Published in final edited form as:

Educ Psychol. 2016 ; 51(3-4): 395–405. doi:10.1080/00461520.2016.1207177.

Quasi-Experimental Designs for Causal Inference

Yongnam Kim and Peter Steiner

Department of Educational Psychology, University of Wisconsin–Madison

Abstract

When randomized experiments are infeasible, quasi-experimental designs can be exploited to evaluate causal treatment effects. The strongest quasi-experimental designs for causal inference are regression discontinuity designs, instrumental variable designs, matching and propensity score designs, and comparative interrupted time series designs. This article introduces for each design the basic rationale, discusses the assumptions required for identifying a causal effect, outlines methods for estimating the effect, and highlights potential validity threats and strategies for dealing with them. Causal estimands and identification results are formalized with the potential outcomes notations of the Rubin causal model.

Causal inference plays a central role in many social and behavioral sciences, including psychology and education. But drawing valid causal conclusions is challenging because they are warranted only if the study design meets a set of strong and frequently untestable assumptions. Thus, studies aiming at causal inference should employ designs and design elements that are able to rule out most plausible threats to validity. Randomized controlled trials (RCTs) are considered as the gold standard for causal inference because they rely on the fewest and weakest assumptions. But under certain conditions quasi-experimental designs that lack random assignment can also be as credible as RCTs (Shadish, Cook, & Campbell, 2002).

This article discusses four of the strongest quasi-experimental designs for identifying causal effects: regression discontinuity design, instrumental variable design, matching and propensity score designs, and the comparative interrupted time series design. For each design we outline the strategy and assumptions for identifying a causal effect, address estimation methods, and discuss practical issues and suggestions for strengthening the basic designs. To highlight the design differences, throughout the article we use a hypothetical example with the following causal research question: What is the effect of attending a summer science camp on students' science achievement?

POTENTIAL OUTCOMES AND RANDOMIZED CONTROLLED TRIAL

Before we discuss the four quasi-experimental designs, we introduce the potential outcomes notation of the Rubin causal model (RCM) and show how it is used in the context of an RCT. The RCM (Holland, 1986) formalizes causal inference in terms of potential outcomes,

which allow us to precisely define causal quantities of interest and to explicate the assumptions required for identifying them. RCM considers a potential outcome for each possible treatment condition. For a dichotomous treatment variable (i.e., a treatment and control condition), each subject i has a potential treatment outcome $Y_i(1)$, which we would observe if subject i receives the treatment ($Z_i=1$), and a potential control outcome $Y_i(0)$, which we would observe if subject i receives the control condition ($Z_i=0$). The difference in the two potential outcomes, $Y_i(1) - Y_i(0)$, represents the individual causal effect.

Suppose we want to evaluate the effect of attending a summer science camp on students' science achievement score. Then each student has two potential outcomes: a potential control score for not attending the science camp, and the potential treatment score for attending the camp. However, the individual causal effects of attending the camp cannot be inferred from data, because the two potential outcomes are never observed simultaneously. Instead, researchers typically focus on average causal effects. The average treatment effect (ATE) for the entire study population is defined as the difference in the expected potential outcomes, $ATE = E[Y_i(1)] - E[Y_i(0)]$. Similarly, we can also define the ATE for the treated subjects (ATT), $ATT = E[Y_i(1) | Z_i = 1] - E[Y_i(0) | Z_i = 1]$. Although the expectations of the potential outcomes are not directly observable because not all potential outcomes are observed, we nonetheless can identify ATE or ATT under some reasonable assumptions. In an RCT, random assignment establishes *independence* between the potential outcomes and the treatment status, which allows us to infer ATE. Suppose that students are randomly assigned to the science camp and that all students comply with the assigned condition. Then random assignment guarantees that the camp attendance indicator Z is independent of the potential achievement scores $Y_i(0)$ and $Y_i(1)$.

The independence assumption allows us to rewrite ATE in terms of observable expectations (i.e., with observed outcomes instead of potential outcomes). First, due to the independence (randomization), the unconditional expectations of the potential outcomes can be expressed as conditional expectations, $E[Y_i(1)] = E[Y_i(1) | Z_i = 1]$ and $E[Y_i(0)] = E[Y_i(0) | Z_i = 0]$. Second, because the potential treatment outcomes are actually observed for the treated, we can replace the potential treatment outcome with the observed outcome such that $E[Y_i(1) | Z_i = 1] = E[Y_i | Z_i = 1]$ and, analogously, $E[Y_i(0) | Z_i = 0] = E[Y_i | Z_i = 0]$. Thus, the ATE is expressible in terms of observable quantities rather than potential outcomes, $ATE = E[Y_i(1)] - E[Y_i(0)] = E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0]$, and we say ATE is identified.

This derivation also rests on the *stable-unit-treatment-value assumption* (SUTVA; Imbens & Rubin, 2015). SUTVA is required to properly define the potential outcomes, that is, (a) the potential outcomes of a subject depend neither on the assignment mode nor on other subjects' treatment assignment, and (b) there is only one unique treatment and one unique control condition. Without further mentioning, we assume SUTVA for all quasi-experimental designs discussed in this article.

REGRESSION DISCONTINUITY DESIGN

Due to ethical or budgetary reasons, random assignment is often infeasible in practice. Nonetheless, researchers may sometimes still retain full control over treatment assignment

as in a regression discontinuity (RD) design where, based on a continuous assignment variable and a cutoff score, subjects are deterministically assigned to treatment conditions.

Suppose that the science camp is a remedial program and only students whose grade point average (GPA) score is less than or equal to 2.0 are eligible to participate. Figure 1 shows a scatterplot of hypothetical data where the x-axis represents the assignment variable (*GPA*) and the y-axis the outcome (*Science Score*). All subjects with a GPA score below the cutoff attended the camp (circles), whereas all subjects scoring above the cutoff do not attend (squares). Because all low-achieving students are in the treatment group and all high-achieving students in the control group, their respective GPA distributions do not overlap, not even at the cutoff. This lack of overlap complicates the identification of a causal effect because students in the treatment and control group are not comparable at all (i.e., they have a completely different distribution of the GPA scores).

