

Chapter 3

Counterfactuals with Experimental and Quasi-Experimental Variation



Erich Battistin and Marco Bertoni

Abstract Inference about the causal effects of a policy intervention requires knowledge of what would have happened to the outcome of the units affected had the policy not taken place. Since this counterfactual quantity is never observed, the empirical investigation of causal effects must deal with a missing data problem. Random variation in the assignment to the policy offers a solution, under some assumptions. We discuss identification of policy effects when participation to the policy is determined by a lottery (randomized designs), when participation is only partially influenced by a lottery (instrumental variation), and when participation depends on eligibility criteria making a subset of participant and non-participant units as good as randomly assigned to the policy (regression discontinuity designs). We offer guidelines for empirical analysis in each of these settings and provide some applications of the methods proposed to the evaluation of education policies.

Learning Objectives

By studying this chapter, you will:

- Learn to speak the language of potential outcomes and counterfactual impact evaluation.
- Grasp different concepts of validity of a research design.
- Understand why randomization helps to detect causal effects.
- Discover how to exploit natural experiments and discontinuities to learn about causality when proper experiments are not feasible.
- Discuss the credibility of the assumption underlying different empirical strategies.

E. Battistin
University of Maryland, College Park, MD, USA
e-mail: ebattist@umd.edu

M. Bertoni (✉)
University of Padova, Padova, Italy
e-mail: marco.bertoni@unipd.it

3.1 Introduction

Do smaller classes yield better school outcomes? To answer this and many similar questions, one needs to compare the outcome in the *status quo* (a large class) to the outcome that would have been observed if the input of interest was set to a different level (a small class). The comparison of students enrolled in small and large classes is always a tempting avenue to answer this causal question. As this comparison involves different students, its validity rests on the assumption that students currently enrolled in small and large classes would have presented the same outcome, on average, had they been exposed to the same number of classmates. This remains an untestable assumption that must be discussed on a case-by-case basis.

The chapter discusses ways to combine policy designs and data to corroborate the validity of this assumption. Sections 3.2 and 3.3 introduce the counterfactual causal analysis talk. They describe the concepts of treatments, potential outcomes and causal effects, and the attributes characterizing the validity of a research design. Section 3.4 is about the beauty and limitations of randomized assignment to “treatment” (e.g., a small class) and paves the way for the discussion in the following sections. Specifically, these sections deal with methods for causal reasoning when randomization is not feasible. Section 3.5 provides an example of instrumental variation in treatment assignment arising from a natural experiment. Section 3.6 is devoted to the closest cousin to randomization, the regression discontinuity design. Section 3.7 offers some concluding remarks.

Our discussion of empirical methods for causal reasoning is far from exhaustive. For example, we do not discuss research designs that exploit longitudinal data and rely on assumptions on pre-treatment outcome trends (e.g., difference-in-differences and synthetic control methods). Similarly, we do not cover matching methods (see Chap. 4 of this volume). In addition, our presentation will mostly focus on the reasoning underlying design-based identification and will only barely touch issues related with estimation. The interested reader can refer to the book by Angrist and Pischke (2008) for a discussion of these topics.

3.2 Causation and Counterfactual Impact Evaluation: The Jargon

It is useful to start by clarifying what we mean by “causes” and “treatment effects.” We consider a population of units indexed by i , with $i = 1, \dots, N$. Although our narrative will often consider individuals as the units of analysis, the same setting extends to other statistical units such as households, villages, schools, or municipalities.

3.2.1 *Causes as Manipulable Treatments*

In the population we study, some units are exposed to a cause, which is a treatment or intervention that manipulates factors that may affect a certain outcome. For instance, we might be interested in studying whether class size at primary school affects student performance. Class size here is the treatment and performance is the outcome, which is typically measured using standardized tests. In many countries, class size formation depends on grade enrollment so that, across cohorts, the number of students in the class may change because enrollment changes or because a specific policy affects the regulation. We will use the words “cause”, “treatment”, or “intervention” interchangeably.

The avenue we take here has some limitations, as not all causes worth considering are manipulable in practice (consider, for example, gender, ethnicity, or genetic traits). Moreover, the design-based approach we describe below may be coarse at times and aimed at shedding light on one particular aspect of a more articulated model. For example, empirical evidence on the causal effects of class size on achievement bundles up the possible contribution of multiple channels that may lead to a better learning environment in small classes. The investigation of channels and mechanisms behind the uncovered effects calls for theories and structural models. The most relevant question to consider turns on the quality of the design-based strategy and on our faith to prop up a more elaborate theoretical framework.

We focus only on binary treatments, that is, we assume that treatment status is described by a binary random variable D_i taking value one if unit i is exposed to treatment (“treated” or “participant”) and zero otherwise (“untreated”, “non-participant”, or “control”). In the class size example, this amounts to considering a setting in which students can be enrolled in small or large classes. The extension to the case of multi-valued or continuous treatment (for example, the number of classmates) is logically identical but requires a more cumbersome notation. More in general, the binary case is always worth of consideration even in a more general context as it helps understand the main challenges in the quest for detecting causal effects. A related issue concerns public policies that are designed as “bundles” of multiple components. In those cases, policy-makers are often interested in disentangling the effect of every component of the policy. We abstract from this problem in our discussion, but emphasize here that the ability to address this question will depend, in general, on the exposure of subjects to different components.

