manipulation when using discrete (noncontinuous) running variables such as family income or standardized tests, see Calcagno and Long (2008) and DesJardins and McCall (2014).

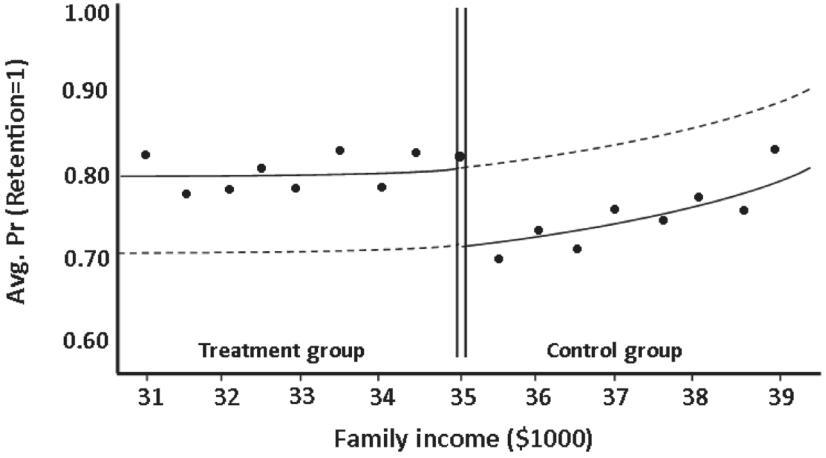
Making Causal Inferences

When they are satisfied that the underlying assumptions discussed earlier cannot be refuted, researchers can begin the process of estimating the causal effect of the treatment on the outcomes of interest. In the case of Midwest University, administrators want to know if the M-Comp grant (the treatment) is effective at inducing students to remain enrolled to their second year (the outcome). The counterfactual conditions they wish to uncover are: (a) what the average retention rate would have been for treated students if these students had not received the grant ($Y_0 \mid T = 1$) and (b) what the average retention rate would have been for untreated students if these students had received the grant ($Y_1 \mid T = 0$)—outcomes that are impossible to observe without our yet-to-be-invented time machine.

An institutional researcher at Midwest University can construct counterfactuals using data from students immediately around the income cutoff and then project what the retention outcomes would have been for treated students if they had not received the M-Comp grant, and vice versa. The logic behind this approach is illustrated in Figure 1.4, where the average predicted probability of being retained to year two is plotted in an $8,000 window around the cut-point (the point at the double line). For simplicity, the figure depicts a sharp RD design where all students comply with the threshold-based policy. Solid regression lines, which represent the relationship between retention and family income, are fitted to observations on each side of the cut-point. We can use these regression lines to infer what the counterfactual outcomes would have been for students who did and did not receive the grant. When these regression lines cross over the threshold, they become dotted to represent that these portions of the lines are extrapolations into areas where factual data do not really exist.

For example, Figure 1.4 demonstrates that the average predicted probability of retention for students with a family income of $35,000 is approximately 0.80. This is evident by reading the value on the *Y* axis at the point where the regression line on the left side of the cut-point (the line fitted to the treatment group) crosses the cut-point threshold. If we extrapolate the regression line on the right side of the cutoff (the line fitted to the control group) into the treated group region, its trajectory suggests what the average outcome would have been for treated students if they had not received the

Figure 1.4. Average Probability of Retention by Selected Values of Family Income, Regression Lines Added, Year One



M-Comp grant. The point where this regression line crosses the threshold and intersects the *Y* axis is 0.72, the counterfactual estimate of the outcome for treated students at the cut-point.

Regression Discontinuity Models. Figure 1.4 is an illustration of some of the concepts that underlie RD estimation. To formally estimate treatment effects, one needs to model the relationship between the treatment, the running variable, and the outcome among individuals who are locally distributed around the cut-point (Lesik, 2008; Murnane & Willett, 2011). Parametric and nonparametric regression techniques have been used to do just that, but due to space limitations we will only cover parametric techniques in this chapter. However, Imbens and Lemieux (2008) and McCall and Bielby (2012) offer detailed discussions of how to use nonparametric techniques, such as local linear regression, to model the running variable/outcome relationship on each side of the cutoff.