One strategy of dealing with the lack of overlap is to rely on the linearity assumption of regression models and to extrapolate into areas of nonoverlap. However, if the linear models do not correctly specify the functional form, the resulting ATE estimate is biased. A safer strategy is to evaluate the treatment effect only at the cutoff score where treatment and control cases almost overlap, and thus functional form assumptions and extrapolation are almost no longer needed. Consider the treatment and control students that score right at the cutoff or just above it. Students with a GPA score of 2.0 participate in the science camp and students with a GPA score of 2.1 are in the control condition (the status quo condition or a different camp). The two groups of students are essentially equivalent because the difference in their GPA scores is negligibly small ($2.1 - 2.0 = .1$) and likely due to random chance (measurement error) rather than a real difference in ability. Thus, in the very close neighborhood around the cutoff score, the RD design is equivalent to an RCT; therefore, the ATE at the cutoff (ATEC) is identified.

CAUSAL ESTIMAND AND IDENTIFICATION

ATEC is defined as the difference in the expected potential treatment and control outcomes for the subjects scoring exactly at the cutoff: $ATEC = E[Y_i(1) | A_i = a_c] - E[Y_i(0) | A_i = a_c]$, where A denotes assignment variable and a_c the cutoff score. Because we observe only treatment subjects and not control subjects right at the cutoff, we need two assumptions in order to identify ATEC (Hahn, Todd, & van Klauw, 2001): (a) the conditional expectations of the potential treatment and control outcomes are continuous at the cutoff (*continuity*), and (b) all subjects comply with treatment assignment (*full compliance*).

The continuity assumption can be expressed in terms of limits as

$$\lim_{a \downarrow a_c} E[Y_i(1) | A_i = a] = E[Y_i(1) | A_i = a_c] = \lim_{a \uparrow a_c} E[Y_i(1) | A_i = a] \text{ and}$$

$$\lim_{a \downarrow a_c} E[Y_i(0) | A_i = a] = E[Y_i(0) | A_i = a_c] = \lim_{a \uparrow a_c} E[Y_i(0) | A_i = a].$$

Thus, we can rewrite ATEC as

the difference in limits, $ATEC = \lim_{a \uparrow a_C} E[Y_i(1) | A_i = a_c] - \lim_{a \downarrow a_C} E[Y_i(0) | A_i = a_c]$, which solves the issue that no control subjects are observed directly at the cutoff. Then, by the full compliance assumption, the potential treatment and control outcomes can be replaced with the observed outcomes such that $ATEC = \lim_{a \uparrow a_C} E[Y_i | A_i = a_c] - \lim_{a \downarrow a_C} E[Y_i | A_i = a_c]$ is identified at the cutoff (i.e., ATEC is now expressed in terms of observable quantities). The difference in the limits represents the discontinuity in the mean outcomes exactly at the cutoff (Figure 1).

Estimating ATEC

ATEC can be estimated with parametric or nonparametric regression methods. First, consider the parametric regression of the outcome Y on the treatment Z , the cutoff-centered assignment variable $A - a_c$, and their interaction: $Y = \beta_0 + \beta_1 Z + \beta_2 (A - a_c) + \beta_3 (Z \times (A - a_c)) + e$. If the model correctly specifies the functional form, then $\hat{\beta}_1$ is an unbiased estimator for ATEC. In practice, an appropriate model specification frequently involves also quadratic and cubic terms of the assignment variable plus their interactions with the treatment indicator.

To avoid overly strong functional form assumptions, semiparametric or nonparametric regression methods like generalized additive models or local linear kernel regression can be employed (Imbens & Lemieux, 2008). These methods down-weight or even discard observations that are not in the close neighborhood around the cutoff. The R packages *rdd* (Dimmery, 2013) and *rdrobust* (Calonico, Cattaneo, & Titiunik, 2015), or the command *rd* in STATA (Nichols, 2007) are useful for estimation and diagnostic purposes.

Practical Issues

A major validity threat for RD designs is the manipulation of the assignment score around the cutoff, which directly results in a violation of the *continuity* assumption (Wong et al., 2012). For instance, if a teacher knows the assignment score in advance and he wants all his students to attend the science camp, the teacher could falsely report a GPA score of 2.0 or below for the students whose actual GPA score exceeds the cutoff value.

Another validity threat is noncompliance, meaning that subjects assigned to the control condition may cross over to the treatment and subjects assigned to the treatment do not show up. An RD design with noncompliance is called a *fuzzy* RD design (instead of a *sharp* RD design with full compliance). A fuzzy RD design still allows us to identify the intention-to-treat effect or the local average treatment effect at the cutoff (LATEC). The intention-to-treat effect refers to the effect of treatment assignment rather than the actual treatment receipt. LATEC estimates ATEC for the subjects who comply with treatment assignment. LATEC is identified if one uses the assignment status as an instrumental variable for treatment receipt (see the upcoming Instrumental Variable section).

Finally, generalizability and statistical power are often mentioned as major disadvantages of RD designs. Because RD designs identify the treatment effect only at the cutoff, ATEC estimates are not automatically generalizable to subjects scoring further away from the cutoff. Statistical power for detecting a significant effect is an issue because the lack of overlap on the assignment variable results in increased standard errors. With semi- or nonparametric regression methods, power further diminishes.

Strengthening RD Designs

To avoid systematic manipulations of the assignment variable, it is desirable to conceal the assignment rule from study participants and administrators. If the assignment rule is known to them, manipulations can hardly be ruled out, particularly when the stakes are high. Researchers can use the McCrary test (McCrary, 2008) to check for potential manipulations. The test investigates whether there is a discontinuity in the distribution of the assignment variable right at the cutoff. Plotting baseline covariates against the assignment variable, and regressing the covariates on the assignment variable and the treatment indicator also help in detecting potential discontinuities at the cutoff.

The RD design's validity can be increased by combining the basic RD design with other designs. An example is the tie-breaking RD design, which uses two cutoff scores. Subjects scoring between the two cutoff scores are randomly assigned to treatment conditions, whereas subjects scoring outside the cutoff interval receive the treatment or control condition according to the RD assignment rule (Black, Galdo & Smith, 2007). This design combines an RD design with an RCT and is advantageous with respect to the correct specification of the functional form, generalizability, and statistical power. Similar benefits can be obtained by adding pretest measures of the outcome or nonequivalent comparison groups (Wing & Cook, 2013).