We must take a stand on the reasons why different units end up having a value of D_i equal to one or zero. This is the so-called “assignment rule” and is at the core of any evaluation study. Assignment to treatment can be totally random. In our class size example, this happens when students are randomized to a small or a large class with equal probability and independently of socio-economic background or past performance. When randomization is not at work, participation to treatment is most likely the result of choices made by the units themselves, administrators of the program, or policy makers. For example, parents can choose to enroll their children in schools with smaller classes in the hope of a better learning environment. Finally,

participation to treatment may depend on admission rules that units must comply with. The case of class size formation based on total enrollment is a good example, as the chance of being enrolled in a small class depends on a school's yearly total recruitment. As we shall see, our ability to assess causal effects grows with knowledge of the assignment rule.

3.2.2 *Effects as Differences Between Factual and Counterfactual Outcomes*

It is essential to set the stage for a transparent definition of the treatment effect. To do so, we define $Y_i(1)$ and $Y_i(0)$ as the potential outcomes experienced if unit i is treated ($D_i = 1$) or untreated ($D_i = 0$), respectively. The unit-level treatment effect of D_i on Y_i is the difference between $Y_i(1)$ and $Y_i(0)$: $\Delta_i = Y_i(1) - Y_i(0)$. Decades of empirical studies using micro-data analyses have taught us that treatment effects most likely vary across units or groups of units with very similar demographics. The notation employed here accommodates for this possibility (the manuals by Angrist & Pischke, 2008, and Imbens & Rubin, 2015, use the same approach).

The definition of Δ_i unveils the fundamental problem that we face when we want to estimate this quantity from the data. While the two potential outcomes can be logically defined for each unit, they can never be observed simultaneously for the same unit. This is true regardless of the assignment rule and the richness or sample size of data we will ever work with. Specifically, the data can reveal only $Y_i(1)$ for units with $D_i = 1$ and $Y_i(0)$ for units with $D_i = 0$. We can, therefore, express the observed outcome Y_i as follows: $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i) = Y_i(0) + D_i(Y_i(1) - Y_i(0))$. As simple as this can be, lack of observability of both potential outcomes implies lack of observability of the unit-level effect Δ_i . We can think of the unit-level causal effect as the difference between an observed (factual) and an unobserved (counterfactual) potential outcome. Factual quantities are those that can be computed from the data. Counterfactual quantities can be logically defined but can never be computed from data. For treated units, we observe $Y_i = Y_i(1)$ and $Y_i(0)$ is the counterfactual. The opposite is true for control units, for whom we observe $Y_i = Y_i(0)$ and $Y_i(1)$ is the counterfactual.

One way to get around this limitation is to settle for less than unit-level effects. We might be interested in considering average treatment effects for the population or only for some sub-groups. For instance, we define the average treatment effect (ATE) as the average of the individual-level treatment effect in the whole population: $ATE = E(Y_i(1) - Y_i(0))$. This parameter reflects our expectation of what would happen if we were to expose to treatment a randomly chosen unit from the population. Alternatively, we can consider the average treatment effect for the treated (ATT), which describes our expectation for units who have been exposed to treatment: $ATT = E(Y_i(1) - Y_i(0) | D_i = 1)$. Analogously, the average treatment effect for

the non-treated (ATNT) is informative about what would have happened to the untreated if they had been exposed to the intervention:

$$ATNT = E(Y_i(1) - Y_i(0) | D_i = 0).$$

Whether any of the above causal parameters can be retrieved from the data will have to be discussed on a case-by-case basis our understanding of the assignment rule plays a key role in this discussion.

3.2.3 What the Data Tell (And When)

Our journey to learn about treatment effects begins by comparing features of the observed outcome Y_i for treated and control units. For instance, the data reveal the average outcomes for treated units, $E(Y_i | D_i = 1)$, and control units, $E(Y_i | D_i = 0)$. Recalling the definition of potential outcomes, the naïve comparison of average outcomes by treatment group, $E(Y_i | D_i = 1) - E(Y_i | D_i = 0) = E(Y_i(1) | D_i = 1) - E(Y_i(0) | D_i = 0)$, conveys the correlation between the treatment, D_i , and the outcome, Y_i .