To facilitate interpretation of the treatment effects of the M-Comp grant, we use a linear probability model. Formally, the model could be defined as

$$\Pr(Y = 1) = \beta_0 + \beta_1(M) + \beta_2(X) + \varepsilon, \tag{1.1}$$

where β_1 represents the effect of the M-Comp grant (M) on the probability of retention ($Y = 1$), β_2 represents the underlying relationship between family income (X) and retention, and ε is a random error term. M is a dichotomous treatment indicator whose value is determined such that $M = 0$ (untreated) if family income (c) $> 35{,}000$ in year one or $c > 37{,}000$ in year two, and $M = 1$ (treated) if $c \leq 35{,}000$ in year one or $c \leq 37{,}000$ in year two.

Pretreatment variables such as SAT score and demographic variables could also be added to the regression model to decrease variance (improve model power) and test for nonrandomness around the cut-point (McCall & Bielby, 2012). If pretreatment characteristics are significant predictors of the outcome, then individuals in the sample may not be randomly distributed around the cut-point.

For ease of interpretation and to facilitate the combining of data from the two academic years, the institutional researcher at Midwest University could transform ("normalize") the family income running variable to indicate a student's relative distance from the cut-point (using $X =$ family income $-$ cutoff score). Centering the running variable on the cut-score allows one to interpret the intercept as the counterfactual outcome for individuals at the treatment threshold (Murnane & Willett, 2011). Thus, β_0 in Equation 1.1 provides an estimate of what the average probability of retention would have been for students with incomes of $35,000 ($37,000) in year one (two) if they had not been eligible for the M-Comp grant.

Choosing a Bandwidth. One way the RD design differs from traditional regression analysis is that the researcher does not necessarily include the full sample of data in her estimation of the treatment effects. You may recall that Thistlethwaite and Campbell's (1960) intuition for the development of RD was that only students near the test score cut-point were randomly distributed across the certificate of merit threshold. But an important question is: What does "near" the cut-point mean?

One of the goals of RD is to identify a group of individuals assigned to the control group who can serve as reasonable counterfactuals for the treated group. In other words, with the exception of their treatment group assignment, the individuals assigned to be treated are, *on average*, identical to the individuals who are assigned to the control group in all observed and unobserved ways. Of course, it is impossible to know if the average values of unobserved characteristics are the same between the two groups. Nonetheless, researchers need to decide which observations to include in the analysis.

Choosing an analytic window around the cut-point (known as a "bandwidth") often involves striking a balance between the need for power and the need for bias reduction (McCall & Bielby, 2012). Smaller bandwidths reduce the possibility of model misspecification and estimation bias, but can result in imprecise (higher variance) treatment effects estimates if there are too few observations located within the observation window. Although statisticians have not yet identified a minimum sample size needed to conduct valid RD analyses, Bloom (2012) describes a formula to approximate the *minimum detectable effect* (MDE), or "the smallest true treatment effect (or effect size) that has an 80% chance (80% power) of producing an estimated treatment effect that is statistically significant" (p. 64). Researchers can input various sample sizes into this formula to help determine a target MDE. Because RD design requires extrapolation into areas where factual

data do not exist, a general rule of thumb is that samples used in RD analyses need to be larger than samples used in randomized trials by a factor of at least 2.72 to produce the same level of precision in their estimates (Bloom, 2012).

After combining the data from year one and year two to (potentially) increase power, the institutional researcher at Midwest University found that there were only 228 students with a family income in a narrow range around the cut-point (arbitrarily defined as + or −$100). Worried that this small sample size would make it difficult to uncover the treatment effect of the M-Comp grant, the IR staff member examined the average values of observable student characteristics in this interval and decided to expand the analytic bandwidth to be $250 on either side of the cutoff. The rationale for increasing the bandwidth was that students assigned to the treatment and control groups within this interval exhibit no statistically significant differences based on their observable characteristics such as sex, race/ethnicity, entering major, and high school GPA. In addition, widening the bandwidth increased the effective sample size to 655 students, more than double the size when the interval around the cut was only $100.