Imbens and Lemieux (2008) and Lee and Lemieux (2010) provided comprehensive introductions to RD designs. Lee and Lemieux also summarized many applications from economics. Angrist and Lavy (1999) applied the design to investigate the effect of class size on student achievement.

INSTRUMENTAL VARIABLE DESIGN

In practice, researchers often have no or only partial control over treatment selection. In addition, they might also lack reliable knowledge of the selection process. Nonetheless, even with limited control and knowledge of the selection process it is still possible to identify a causal treatment effect if an instrumental variable (IV) is available. An IV is an exogenous variable that is related to the treatment but is completely unrelated to the outcome, except via treatment. An IV design requires researchers either to create an IV at the design stage (as in an encouragement design; see next) or to find an IV in the data set at hand or a related data base.

Consider the science camp example, but instead of random or deterministic treatment assignment, students decide on their own or together with their parents whether to attend the camp. Many factors may determine the decision, for instance, students' science ability and

motivation, parents' socioeconomic status, or the availability of public transportation for the daily commute to the camp. Whereas the first three variables are presumably also related to the science outcome, public transportation might be unrelated to the science score (except via camp attendance). Thus, the availability of public transportation may qualify as an IV. Figure 2 illustrates such IV design: Public transportation (IV) directly affects camp attendance but has no direct or indirect effect on science achievement (outcome) other than through camp attendance (treatment). The question mark represents unknown or unobserved confounders, that is, variables that simultaneously affect both camp attendance and science achievement. The IV design allows us to identify a causal effect even if some or all confounders are unknown or unobserved.

The strategy for identifying a causal effect is based on exploiting the variation in the treatment variable explained by IV. In Figure 2, the total variation in the treatment consists of (a) the variation induced by the IV and (b) the variation induced by confounders (question mark) and other exogenous variables (not shown in the figure). The identification of the camp's effect requires us to isolate the treatment variation that is related to public transportation (IV), and then to use the isolated variation to investigate the camp's effect on the science score. Because we exploit the treatment variation exclusively induced by the IV but ignore the variation induced by unobserved or unknown confounders, the IV design identifies the ATE for the sub-population of compliers only. In our example, the compliers are the students who attend the camp because public transportation is available and do not attend because it is unavailable. For students whose parents always use their own car to drop them off and pick them up at the camp location, we cannot infer the causal effect, because their camp attendance is completely unrelated to the availability of public transportation.

Causal Estimand and Identification

The complier average treatment effect (CATE) is defined as the expected difference in potential outcomes for the sub-population of compliers: $CATE = E[Y(1) | Complier] - E[Y(0) | Complier] = \tau_C$.

Identification requires us to distinguish between four latent groups: compliers (C), who attend the camp if public transportation is available but do not attend if unavailable; always-takers (A), who always attend the camp regardless of whether or not public transportation is available; never-takers (N), who never attend the camp regardless of public transportation; and defiers (D), who do not attend if public transportation is available but attend if unavailable. Because group membership is unknown, it is impossible to directly infer CATE from the data of compliers. However, CATE is identified from the entire data set if (a) the IV is predictive of the treatment (*predictive first stage*), (b) the IV is unrelated to the outcome except via treatment (*exclusion restriction*), and (c) no defiers are present (*monotonicity*; Angrist, Imbens, & Rubin, 1996; see Steiner, Kim, Hall, & Su, 2015, for a graphical explanation).

First, notice that the IV's effects on the treatment (γ) and the outcome (δ) are directly identified from the observed data because the IV's relation with the treatment and outcome is unconfounded. In our example (Figure 2), γ denotes the effect of public transportation on camp attendance and δ the indirect effect of public transportation on the science score. Both

effects can be written as weighted averages of the corresponding group-specific effects (γ_C , γ_A , γ_N , γ_D and δ_C , δ_A , δ_N , δ_D for compliers, always-takers, never-takers, and defiers, respectively): $\gamma = p(C)\gamma_C + p(A)\gamma_A + p(N)\gamma_N + p(D)\gamma_D$ and $\delta = p(C)\delta_C + p(A)\delta_A + p(N)\delta_N + p(D)\delta_D$ where $p(\cdot)$ represents the portion of the respective latent group in the population and $p(C) + p(A) + p(N) + p(D) = 1$. Because the treatment choice of always-takers and never-takers is entirely unaffected by the instrument, the IV's effect on the treatment is zero, $\gamma_A = \gamma_N = 0$, and together with the *exclusion restriction*, we also know $\delta_A = \delta_N = 0$, that is, the IV has no effect on the outcome. If no defiers are present, $p(D) = 0$ (*monotonicity*), then the IV's effects on the treatment and outcome simplify to $\gamma = p(C)\gamma_C$ and $\delta = p(C)\delta_C$, respectively. Because $\delta_C = \gamma_C\tau_C$ and $\gamma \neq 0$ (*predictive first stage*), the ratio of the observable IV effects, γ and δ , identifies CATE: $\frac{\delta}{\gamma} = \frac{p(C)\gamma_C\tau_C}{p(C)\gamma_C} = \tau_C$.

Estimating CATE

A two-stage least squares (2SLS) regression is typically used for estimating CATE. In the first stage, treatment Z is regressed on the IV, $Z = \beta_0 + \beta_1 IV + e$. The linear first-stage model applies with a dichotomous treatment variable (linear probability model). The second stage then regresses the outcome Y on the predicted values $\hat{Z}$ from the first stage model, $Y = \pi_0 + \pi_1 \hat{Z} + r$, where $\hat{\pi}_1$ is the CATE estimator. The two stages are automatically performed by the 2SLS procedure, which also provides an appropriate standard error for the effect estimate. The STATA commands *ivregress* and *ivreg2* (Baum, Schaffer, & Stillman, 2007) or the *sem* package in R (Fox, 2006) perform the 2SLS regression.