The causal interpretation of such naïve comparison is controversial in most cases. To see why, we can add and subtract from the right-hand side of the previous equation the quantity $E(Y_i(0) | D_i = 1)$. This is a counterfactual quantity, as the outcome $Y_i(0)$ cannot be observed for treated units, and represents what would have happened to treated units had they not participated to treatment. We can arrange the terms and write:

$$\begin{aligned} E(Y_i | D_i = 1) - E(Y_i | D_i = 0) &= E(Y_i(1) - Y_i(0) | D_i = 1) + E(Y_i(0) | D_i = 1) \\ &\quad - E(Y_i(0) | D_i = 0). \end{aligned} \tag{3.1}$$

It follows that the naïve comparison on the left-hand side of Eq. 3.1 is equal to the sum of the ATT and the term $E(Y_i(0) | D_i = 1) - E(Y_i(0) | D_i = 0)$, which is often called “selection bias”. It is worth noting that this representation does not hinge on any assumptions. It is the result of a simple algebraic trick and, as such, is always true.

Selection bias is an error in the causal reasoning. It is different from zero when, in the absence of treatment, the group with $D_i = 1$ would have performed differently from the group with $D_i = 0$. The same concept is conveyed by the “correlation is not causation” *motto*: correlation (the naïve treatment–control comparison) has no causal interpretation (that is, it does not coincide with the ATT) unless the selection bias is zero. This reframes the quest for causal effects as a discussion on the existence of selection bias. A non-zero bias follows from having groups defined by $D_i = 1$ and $D_i = 0$ that are not representative of the same population, in the sense that participation to treatment depends on non-random selection. At the end of the day, selection bias reflects compositional differences between treatment and control

units. Taking up our class size example, parents with a strong preference for smaller classes are most likely selected in terms of socio-economic background and demographics. If this selection translates into a better learning potential of their children, forming classes as a reflection of parental preference must create dis-homogenous groups of students. In this case, detecting a correlation between class size and achievement might just reveal dis-homogeneity across classes rather than a true causal effect of class size.

Importantly, for the time being, we are agnostic about whether this dis-homogeneity concerns characteristics of units that are observed in the data at hand or not. In fact, any strategy that can adjust for compositional differences between treated and control units also corrects for this bias. One leading example to consider here is randomization. When classes are formed by a coin toss, composition is the same. Even when it is because of sampling variability, differences in composition must be as good as random. We will formalize this idea in Sect. 3.4, below. Instead, Chapters 4 and 5 in this volume present methods to alleviate imbalances along observable dimensions and discuss the identifying assumptions that permit to reach causal conclusions once these differences are eliminated.

3.3 Shades of Validity

The assessment of a causal channel from treatment to the outcome depends on the properties of the research design. In short, this is the toolbox of empirical methods that allows one to distinguish between correlation and causality. Any strategy falling short on this minimum requirement is not a valid option to consider for a good researcher. On the other hand, a good research design must be able to detect precisely the causal relationship of interest. That is, you do not want your design to be underpowered for the size of the treatment effect. Finally, the ideal research design should be able to provide causal statements that apply to the largest share of units in the population and extend to other contexts and times. The concern here is one of generalizability, which is of fundamental importance for offering evidence-based policy recommendations. Causal talk makes use of these three ideas of validity in the development of a research design. This is what we will discuss briefly next. The seminal textbook by Cook and Campbell (1979) provides a deeper treatment of these topics.

3.3.1 *Internal Validity: The Ability to Make a Causal Claim from a Pattern Documented in the Data*

Internal validity concerns the ability of assessing whether the correlation between treatment and outcome depicts a causal relationship or if it could have been observed even in the absence of the treatment. Therefore, internal validity is solely concerned

with the presence of selection bias. It is achieved under a *ceteris paribus* comparison of units, when all else but the treatment is kept constant between treated and control units. As we discussed above, this calls for the same composition of treatment (small class) and control (large class) units. An internally valid conclusion is the one without selection bias. One of the main advantages of using randomization is that such *ceteris paribus* condition is met by design. Because of this, a properly conducted randomization yields internally valid causal estimates.

3.3.2 *Statistical Validity: Measuring Precisely the Relationship Between Causes and Outcomes in the Data*

Statistical validity refers to the appropriate use of statistical tools to assess the extent of correlation between treatment and outcomes. It is fundamentally concerned with standard errors and accuracy in assessing a statistical relationship. The main question addressed by statistical validity is whether the chosen data and techniques of statistical inference can produce precise estimates of very small treatment effects (a statistically precise zero) or if, instead, the research design will likely produce statistical zeros (a statistically insignificant effect). An insignificant effect that is statistically different from zero is a powerful oxymoron to summarize the idea underlying statistical validity.

3.3.3 *External Validity: The Ability to Extend Conclusions to a Larger Population, over Time and Across Contexts*

External validity is about the predictive value of a particular causal estimate for times, places, and units beyond those represented in the study that produced it. The concern posed by external validity is one of generalizability and out-of-sample prediction. For example, an internally valid estimate for a given sub-group of the population might not be informative about the treatment effect for other (potentially different and policy-relevant) sub-groups. Similarly, ATT is, in general, different from ATE. Replicability of the same results in other contexts and times is of fundamental interest for providing policy recommendations.