Specifying a Functional Form. Why is a regression model used in RD, rather than simply approximating a random experiment by taking the difference in the average (mean) outcomes among the treated and untreated students at or near the cut-point? The answer is that it is important to account for the relationship between the running variable and the outcome in order to accurately measure the effect of the treatment on the outcome (Murnane & Willett, 2011). Indeed, many scholars have documented that there is a positive relationship between student family income or social class and college retention (Chen & DesJardins, 2008; Kim, 2007; Walpole, 2003). Most likely, students at Midwest University are no different from students elsewhere in the United States—making those with the greatest access to material resources the most likely to be retained in college. Thus, even if the M-Comp grant had never been implemented, differences in the average probability of retention could have existed between students whose family incomes are just below and above the threshold. By modeling the relationship between the running variable and the outcome, we help ensure that the M-Comp treatment effect (β_1) on Equation 1.1 is purged of underlying relationships such as the income and retention correlation.

However, including the running variable in the analysis requires that the researcher accurately model the functional form of the regression (Lesik, 2008; McCall & Bielby, 2012; Murnane & Willett, 2011). Relationships may not always be linear; for example, family income may have differential impacts on student education outcomes at lower levels than at higher levels of income. Thus, it is important to consider the possibility that the slope of the relationship between the running variable and the outcome may differ on opposite sides of the cut-point and across the values of the running variable.

Researchers can use several methods to attempt to identify the correct functional form between the running variable and the outcome, thereby helping to avoid model misspecification bias (Lesik, 2008). One useful approach is to begin by graphing the relationship between the running variable and the outcome and then fitting a line that appears to best approximate the relationship (Lesik, 2007, 2008; McCall & Bielby, 2012). For example, Lesik (2007) used a nonparametric *lowess smoothing* technique to graph the relationship between a course placement test score and college retention. The graph indicated that, in her sample, a linear logit transformation was appropriate for data on both sides of the cut-point.

Another approach to find the correct functional form is to include higher orders of the running variable—such as quadratic or cubic—or interactions between the running variable and the treatment variable in the regression (Lesik, 2008; Murnane & Willett, 2011). Using this approach, DesJardins and McCall (2014) found that a model with a quadratic form of the noncognitive test score used to assign GMS scholarships was the most appropriate specification.

Although it is advisable to check for the appropriate functional form, in practice it may not be as critical when one is modeling using data very close to the cut-point—especially if there are very large samples in this interval. This is because the smaller the analytic bandwidth, the more likely it is that the slope of the regression line is approximately linear in this smaller interval (McCall & Bielby, 2012).

Estimating Treatment Effects. In a sharp RD design, the treatment effect is the parameter associated with the dichotomous treatment indicator. This treatment effect is estimated using only data in a window around the cut-point. In Equation 1.1, the parameter β_1 reflects what the ATE of the M-Comp grant would have been for students at the cut-point had all students and staff at Midwest University complied with the threshold-based policy. Therefore, the treatment effect is referred to as the *local average treatment effect* (LATE; McCall & Bielby, 2012). The LATE can be visualized in Figure 1.4 as the vertical distance between the two regression lines evaluated at the cut-point. It is important to note that the LATE is a measurement of the estimated treatment effect for students *at the margin* of receiving the intervention being studied. In our M-Comp grant example, the marginal students are those whose family income places them directly at the cut-point (those with family incomes of $35,000 and $37,000). We can also make the assumption that the LATE reflects the effect of the treatment on those who are included in the analytical bandwidth (for instance, students with a family income of $34,800 in year one). However, as individuals far from the cut-point are often not included in an RD analysis, no inferences should be made as to how the treatment affects their behavior (Murnane & Willett, 2011). Given that the treatment effect (LATE) is local, it may not be an accurate estimate of the M-Comp grant's effect on students from families with very low income levels (e.g., $0 to $5,000).

But as noted earlier in the chapter, the mechanisms underlying the M-Comp grant may not be amenable to a sharp RD design. Some low-income students assigned to the treatment group based on their family income did not actually receive the grant because they had already received the maximum financial aid allowable. Also, some students with incomes above the threshold (those who were assigned to the control group) received the grant (were treated) due to an administrative intervention. Individuals whose assigned treatment group and actual treatment group differ are referred to as *crossovers* (Shadish et al., 2002).