Practical Issues

One challenge in implementing an IV design is to find a valid instrument that satisfies the assumptions just discussed. In particular, the exclusion restriction is untestable and frequently hard to defend in practice. In our example, if high-income families live in suburban areas with bad public transportation connections, then the availability of the public transportation is likely related to the science score via household income (or socioeconomic status). Although conditioning on the observed household income can transform public transportation into a conditional IV (see next), one can frequently come up with additional scenarios that explains why the IV is related to the outcome and thus violates the exclusion restriction.

Another issue arises from “weak” IVs that are only weakly related to treatment. Weak IVs cause efficiency problems (Wooldridge, 2012). If the availability of public transportation barely affects camp attendance because most parents give their children a ride anyway, the IV's effect on the treatment (γ) is close to zero. Because $\hat{\gamma}$ is the denominator in the CATE estimator, $\hat{\tau}_C = \hat{\delta}/\hat{\gamma}$, an imprecisely estimated $\hat{\gamma}$ results in a considerable over- or underestimation of CATE. Moreover, standard errors will be large.

One also needs to keep in mind that the substantive meaning of CATE depends on the chosen IV. Consider two slightly different IVs with respect to public transportation: the availability of (a) a bus service and (b) subway service. For the first IV, the complier

population consists of students who choose to (not) attend the camp depending on the availability of a bus service. For the second IV, the complier population refers to the availability of a subway service. Because the two complier populations are very likely different from each other (students who are willing to take the subway might not be willing to take the bus), the corresponding CATEs refer to different subpopulations.

Strengthening IV Designs

Given the challenges in identifying a valid instrument from observed data, researchers should consider creating an IV at the design stage of a study. Although it might be impossible to directly assign subjects to treatment conditions, one might still be able to *encourage* participants to take the treatment. Subjects are randomly encouraged to sign up for treatment, but whether they actually comply with the encouragement is entirely their own decision (Imai et al., 2011). Random encouragement qualifies as an IV because it very likely meets the exclusion restriction. For example, instead of collecting data on public transportation, researchers may advertise and recommend the science camp in a letter to the parents of a randomly selected sample of students.

With observational data it is hard to identify a valid IV because covariates that strongly predict the treatment are usually also related to the outcome. However, these covariates can still qualify as an IV if they affect the outcome only indirectly via other observed variables. Such covariates can be used as conditional IVs, that is, they meet the IV requirements conditional on the observed variables (Brito & Pearl, 2002). Assume the availability of public transportation (IV) is associated with the science score via household income. Then, controlling for the reliably measured household income in both stages of the 2SLS analysis blocks the IV's relation to the science score and turns public transportation into a conditional IV. However, controlling for a large set of variables does not guarantee that the exclusion restriction is more likely met. It may even result in more bias as compared to an IV analysis with fewer covariates (Ding & Miratrix, 2015; Steiner & Kim, in press). The choice of a valid conditional IV requires researchers to carefully select the control variables based on subject-matter theory.

The seminal article by Angrist et al. (1996) provides a thorough discussion of the IV design, and Steiner, Kim, et al. (2015) proved the identification result using graphical models. Excellent introductions to IV designs can be found in Angrist and Pischke (2009, 2015). Angrist and Krueger (1992) is an example of a creative application of the design with birthday as the IV. For encouragement designs, see Holland (1988) and Imai et al. (2011).

MATCHING AND PROPENSITY SCORE DESIGN

This section considers quasi-experimental designs in which researchers lack control over treatment selection but have good knowledge about the selection mechanism or at least the confounders that simultaneously determine the treatment selection and the outcome. Due to self or third-person selection of subjects into treatment, the resulting treatment and control groups typically differ in observed but also unobserved baseline covariates. If we have reliable measures of all confounding covariates, then matching or propensity score (PS) designs balance groups on observed baseline covariates and thus enable the identification of

causal effects (Imbens & Rubin, 2015). Regression analysis and the analysis of covariance can also remove the confounding bias, but because they rely on functional form assumptions and extrapolation we discuss only nonparametric matching and PS designs.

Suppose that students decide on their own whether to attend the science camp. Although many factors can affect students' decision, teachers with several years of experience of running the camp may know that selection is mostly driven by students' science ability, liking of science, and their parents' socioeconomic status. If all the selection-relevant factors that also affect the outcome are known, the question mark in Figure 2 can be replaced by the known confounding covariates.

Given the set of confounding covariates, causal inference with matching or PS designs is straightforward, at least theoretically. The basic one-to-one matching design matches each treatment subject to a control subject that is equivalent or at least very similar in observed covariates. To illustrate the idea of matching, consider a camp attendee with baseline measures of 80 on the science pre-test, 6 on liking science, and 50 on the socioeconomic status. Then a multivariate matching strategy tries to find a nonattendee with exactly the same or at least very similar baseline measures. If we succeed in finding close matches for all camp attendee, the matched samples of attendees and nonattendees will have almost identical covariate distributions.

Although multivariate matching works well when the number of confounders is small and the pool of control subjects is large relative to the number of treatment subjects, it is usually difficult to find close matches with a large set of covariates or a small pool of control subjects. Matching on the PS helps to overcome this issue because the PS is a univariate score computed from the observed covariates (Rosenbaum & Rubin, 1983). The PS is formally defined as the conditional probability of receiving the treatment given the set of observed covariates $\mathbf{X}$: $PS = \Pr(Z = 1 \mid \mathbf{X})$.

Causal Estimand and Identification

Matching and PS designs usually investigate $ATE = E[Y_i(1)] - E[Y_i(0)]$ or $ATT = E[Y_i(1) \mid Z_i = 1] - E[Y_i(0) \mid Z_i = 1]$. Both causal effects are identified if (a) the potential outcomes are statistically independent of the treatment indicator given the set of observed confounders $\mathbf{X}$, $\{Y(1), Y(0)\} \perp Z \mid \mathbf{X}$ (*unconfoundedness*; $\perp$ denotes independence), and (b) the treatment probability is strictly between zero and one, $0 < \Pr(Z = 1 \mid \mathbf{X}) < 1$ (*positivity*).