3.4 Random Assignment Strengthens Internal Validity

As Andrew Leigh puts it in his book “*Randomistas: How Radical Researchers Are Changing the World*,” (Leigh, 2018) randomized controlled trials (RCTs) use “the power of chance” to assign the groups. Randomization can be achieved by flipping a coin, drawing the shorter straw, or using a computer to randomly assign statistical

units to groups. In any of these cases, the result would be the same: the treatment and the control group are random samples from the same population.

Random assignment ensures that treatment and control units are the same in every respect, including their expected $Y_i(0)$. It follows that, in RCTs, selection bias must be zero since $E(Y_i(0) | D_i = 1) = E(Y_i(0) | D_i = 0)$. In other words, what we observe for control units approximates what would have happened to treated units in the absence of treatment. It is worth noting that random assignment does not work by eliminating individual differences, but it rather ensures that the composition of units being compared is the same.

RCTs ensure a *ceteris paribus* (i.e., without confounds) comparison of treatment and control groups. Because of this, an RCT provides an internally valid research design for assessing causality. Evidence in support of this validity can be obtained using pre-intervention measurements. In fact, it is a good practice to collect this information and test the validity of the design by carrying out a battery of “balancing” tests. In a properly implemented randomization, there are no selective differences in the distribution of pre-intervention measurements between treated and control units. This statement does not rule out the possibility of between-group differences arising from sampling variability, which is a problem concerning the statistical validity (that is, the precision of point estimates) of RCTs.

Finally, under random assignment, the naïve comparison will provide internally valid conclusions about the average treatment effect on the treated (ATT), as we have that $E(Y_i | D_i = 1) - E(Y_i | D_i = 0) = E(Y_i(1) - Y_i(0) | D_i = 1)$. In addition, under randomization, the groups with $D_i = 1$ and $D_i = 0$ are representative of the same population so that $E(Y_i(1) - Y_i(0) | D_i = 1) = E(Y_i(1) - Y_i(0))$. This means that the causal conclusions hold for any unit randomly selected from the population.

Random assignment to treatment is not uncommon in numerous fields of the social sciences. One such example is the lottery-based allocation of pupils to schools that are oversubscribed. This alternative to the traditional priority criterion based on proximity should dampen school stratification caused by wealthy parents buying houses in the close vicinity of high-quality schools. As a result, among the pool of applicants to a school where oversubscription is resolved by a lottery, getting a seat or not is completely random. Some researchers (see Cullen et al., 2006, for an example) have exploited this to evaluate the educational effects of attending one’s preferred school.

Another example is the Oregon Health Insurance Experiment (see Finkelstein et al., 2012). Medicaid is one of the landmark US public health insurance programs and provides care for millions of low-income families. In 2008, the state of Oregon extended coverage of Medicaid by selecting eligible individuals with a lottery. This gave researchers the unique opportunity to provide credible causal estimates of the effect of health insurance eligibility on health care utilization, medical expenditure, medical debt, health status, earnings, and employment.

Although RCTs are considered as the “gold standard” for providing internally valid estimates of causal effects, they are not without shortcomings (see the excellent surveys by Duflo et al., 2008 and Peters et al., 2018). External validity is often perceived as the main limitation and more so for small-scale experiments on very

specific subpopulations. Bates and Glennerster (2017) propose a framework to discuss generalizability based on four steps: identify the theory behind the program; check if local conditions hold for that theory to apply; evaluate the strength of the evidence for the required general behavioral change; evaluate whether the implementation process can be carried out well. External validity is granted if these four conditions apply in a context different from the one where the experiment was conducted. Statistical validity as well may challenge the significance of many small-scale experiments (see Young, 2019).

RCTs have other limitations. Many RCTs are carried out as small-scale pilots that shall be eventually scaled up to the entire population. Causal reasoning in this context must consider the general equilibrium effects arising from this change in scope. These effects are concerned with the possible externalities for non-participants when the policy is implemented on a larger scale and the implications for market equilibria. An additional concern about RCTs is that the sole fact of being “under evaluation” may generate some behavioral response that has nothing to do with a treatment effect.¹ Replicability of experiments also has been called into question in many fields of the social sciences (see Open Science Collaboration, 2015, for psychology and Camerer et al., 2016, for economics).

What happens when randomization is not a feasible option? This is the question to which we turn next.