There are several approaches that researchers can use to account for crossovers in an RD design. One approach is to estimate a sharp RD model, such as in Equation 1.1, and redefine the study's research question to be an examination of the effect of being *offered* a treatment rather than the effect of actually having *received* a treatment. This is known as an *intent-to-treat* analysis (Shadish et al., 2002). If the institutional researcher at Midwest University were to take this approach, her analysis would produce an estimate of the effect of having an *income at or below the M-Comp cut-point* on an individual's probability of retention, not an estimate of M-Comp *receipt* on retention.

However, an intent-to-treat analysis does not always answer the substantive questions that educational stakeholders have about the effectiveness of an intervention. For example, the administrators at Midwest University want to know if the M-Comp grant program improved the likelihood of retention for actual recipients, not students who were just income-eligible to receive the grant. Thus, another approach the IR staff member could take is to eliminate crossovers from the sample and estimate Equation 1.1. This approach is only appropriate if crossovers constitute a small proportion of the sample—typically 5% or less—or are confined to a narrow range of the running variable immediately around the cut-point (Shadish et al., 2002; Trochim, 1984). If eliminating crossovers is not an option, then one could use instrumental variable (IV) methods. Space restrictions prevent us from illustrating how to employ IV estimation here, but an example of IV use in educational evaluation can be found in Bielby et al. (2013).

Sensitivity Analysis. A crucial step in conducting an RD analysis is to check the sensitivity of the estimated treatment effects to variations in the model specification. Point estimates that are stable across model specifications provide more believable evidence of the LATE than is the case when such estimates change depending on the regressors included, their form (quadratic, cubic), or the estimation bandwidth chosen. In particular, it is important to check how sensitive the estimates are to changes in the bandwidth and the specification of the functional form of the running variable/outcome relationship. Table 1.1 presents the institutional researcher's (hypothetical) parameter estimates of M-Comp's ATE on recipients' probability of first-year retention, using various bandwidths. Administrators may

Table 1.1. Parameter Estimates of M-Comp ATE on One-Year Retention Across Various Bandwidths

	Cutoff Score ±$250 (Model 1)	*Cutoff Score* ±$350 (Model 2)	*Cutoff Score* ±$450 (Model 3)	*Cutoff Score* ±$550 (Model 4)
ATE	0.079	0.085	0.100	0.111
SE	0.038	0.032	0.028	0.027
N	655	827	1101	1290

be relieved to note that the treatment estimates across all bandwidths suggest a positive effect of the grant on retention to the sophomore year.

Note that Model 1, which has the most restrictive bandwidth, estimates that the M-Comp grant increases recipients' probability of first-year retention by 7.9 percentage points, and this result is significant at conventional levels ($t = 2.08$; $p < .05$). The results from Models 2, 3, and 4, each of which has wider bandwidths, provide estimates that the grant increases retention by a statistically significant 8.5, 10.0, and 11.1 percentile points, respectively.

In addition to checking for robustness across bandwidths, one should test the sensitivity of the results to the functional form of the model. Shadish et al. (2002) recommend overfitting a test case model by including higher order polynomial terms of the running variable (quadratic/cubic terms) and interactions of these with the treatment indicator. Then the analyst successively removes these variables from the regression model and uses fit statistics, such as an F test when using ordinary least-squares regression or likelihood ratio tests when using logistic regression, to test whether the constrained models (the ones with the higher order/interaction terms) are a better fit to the data than the models that do not include these variables (the unconstrained models). (Interested readers should consult Lesik [2008] for more details about this approach, and see pp. 269 and 270 in McCall and Bielby [2012] for an example of how to present and interpret the results of a model that is estimated with linear, quadratic, and cubic terms.)