By the positivity assumption we get $E[Y_i(1)] = E_X[E[Y_i(1) \mid \mathbf{X}]]$ and $E[Y_i(0)] = E_X[E[Y_i(0) \mid \mathbf{X}]]$. If the unconfoundedness assumption holds, we can write the inner expectations as $E[Y_i(1) \mid \mathbf{X}] = E[Y_i(1) \mid Z_i = 1; \mathbf{X}]$ and $E[Y_i(0) \mid \mathbf{X}] = E[Y_i(0) \mid Z_i = 0; \mathbf{X}]$. Finally, because the treatment (control) outcomes of the treatment (control) subjects are actually observed, ATE is identified because it can be expressed in terms of observable quantities: $ATE = E_X[E[Y_i \mid Z_i = 1; \mathbf{X}]] - E_X[E[Y_i \mid Z_i = 0; \mathbf{X}]]$. The same can be shown for ATT. The unconfoundedness and positivity assumption are frequently referred to jointly as the *strong ignorability* assumption. Rosenbaum and Rubin (1983) proved that if the assignment is strongly ignorable given $\mathbf{X}$, then it is also strongly ignorable given the PS alone.

Estimating ATE and ATT

Matching designs use a distance measure for matching each treatment subject to the closest control subject. The Mahalanobis distance is usually used for multivariate matching and the Euclidean distance on the logit of the PS for PS matching. Matching strategies differ with respect to the matching ratio (one-to-one or one-to-many), replacement of matched subjects (with or without replacement), use of a caliper (treatment subjects that do not have a control subject within a certain threshold remain unmatched), and the matching algorithm (greedy, genetic, or optimal matching; Sekhon, 2011; Steiner & Cook, 2013). Because we try to find at least one control subject for each treatment subject, matching estimators typically estimate ATT. Once treatment and control subjects are matched, ATT is computed as the difference in the mean outcome of the treatment and control group. An alternative matching strategy that allows for estimating ATE is full matching, which stratifies all subjects into the maximum number of strata, where each stratum contains at least one treatment and one control subject (Hansen, 2004).

The PS can also be used for PS stratification and inverse-propensity weighting. PS stratification stratifies the treatment and control subjects into at least five strata and estimates the treatment effect within each stratum. ATE or ATT is then obtained as the weighted average of the stratum-specific treatment effects. Inverse-propensity weighting follows the same logic as inverse-probability weighting in survey research (Horvitz & Thompson, 1952) and requires the computation of weights that refer to either the overall population (ATE) or the population of treated subjects only (ATT). Given the inverse-propensity weights, ATE or ATT is usually estimated via weighted least squares regression.

Because the true PSs are unknown, they need to be estimated from the observed data. The most common method for estimating the PS is logistic regression, which regresses the binary treatment indicator Z on predictors of the observed covariates. The PS model is specified according to balance criteria (instead of goodness of fit criteria), that is, the estimated PSs should remove all baseline differences in observed covariates (Imbens & Rubin, 2015). The predicted probabilities from the PS model represent the estimated PSs.

All three PS designs—matching, stratification, and weighting—can benefit from additional covariance adjustments in an outcome regression. That is, for the matched, stratified or weighted data, the outcome is regressed on the treatment indicator and the additional covariates. Combining the PS design with a covariance adjustment gives researchers two chances to remove the confounding bias, by correctly specifying either the PS model or the outcome model. These combined methods are said to be doubly robust because they are robust against either the misspecification of the PS model or the misspecification of the outcome model (Robins & Rotnitzky, 1995). The R packages *optmatch* (Hansen & Klopfer, 2006) and *MatchIt* (Ho et al., 2011) and the STATA command *teffects*, in particular *teffects psmatch* (StataCorp, 2015), can be useful for matching or PS analyses.

Practical Issues

The most challenging issue with matching and PS designs is the selection of covariates for establishing unconfoundedness. Ideally, subject-matter theory about the selection process

and the outcome-generating model is used for selecting a set of covariates that removes all the confounding (Pearl, 2009). If strong subject-matter theories are not available, selecting the right covariates is difficult. In the hope to remove a major part of the confounding bias—if not all of it—a frequently applied strategy is to match on as many covariates as possible. However, recent literature shows that thoughtless inclusion of covariates may increase rather than reduce the confounding bias (Pearl, 2010; Steiner & Kim, in press). The risk of increasing bias can be reduced if the observed covariates cover a broad range of heterogeneous construct domains, including at least one reliable pretest measure of the outcome (Steiner, Cook, et al., 2015). Besides having the right covariates, they also need to be reliably measured. The unreliable measurement of confounding covariates has a similar effect as the omission of a confounder: It results in a violation of the unconfoundedness assumption and thus in a biased effect estimate (Steiner, Cook, & Shadish, 2011; Steiner & Kim, in press).

Even if the set of reliably measured covariates establishes unconfoundedness, we still need to correctly specify the functional form of the PS model. Although parametric models like logistic regression, including higher order terms, might frequently approximate the correct functional form, they still rely on the linearity assumption. The linearity assumption can be relaxed if one estimates the PS with statistical learning algorithms like classification trees, neural networks, or the LASSO (Keller, Kim, & Steiner, 2015; McCaffrey, Ridgeway, & Morral, 2004).

Strengthening Matching and PS Designs

The credibility of matching and PS designs heavily relies on the unconfoundedness assumption. Although empirically untestable, there are indirect ways for assessing unconfoundedness. First, unaffected (nonequivalent) outcomes that are known to be unaffected by the treatment can be used (Shadish et al., 2002). For instance, we may expect that attendance in the science camp does not significantly affect the reading score. Thus, if we observe a significant group difference in the reading score after the PS adjustment, bias due to unobserved confounders (e.g., general intelligence) is still likely. Second, adding a second but conceptually different control group allows for a similar test as with the unaffected outcome (Rosenbaum, 2002).

Because researchers rarely know whether the unconfoundedness assumption is actually met with the data at hand, it is important to assess the effect estimate's sensitivity to potentially unobserved confounders. Sensitivity analyses investigate how strongly an estimate's magnitude and significance changes if a confounder of a certain strength would have been omitted from the analyses. Causal conclusions are much more credible if the effect's direction, magnitude, and significance is rather insensitive to omitted confounders (Rosenbaum, 2002). However, despite the value of sensitivity analyses, they are not informative about whether hidden bias is actually present.