3.5 Internally Valid Reasoning Without RCTs: Instrumental Variation

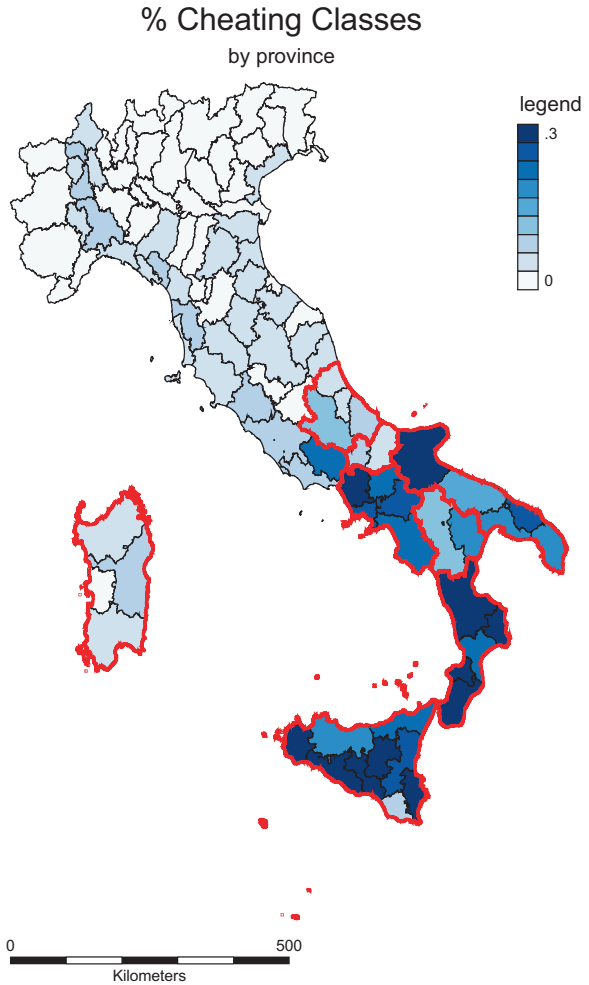
3.5.1 *A Tale of Pervasive Manipulation*

Randomizations obtained by design are not the only way to ensure *ceteris paribus* comparisons. Randomness in the assignment to treatment may arise indirectly from natural factors or events independently of the causal channel of interest. Under assumptions that we shall discuss, these factors can be used instrumentally to pin down a meaningful causal parameter. The most important takeaway message here is that we must use assumptions to make up for the lack of randomization. Because of this, much of the simplicity of the research design is lost, and internal validity must be addressed on a case-by-case basis. We will present an example of the toolbox for good empirical investigations using administrative data on student achievement and, further below, class size.

Our working example makes use of standardized tests from INVALSI (a government agency charged with educational assessment) for second and fifth graders in Italian schools for the years 2009–2011. Italy is an interesting case study as it is

¹Such quirky responses are called “Hawthorne” effects for treated subjects and “John Henry” effects for controls.

Fig. 3.1 Manipulation by province (Angrist et al., 2017). (Mezzogiorno regions are bordered with dashed lines)



characterized by a sharp North–South divide along many dimensions, among which school quality. This divide motivates public interventions to improve school inputs in the South. As testing regimes have proliferated in the country, so has the temptation to cut corners or cheat at the national exam.² As shown in Fig. 3.1, the South is distinguished by widespread manipulation on standardized tests. INVALSI tests are usually proctored and graded by teachers from the same school, and past work by Angrist et al. (2017) has shown that manipulation takes place during the grading process. Classes with manipulated scores are those where teachers did not grade exams honestly.

Consider the causal effect of manipulation on test scores. As scores are inflated, the sign of this effect is obvious. However, the size of the causal effect (that is, by

²Cheating or manipulation is not unique to Italy, as discussed in Battistin (2016).

how much scores are inflated) is difficult to measure because manipulation is not the result of random factors. The incentive to manipulate likely decreases as true scores increase so that the distribution of students' true scores is not the same across classes with teachers grading honestly or dishonestly. Again, this is a problem about the composition of the two groups, as treatment classes (with manipulated scores) and control classes (with honest scores) need not be representative of the same population.

When empirical work is carried out using observational data, as it is the case here, it is always illuminating to start from the thought experiment. This is the hypothetical experiment that would be used to measure the causal effect of interest if we had the possibility to randomize units. With observational data, the identification strategy consists of the assumptions that we must make to replicate the experimental ideal. The thought experiment in the case of INVALSI data corresponds to distributing manipulation (the treatment) across classes at random. The identification strategy here amounts to the set of assumptions needed to mimic the very same experimental ideal *even if* manipulation is not random. How can this be possible?

Econometrics combined with the institutional context come to the rescue. It turns out that about 20% of primary schools in Italy are randomly assigned to external monitors, who supervise test administration and the grading of exams from local teachers in selected classes within the school (see Bertoni et al., 2013, and Angrist et al., 2017, for details on the institutional context). Table 3.1 shows that monitors are indeed assigned to schools using a lottery. Schools with monitors are statistically indistinguishable from the others along several dimensions, including average class size and grade enrollment. For example, the table shows that the average class size in unmonitored classes of the country is 19.812 students. The difference between treated and control classes is as small as 0.035 students and statistically indistinguishable from zero. Additional evidence on the lack of imbalance between schools with and without monitors is in Angrist et al. (2017). In the next section, we discuss how to use the monitoring randomization to learn about the effects of manipulation on scores.

3.5.2 *General Formulation of the Problem*

In our example, the class is the statistical unit of analysis and the treatment is manipulation ($D_i = 1$ if class scores are manipulated and $D_i = 0$ if they are honestly reported). INVALSI has developed a procedure to reveal D_i , so treatment status is observed in the data. Scores (standardized by grade, year, and subject) are the class-level outcome, Y_i . The presence of external monitors is described by a binary random variable Z_i , with $Z_i = 1$ for classes in schools with monitors and $Z_i = 0$ otherwise. In the applied econometrics parlance, variables like Z_i —which is randomly assigned and can influence treatment status—are called “instruments.”