Conclusion

Institutional researchers have been very successful in informing education decision makers about the factors that are *correlated* with educational outcomes, but we have been less successful in determining whether there are *causal* linkages among interventions and educational outcomes. One reason may be researchers' failure to apply designs and methods that can help unravel the causal effects of educational treatments on outcomes. Given the push for researchers to better understand the causal mechanisms underlying much of what we study, it is incumbent on us to employ the types of

methods that will allow us to better understand what programs, policies, and practices are truly effective. Doing so has the potential to improve institutional decision making and, by extension, the prospects of the varied stakeholders we serve. We hope this chapter and the references provided throughout it, along with the other chapters in this volume, will help inform institutional researchers about the utility and proper applications of these methods.

References

Bielby, R. M., House, E., Flaster, A., & DesJardins, S. L. (2013). Instrumental variables: Conceptual issues and an application considering high school course taking. In M. B. Paulsen (Ed.), *Higher education: Handbook of theory and research* (pp. 263–321). Dordrecht, the Netherlands: Springer.

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

Calcagno, J. C., & Long, B. T. (2008). *The impact of postsecondary remediation using a regression discontinuity approach: Addressing endogenous sorting and noncompliance* (Working Paper No. W14194). Washington, DC: National Bureau of Economic Research.

Chen, R., & DesJardins, S. L. (2008). Exploring the effects of financial aid on the gap in student dropout risks by income level. *Research in Higher Education*, *49*(1), 1–18.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24*(3), 175–199.

Cook, T. D. (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.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*(4), 724–750.

DesJardins, S. L., & Flaster, A. (2013). Non-experimental designs and causal analyses of college access, persistence, and completion. In L. W. Perna & A. P. Jones (Eds.), *The state of college access and completion* (pp. 190–207). New York, NY: Routledge.

DesJardins, S. L., & McCall, B. P. (2014). The impact of the Gates Millennium Scholars Program on college and post-college related choices of high ability, low-income minority students. *Economics of Education Review*, *38*, 124–138.

DesJardins, S. L., McCall, B. P., Ott, M., & Kim, J. (2010). A quasi-experimental investigation of how the Gates Millennium Scholars Program is related to college students' time use and activities. *Educational Evaluation and Policy Analysis*, *32*(4), 456–475.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–970.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635.

Kim, D. (2007). The effect of loans on students' degree attainment: Differences by student and institutional characteristics. *Harvard Educational Review*, *77*(1), 64–100.

Lee, D. S. (2008). Randomized experiments from non-random selection in US House elections. *Journal of Econometrics*, *142*(2), 675–697.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355.

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

Lesik, S. A. (2008). Studying the effectiveness of programs and initiatives in higher education using the regression-discontinuity design. In J. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 277–297). Dordrecht, the Netherlands: Springer.

McCall, B. P., & Bielby, R. M. (2012). Regression discontinuity design: Recent developments and a guide to practice for researchers in higher education. In J. Smart & M. Paulsen (Eds.), *Higher education: Handbook of theory and research* (pp. 249–290). Dordrecht, the Netherlands: Springer.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698–714.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. New York, NY: Oxford University Press.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs*. Washington, DC: American Educational Research Association.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth Cengage Learning.

Steiner, P. M., Wroblewski, A., & Cook, T. D. (2009). Randomized experiments and quasi-experimental designs in educational research. In K. E. Ryan & J. B. Cousins, (Ed.), *The Sage handbook of educational evaluation* (pp. 75–95). Thousand Oaks, CA: Sage.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology*, *51*(6), 309–317.

Trochim, W. M. (1984). *Research design for program evaluation: The regression-discontinuity approach*. Beverly Hills, CA: Sage.

U.S. Department of Education. (2011). *What Works Clearinghouse: Procedures and standards handbook* (version 2.1). Washington, DC: Author. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf

Walpole, M. (2003). Socioeconomic status and college: How SES affects college experiences and outcomes. *The Review of Higher Education*, *27*(1), 45–73.

ALLYSON FLASTER is a doctoral candidate and research assistant in the Center for the Study of Higher and Postsecondary Education at the University of Michigan.

STEPHEN L. DESJARDINS is professor in the Center for the Study of Higher and Postsecondary Education and professor in the Gerald R. Ford School of Public Policy at the University of Michigan.