Schafer and Kang (2008) and Steiner and Cook (2013) provided a comprehensive introduction. Rigorous formalization and technical details of PS designs can be found in Imbens and Rubin (2015). Rosenbaum (2002) discussed many important design issues in these designs.

COMPARATIVE INTERRUPTED TIME SERIES DESIGN

The designs discussed so far require researchers to have either full control over treatment assignment or reliable knowledge of the exogenous (IV) or endogenous part of the selection mechanism (i.e., the confounders). If none of these requirements are met, a comparative interrupted time series (CITS) design might be a viable alternative if (a) multiple measurements of the outcome (*time series*) are available for both the treatment and a comparison group and (b) the treatment group's time series has been interrupted by an intervention.

Suppose that all students of one class in a school (say, an advanced science class) attend the camp, whereas all students of another class in the same school do not attend. Also assume that monthly measures of science achievement before and after the science camp are available. Figure 3 illustrates such a scenario where the x -axis represents time in *Months* and the y -axis the *Science Score* (aggregated at the class level). The filled symbols indicate the treatment group (science camp), open symbols the comparison group (no science camp). The science camp intervention divides both time series into a preintervention time series (circles) and a postintervention time series (squares). The changes in the levels and slopes of the pre- and postintervention regression lines represent the camp's impact but possibly also the effect of other events that co-occur with the intervention. The dashed lines extrapolate the preintervention growth curves into the postintervention period, and thus represent the counterfactual situation where the intervention but also other co-occurring events are absent.

The strength of a CITS design is its ability to discriminate between the intervention's effect and the effects of co-occurring events. Such events might be other potentially competing interventions (history effects) or changes in the measurement of the outcome (instrumentation), for instance. If the co-occurring events affect the treatment and comparison group to the same extent, then subtracting the changes in the comparison group's growth curve from the changes in the treatment group's growth curve provides a valid estimate of the intervention's impact. Because we investigate the difference in the changes (= differences) of the two growth curves, the CITS design is a special case of the difference-in-differences design (Somers et al., 2013).

Assume that a daily TV series about Albert Einstein was broadcast in the evenings of the science camp week and that students of both classes were exposed to the same extent to the TV series. It follows that the comparison group's change in the growth curve represents the TV series' impact. The comparison group's time series in Figure 3 indicates that the TV series might have had an immediate impact on the growth curve's level but almost no effect on the slope. On the other hand, the treatment group's change in the growth curve is due to both the science camp and the TV series. Thus, in differencing out the TV series' effect (estimated from the comparison group) we can identify the camp effect.

Causal Estimand and Identification

Let t_c denote the time point of the intervention, then the intervention's effect on the treated (ATT) at a postintervention time point $t > t_c$ is defined as $\tau_t = E[Y_{it}^T(1)] - E[Y_{it}^T(0)]$, where

$Y_{it}^T(0)$ and $Y_{it}^T(1)$ are the potential control and treatment outcomes of subject i in the treatment group (T) at time point t . The time series of the expected potential outcomes can be formalized as sum of nonparametric but additive time-dependent functions. The treatment group's expected potential control outcome can be represented as $E[Y_{it}^T(0)] = f_0^T(t) + f_E^T(t)$, where the control function $f_0^T(t)$ generates the expected potential control outcomes in absence of any interventions (I) or co-occurring events (E), and the event function $f_E^T(t)$ adds the effects of co-occurring events. Similarly, the expected potential treatment outcome can be written as $E[Y_{it}^T(1)] = f_0^T(t) + f_E^T(t) + f_I^T(t)$, which adds the intervention's effect $\tau_t = f_I^T(t)$ to the control and event function. In the absence of a comparison group, we can try to identify the impact of the intervention by comparing the observable postintervention outcomes to the extrapolated outcomes from the preintervention time series (dashed line in Figure 3). Extrapolation is necessary because we do not observe any potential control outcomes in the postintervention period (only potential treatment outcomes are observed). Let $\hat{f}_0^T(t)$ denote the parametric extrapolation of the preintervention control function $f_0^T(t)$, then the observable pre-post-intervention difference (PP^T) in the expected control outcome is $PP_t^T = f_0^T(t) + f_E^T(t) + f_I^T(t) - \hat{f}_0^T(t) = f_I^T(t) + (f_0^T(t) - \hat{f}_0^T(t)) + f_E^T(t)$. Thus, in the absence of a comparison group, ATT is identified (i.e., $PP_t^T = f_I^T(t) = \tau_t$) only if the control function is correctly specified ($f_0^T(t) = \hat{f}_0^T(t)$) and if no co-occurring events are present ($f_E^T(t) = 0$).

The comparison group in a CITS design allows us to relax both of these identifying assumptions. In order to see this, we first define the expected control outcomes of the comparison group (C) as a sum of two time-dependent functions as before: $E[Y_{it}^C(0)] = f_0^C(t) + f_E^C(t)$. Then, in extrapolating the comparison group's preintervention function into the postintervention period, $\hat{f}_0^C(t)$, we can compute the pre-post-intervention difference for the comparison group: $PP_t^C = f_0^C(t) + f_E^C(t) - \hat{f}_0^C(t) = f_E^C(t) + (f_0^C(t) - \hat{f}_0^C(t))$. If the control function is correctly specified $f_0^C(t) = \hat{f}_0^C(t)$, the effect of co-occurring events is identified $PP_t^C = f_E^C(t)$. However, we do not necessarily need a correctly specified control function, because in a CITS design we focus on the difference in the treatment and comparison group's pre-post-intervention differences, that is, $PP_t^T - PP_t^C = f_I^T(t) + \{(f_0^T(t) - \hat{f}_0^T(t)) - (f_0^C(t) - \hat{f}_0^C(t))\} + \{f_E^T(t) - f_E^C(t)\}$. Thus, ATT is identified, $PP_t^T - PP_t^C = f_I^T(t) = \tau_t$, if (a) both control functions are either correctly specified or misspecified to the same additive extent such that $(f_0^T(t) - \hat{f}_0^T(t)) = (f_0^C(t) - \hat{f}_0^C(t))$ (*no differential misspecification*) and (b) the effect of co-occurring events is identical in the treatment and comparison group, $f_E^T(t) = f_E^C(t)$ (*no differential event effects*).