The ordinary least squares (OLS) regression of Y_i on D_i summarizes the correlation between manipulation and reported scores. Estimation results obtained from

Table 3.1 Covariate balance in the monitoring experiment (Angrist et al., 2017)

	Italy		North/Center		South	
	Control mean	Treatment difference	Control mean	Treatment difference	Control mean	Treatment difference
	(1)	(2)	(3)	(4)	(5)	(6)
Class size	19.812 [3.574]	0.0348 (0.0303)	20.031 [3.511]	0.0179 (0.0374)	19.456 [3.646]	0.0623 (0.0515)
Grade enrollment at school	53.119 [30.663]	-0.4011 (0.3289)	49.804 [27.562]	-0.5477 (0.3913)	58.483 [34.437]	-0.1410 (0.5909)
% in class sitting the test	0.939 [0.065]	0.0001 (0.0005)	0.934 [0.066]	0.0006 (0.0006)	0.947 [0.062]	-0.0007 (0.0008)
% in school sitting the test	0.938 [0.054]	-0.0001 (0.0005)	0.933 [0.055]	0.0005 (0.0006)	0.946 [0.051]	-0.0010 (0.0008)
% in institution sitting the test	0.937 [0.045]	-0.0001 (0.0004)	0.932 [0.043]	0.0005 (0.0005)	0.945 [0.045]	-0.0010 (0.0007)
N	140,010		87,498		52,512	

Columns 1, 3, and 5 show means and standard deviations for variables listed at the left. Other columns report coefficients from regressions of each variable on a treatment dummy (indicating classroom monitoring), grade and year dummies, and sampling strata controls (grade enrollment at institution, region dummies, and their interactions). Standard deviations for the control group are in square brackets; robust standard errors are in parentheses

^a $p < 0.01$, ^b $p < 0.05$, ^c $p < 0.1$

OLS are reported in Table 3.2, and a positive correlation between cheating and test score is revealed in all columns. For instance, the value of the coefficient reported in Column (1) of Panel A implies that when we consider data for the whole of Italy, the average math score in classes with manipulated scores is 1.414 standard deviations higher than in classes where teachers did not manipulate scores.³ However, as discussed above, this result cannot be given any causal interpretation, as the samples with $D_i = 0$ and $D_i = 1$ are non-randomly selected.

Unlike D_i , the status Z_i is randomly assigned. So, it is can be instructive to consider the regression of Y_i on Z_i , summarizing the correlation between manipulation and monitoring. As Z_i is randomly assigned, the latter regression yields the causal effect of monitoring on scores (orthodox empiricists often call this regression the “reduced form equation”). Results in Columns (1)–(3) of Table 3.3 show a negative effect of monitoring on test scores in all columns (see Bertoni et al., 2013). For example, from Column (1) of Panel A, we learn that the average math score in schools with external monitors is 0.112 standard deviations lower than in schools without monitors. Arguably, the negative effect of monitoring on scores passes through a reduction of manipulation.

We need to enrich our causal inference vocabulary to consider potential outcomes based on the 2×2 scenarios that result from the cross-tabulation of D_i and Z_i : $Y_i(D_i, Z_i)$. Similarly, we need to adjust the notation to express the idea that Z_i

³Here and in what follows, INVALSI scores are standardized to have zero mean and unit variance by subject and year.

Table 3.2 Correlation between score manipulation and test scored

	Test scores		
	Italy	North/Center	South
	(1)	(2)	(3)
	A. Math		
Score manipulation	1.414 ^a (0.006)	1.404 ^a (0.009)	1.413 ^a (0.007)
Means	0.007	-0.074	0.141
(sd)	(0.637)	(0.502)	(0.796)
N	139,996	87,491	52,505
	B. Language		
Score manipulation	1.179 ^a (0.005)	1.085 ^a (0.007)	1.213 ^a (0.006)
Means	0.01	-0.005	0.035
(sd)	(0.523)	(0.428)	(0.649)
N	140,003	87,493	52,510

All models control for a quadratic polynomial in grade enrollment, segment dummies, and their interactions. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and the proportions of missing values in these variables. All regressions additionally include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). ^ap<0.01, ^bp<0.05, ^cp<0.1

Table 3.3 Monitoring effects on test scores and score manipulation (Angrist et al., 2017)

	Test scores			Score manipulation		
	Italy	North/Center	South	Italy	North/Center	South
	(1)	(2)	(3)	(4)	(5)	(6)
	A. Math					
Monitor at institution (M_{igkt})	-0.112 ^a (0.006)	-0.075 ^a (0.005)	-0.180 ^a (0.012)	-0.029 ^a (0.002)	-0.010 ^a (0.001)	-0.062 ^a (0.004)
Means	0.007	-0.074	0.141	0.064	0.02	0.139
(sd)	(0.637)	(0.502)	(0.796)	(0.246)	(0.139)	(0.346)
N	140,010	87,498	52,512	139,996	87,491	52,505
	B. Language					
Monitor at institution (M_{igkt})	-0.081 ^a (0.004)	-0.054 ^a (0.004)	-0.131 ^a (0.009)	-0.025 ^a (0.002)	-0.012 ^a (0.001)	-0.047 ^a (0.004)
Means	0.01	-0.005	0.035	0.055	0.023	0.11
(sd)	(0.523)	(0.428)	(0.649)	(0.229)	(0.149)	(0.313)
N	140,010	87,498	52,512	140,003	87,493	52,510

Columns 1–3 report the reduced form effects of having a monitor at the institution on test scores. Columns 4–6 show the first-stage estimates of the effect of having a monitor at the institution on score manipulation. All models control for a quadratic polynomial in grade enrollment, segment dummies, and their interactions. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and proportions of missing values in these variables. All regressions additionally include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). ^ap<0.01, ^bp<0.05, ^cp<0.1

affects D_i . We define potential treatments $D_i(0)$ and $D_i(1)$ as the treatment status that individual i has when exposed to $Z_i = 0$ and $Z_i = 1$, respectively. In our running example, the realized score Y_i corresponds to the potential score realized for the observed combination $\{D_i = d, Z_i = z\}$, while the realized manipulation D_i coincides with the potential manipulation realized for the observed value of $Z_i = z$. For example, $Y_i(1, 1)$ represents the score that would be recorded for class i if teacher grading was dishonest ($D_i = 1$) and the school had an INVALSI monitor ($Z_i = 1$). Recall that, since only selected classes within the school are monitored, dishonest behavior from teachers in unmonitored classes within the school is always possible (see Bertoni et al., 2013).

Depending on the values taken by $D_i(0)$ and $D_i(1)$, we can divide classes into four groups depending on the behavior of teachers grading the exams (see Battistin et al., 2017, for a similar approach):

- Complying dishonest teachers (C), who grade dishonestly without monitors and grade honestly with monitors: $D_i(0)=1$ and $D_i(1) = 0$.
- Always dishonest teachers (A), who always grade dishonestly regardless of the presence of monitors: $D_i(0)=1$ and $D_i(1) = 1$.
- Never dishonest teachers (N), who always grade honestly regardless of the presence of monitors: $D_i(0)=0$ and $D_i(1) = 0$.
- Non-complying dishonest teachers (D), who grade honestly without monitors and grade dishonestly with monitors: $D_i(0)=0$ and $D_i(1) = 1$.

This classification does not hinge on any assumptions and represents the taxonomy of all possible behavioral responses from teachers arising from the monitoring status of the school. The fact that both D_i and Z_i are binary limits to four the number of such responses.

3.5.3 Assumptions

The identification strategy for the analysis of natural experiment builds on four assumptions. We now discuss each of them with reference to our specific running example on the effect of manipulation on test scores. We refer the reader to Angrist and Pischke (2008) for a more general discussion.

3.5.3.1 The “Monotonicity” Assumption

We begin our investigation by assuming lack of non-complying dishonest teachers (D -teachers) in the data. This is a rather innocuous assumption in our context. A violation would represent a quirky behavioral response to the presence of monitors. This assumption is also known as monotonicity condition. It is a restriction on the behavior of units stating that when we move the instrument Z_i from z' to z'' , all agents respond by changing their D_i in the same direction or by leaving it unaltered. In our

case, this assumption implies that (a) honest teachers without monitors at school would be honest teachers even with a monitor and (b) dishonest teachers without monitors at school might grade honestly under the threat of a monitor at school. In the former case, the value of D_i is unchanged by monitoring and remains zero; in the latter case, the value of D_i may remain one or turn to zero with monitoring. The events (a) and (b) imply that the distribution of the variable D_i must move toward zero in the presence of school monitoring. Ruling out the presence of D -teachers implies that monitors cannot change the variable D_i in the opposite direction, from zero to one. This exemplifies why the variable Z_i must induce a monotone (towards zero) behavior for all teachers.

Monotonicity plays a crucial role in natural experiments: under this assumption, we are left with three compliance types— C , A , and N —whose shares in the populations can be represented by π_C , π_A , π_N , respectively. Manipulators are a mixture of always dishonest teachers (A -teachers) and complying dishonest teachers (C -teachers) without monitors. Honest teachers are composed of never dishonest teachers (N -teachers) and complying dishonest teachers (C -teachers) with monitors.

3.5.3.2 The “As Good as Random” Assumption

A second key relationship among the variables involved arises because schools are randomly assigned to either $Z_i = 1$ or $Z_i = 0$. Because of the monitoring experiment, the two groups of schools must have the same composition with respect to any variable, including potential outcomes and potential treatment statuses. It, therefore, follows that $\{Y_i(1, 1), Y_i(0, 1), Y_i(1, 0), Y_i(0, 0), D_i(0), D_i(1)\} \perp Z_i$. In our case, this “as good as random” assumption holds by design, because monitors have been explicitly assigned at random to schools.