Estimating ATT

CITS designs are typically analyzed with linear regression models that regress the outcome Y on the centered time variable $(T - t_c)$, the intervention indicator Z ($Z = 0$ if $t < t_c$, otherwise $Z = 1$), the group indicator G ($G = 1$ for the treatment group and $G = 0$ for the control group), and the corresponding two-way and three-way interactions:

$$Y = \beta_0 + \beta_1(T - t_c) + \beta_2Z + \beta_3G + \beta_4(Z \times (T - t_c)) + \beta_5(Z \times G) + \beta_6(G \times (T - t_c)) + \beta_7(Z \times G \times (T - t_c)) + e.$$

Depending on the number of subjects in each group, fixed or random effects for the subjects are included as well (time fixed or random effect can also be considered). $\hat{\beta}_5$ estimates the intervention's immediate effect at the onset of the intervention (change in intercept) and $\hat{\beta}_7$ the intervention's effect on the growth rate (change in slope). The inclusion of dummy variables for each postintervention time point (plus their interaction with the intervention and group indicators) would allow for a direct estimation of the time-specific effects. If the time series are long enough (at least 100 time points), then a more careful modeling of the autocorrelation structure via time series models should be considered.

Practical Issues

Compared to other designs, CITS designs heavily rely on extrapolation and thus on functional form assumptions. Therefore, it is crucial that the functional forms of the pre- and postintervention time series (including their extrapolations) are correctly specified or at least not differentially misspecified. With short time series or measurement points that inadequately capture periodical variations, the correct specification of the functional form is very challenging. Another specification aspect concerns serial dependencies among the data points. Failing to model serial dependencies can bias effect estimates and their standard errors such that significance tests might be misleading. Accounting for serial dependencies requires autoregressive models (e.g., ARIMA models), but the time series should have at least 100 time points (West, Biesanz, & Pitts, 2000). Standard fixed effects or random effects models deal at least partially with the dependence structure. Robust standard errors (e.g., Huber-White corrected ones) or the bootstrap can also be used to account for dependency structures.

Events that co-occur with the intervention of interest, like history or instrumentation effects, are a major threat to the time series designs that lack a comparison group (Shadish et al., 2002). CITS designs are rather robust to co-occurring events as long as the treatment and comparison groups are affected to the same additive extent. However, there is no guarantee that both groups are exposed to the same events and affected to the same extent. For example, if students who do not attend the camp are less likely to watch the TV series, its effect cannot be completely differenced out (unless the exposure to the TV series is measured). If one uses aggregated data like class or school averages of achievement scores, then differential compositional shifts over time can also invalidate the CITS design. Compositional shifts occur due to dropouts or incoming subjects over time.

Strengthening CITS Designs

If the treatment and comparison group's preintervention time series are very different (different levels and slopes), then the assumption that history or instrumentation threats affect both groups to the same additive extent may not hold. Matching treatment and comparison subjects prior to the analysis can increase the plausibility of this assumption. Instead of using all nonparticipating students of the comparison class, we may select only those students who have a similar level and growth in the preintervention science scores as the students participating in the camp. We can also match on additional covariates like socioeconomic status or motivation levels. Multivariate or PS matching can be used for this purpose. If the two groups are similar, it is more likely that they are affected by co-occurring events to the same extent.

As with the matching and PS designs, using an unaffected outcome in CITS designs helps to probe the untestable assumptions (Coryn & Hobson, 2011; Shadish et al., 2002). For instance, we might expect that attending the science camp does not affect students' reading scores but that some validity threats (e.g., attrition) operate on both the reading and science outcome. If we find a significant camp effect on the reading score, the validity of the CITS design for evaluating the camp's impact on the science score is in doubt.

Another strategy to avoid validity threats is to control the time point of the intervention if possible. Researchers can wait with the implementation of the treatment until they have enough preintervention measures for reliably estimating the functional form. They can also choose to intervene when threats to validity are less likely (avoiding the week of the TV series). Control over the intervention also allows researchers to introduce and remove the treatment in subsequent time intervals, maybe even with switching replications between two (or more) groups. If the treatment is effective, we expect that the pattern of the intervention scheme is directly reflected in the time series of the outcome (for more details, see Shadish et al., 2002; for the literature on single case designs, see Kazdin, 2011).

A comprehensive introduction to CITS design can be found in Shadish et al. (2002), which also addresses many classical applications. For more technical details of its identification, refer to Lechner (2011). Wong, Cook, and Steiner (2009) evaluated the effect of No Child Left Behind using a CITS design.

CONCLUDING REMARKS

This article discussed four of the strongest quasi-experimental designs for causal inference when randomized experiments are not feasible. For each design we highlighted the identification strategies and the required assumptions. In practice, it is crucial that the design assumptions are met, otherwise biased effect estimates result. Because most important assumptions like the exclusion restriction or the unconfoundedness assumption are not directly testable, researchers should always try to assess their plausibility via indirect tests and investigate the effect estimates' sensitivity to violations of these assumptions.

Our discussion of RD, IV, PS, and CITS designs made it also very clear that, in comparison to RCTs, quasi-experimental designs rely on more or stronger assumptions. With perfect

control over treatment assignment and treatment implementation (as in an RCT), causal inference is warranted by a minimal set of assumptions. But with limited control over and knowledge about treatment assignment and implementation, stronger assumptions are required and causal effects might be identifiable only for local subpopulations. Nonetheless, observational data sometimes meet the assumptions of a quasi-experimental design, at least approximately, such that causal conclusions are credible. If so, the estimates of quasi-experimental designs—which exploit naturally occurring selection processes and real-world implementations of the treatment—are frequently better generalizable than the results from a controlled laboratory experiment. Thus, if external validity is a major concern, the results of randomized experiments should always be complemented by findings from valid quasi-experiments.

REFERENCES

- Angrist JD , Imbens GW , & Rubin DB (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91, 444–455.
- Angrist JD , & Krueger AB (1992). The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, 87, 328–336.
- Angrist JD , & Lavy V (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, 114, 533–575.
- Angrist JD , & Pischke JS (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.
- Angrist JD , & Pischke JS (2015). *Mastering' metrics: The path from cause to effect*. Princeton, NJ: Princeton University Press.
- Baum CF , Schaffer ME , & Stillman S (2007). Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *The Stata Journal*, 7, 465–506.
- Black D , Galdo J , & Smith JA (2007). Evaluating the bias of the regression discontinuity design using experimental data (Working paper). Chicago, IL: University of Chicago.
- Brito C , & Pearl J (2002). Generalized instrumental variables In Darwiche A & Friedman N (Eds.), *Uncertainty in artificial intelligence* (pp. 85–93). San Francisco, CA: Morgan Kaufmann.
- Calonico S , Cattaneo MD , & Titiunik R (2015). *rdrobust: Robust data-driven statistical inference in regression-discontinuity designs* (R package ver. 0.80). Retrieved from <http://CRAN.R-project.org/package=rdrobust>
- Coryn CLS , & Hobson KA (2011). Using nonequivalent dependent variables to reduce internal validity threats in quasi-experiments: Rationale, history, and examples from practice. *New Directions for Evaluation*, 131, 31–39.
- Dimmery D (2013). *rdd: Regression discontinuity estimation* (R package ver. 0.56). Retrieved from <http://CRAN.R-project.org/package=rdd>
- Ding P , & Miratrix LW (2015). To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. *Journal of Causal Inference*, 3(1), 41–57.
- Fox J (2006). Structural equation modeling with the sem package in R. *Structural Equation Modeling*, 13, 465–486.
- Hahn J , Todd P , & Van der Klaauw W (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1), 201–209.
- Hansen BB (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99, 609–618.
- Hansen BB , & Klopfer SO (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15, 609–627.

- Ho D , Imai K , King G , & Stuart EA (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8), 1–28. Retrieved from <http://www.jstatsoft.org/v42/i08/>
- Holland PW (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–960.
- Holland PW (1988). Causal inference, path analysis and recursive structural equations models. ETS Research Report Series. doi:10.1002/j.2330-8516.1988.tb00270.x
- Horvitz DG , & Thompson DJ (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.
- Imai K , Keele L , Tingley D , & Yamamoto T (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105, 765–789.
- Imbens GW , & Lemieux T (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142, 615–635.
- Imbens GW , & Rubin DB (2015). *Causal inference in statistics, social, and biomedical sciences*. New York, NY: Cambridge University Press.
- Kazdin AE (2011). *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.
- Keller B , Kim JS , & Steiner PM (2015). Neural networks for propensity score estimation: Simulation results and recommendations In van der Ark LA , Bolt DM , Chow S-M , Douglas JA , & Wang W-C (Eds.), *Quantitative psychology research* (pp. 279–291). New York, NY: Springer.
- Lechner M (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, 4, 165–224.
- Lee DS , & Lemieux T (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48, 281–355.
- McCaffrey DF , Ridgeway G , & Morral AR (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9, 403–425.15598095
- McCrary J (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142, 698–714.
- Nichols A (2007). rd: Stata modules for regression discontinuity estimation. Retrieved from <http://ideas.repec.org/c/boc/bocode/s456888.html>
- Pearl J (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York, NY: Cambridge University Press.
- Pearl J (2010). On a class of bias-amplifying variables that endanger effect estimates In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* (pp. 425–432). Corvallis, OR: Association for Uncertainty in Artificial Intelligence.
- Robins JM , & Rotnitzky A (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429), 122–129.
- Rosenbaum PR (2002). *Observational studies*. New York, NY: Springer.
- Rosenbaum PR , & Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
- Schafer JL , & Kang J (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, 13, 279–313.19071996
- Sekhon JS (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, 42(7), 1–52.
- Shadish WR , Cook TD , & Campbell DT (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.
- Somers M , Zhu P , Jacob R , & Bloom H (2013). The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation (MDRC working paper in research methodology). New York, NY: MDRC.
- StataCorp. (2015). *Stata treatment-effects reference manual: Potential outcomes/counterfactual outcomes*. College Station, TX: Stata Press Retrieved from <http://www.stata.com/manuals14/te.pdf>

- Steiner PM , & Cook D (2013). Matching and propensity scores In Little T (Ed.), The Oxford handbook of quantitative methods in psychology (Vol. 1, pp. 237–259). New York, NY: Oxford University Press.
- Steiner PM , Cook TD , Li W , & Clark MH (2015). Bias reduction in quasi-experiments with little selection theory but many covariates. *Journal of Research on Educational Effectiveness*, 8, 552–576.
- Steiner PM , Cook TD , & Shadish WR (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, 36, 213–236.
- Steiner PM , & Kim Y (in press). The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. *Journal of Causal Inference*.
- Steiner PM , Kim Y , Hall CE , & Su D (2015). Graphical models for quasi-experimental designs. *Sociological Methods & Research*. Advance online publication. doi:10.1177/0049124115582272
- West SG , Biesanz JC , & Pitts SC (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs In Reis HT & Judd CM (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–84). New York, NY: Cambridge University Press.
- Wing C , & Cook TD (2013). Strengthening the regression discontinuity design using additional design elements: A within-study comparison. *Journal of Policy Analysis and Management*, 32, 853–877.
- Wong M , Cook TD , & Steiner PM (2009). No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series (Working Paper No. WP-09–11). Evanston, IL: Institute for Policy Research, Northwestern University.
- Wong VC , Wing C , Steiner PM , Wong M , & Cook TD (2012). Research designs for program evaluation. *Handbook of Psychology*, 2, 316–341.
- Wooldridge J (2012). *Introductory econometrics: A modern approach* (5th ed.). Mason, OH: South-Western Cengage Learning.

FIGURE 1.

A hypothetical example of regression discontinuity design. *Note.* GPA = grade point average.

FIGURE 2.

A diagram of an example of instrumental variable design.

FIGURE 3.
A hypothetical example of comparative interrupted time series design.