3.5.3.3 The “Exclusion Restriction”

The causal reasoning builds upon an exclusion restriction. This formalizes the causal construct that the effect of Z_i on Y_i shall be solely because of the effect of Z_i on D_i . In the example considered here, this restriction can be put across considering the following equations:

$$\begin{aligned} Y_i(0,1) &= Y_i(0,0), \\ Y_i(1,1) &= Y_i(1,0). \end{aligned}$$

Therefore, the exclusion restriction implies that there are only two potential outcomes, indexed against D_i : $Y_i(D_i)$. For example, the first equation implies that scores under honest grading ($D_i = 0$) would be the same irrespective of the presence of monitors. Similarly, the second equation implies that dishonest grading ($D_i = 1$) would yield the same score independently of school monitoring. The latter

condition would be violated if, for example, always dishonest teachers cheated differently in the presence of external monitors at school. This possibility is discussed in Battistin et al. (2017) and is ruled out in the case of INVALSI data by results in Angrist et al. (2017).

3.5.3.4 The “First-Stage” Requirement

The assumed causal link from D_i to Z_i can be verified in the data by running an OLS regression of D_i on Z_i . In fact, it is a good practice to verify the size and statistical strength of this “first-stage” regression in any study based on quasi-experimental variation, as the causal chain we have in mind originates from the effect of Z_i on D_i . Should we observe any effect of Z_i on Y_i but no effect of Z_i on D_i , it would be hard to justify that the random variation in Z_i affected Y_i via the ability of Z_i to move D_i . Estimates of the “first-stage” relationship between exposure to monitors and manipulation are reported in Columns (4)–(6) of Table 3.3. As expected, score manipulation is less likely in schools where monitors are present. For example, Column (4) of Panel A indicates that the probability of score manipulation is 2.9 percentage points lower in schools of the country with monitors. This is equivalent to a 36% decrease in the probability of manipulation with respect to the mean in non-monitored schools (equal to 6.4%). As demonstrated by the estimates in Columns (5) and (6) of Table 3.3, this decrease is stronger in Southern Italy than in the North and Center of the country and strongly statistically significant.

3.5.4 Better LATE than Never

To nail down the causal effect of manipulation on scores, we proceed by comparing the expected value of the product $Y_i D_i$ for schools with and without monitors. This product is equal to Y_i for units with $D_i = 1$ and to 0 for units with $D_i = 0$. Given all the assumptions made so far, we have that:

$$E(Y_i D_i | Z_i = 1) = \pi_A * E(Y_i(1) | A),$$

$$E(Y_i D_i | Z_i = 0) = \pi_C * E(Y_i(1) | C) + \pi_A * E(Y_i(1) | A).$$

In the first equation, neither C -teachers or N -teachers show up, because for them $D_i = 0$ when $Z_i = 1$ so that $Y_i D_i = 0$.⁴ Because of the monotonicity assumptions, there

⁴A consequence of random assignment of Z_i and of the exclusion restriction is that conditional on the compliance types defined above, potential outcomes are independent of Z_i , that is, $\{Y_i(1), Y_i(0)\} \perp Z_i | \{D_i(0), D_i(1)\}$. In fact, conditional on a given compliance type, there is a one-to-one mapping between Z_i and D_i , and therefore, knowledge of Z_i implies knowledge of D_i .

are no D -type teachers either. Therefore, the only group left is that of A -teachers, whose fraction in the population is π_A and for whom we always observe $Y_i(1)$. In a similar fashion, we do not see N -teachers in the second line, as for them, $D_i = 0$ when $Z_i = 0$. Consequently, after ruling out the presence of D -teachers by monotonicity, only C - and A -teachers show up in this equation. C -teachers account for a fraction π_C of the population, and for them, we observe $Y_i(1)$ as in this case $Z_i = 0$, and therefore, $D_i = 1$.

For these very same reasons, if we compare the share of manipulators in schools with and without external monitors, we obtain:

$$E(D_i | Z_i = 1) = \pi_A,$$

$$E(D_i | Z_i = 0) = \pi_C + \pi_A.$$

The former expression suggests that only A -teachers have $D_i = 1$ when $Z_i = 1$; the latter that are both C - and A -teachers have $D_i = 1$ when $Z_i = 0$. Analogous expressions can be derived for $E(Y_i(0) | C)$, $E(Y_i(0) | N)$ and for π_N if one substitutes D_i with $(1 - D_i)$ in the above. We have that:

$$E(Y_i(1 - D_i) | Z_i = 1) = \pi_C * E(Y_i(0) | C) + \pi_N * E(Y_i(0) | N),$$

$$E(Y_i(1 - D_i) | Z_i = 0) = \pi_N * E(Y_i(0) | N),$$

$$E((1 - D_i) | Z_i = 1) = \pi_C + \pi_N,$$

$$E((1 - D_i) | Z_i = 0) = \pi_N.$$

In the first and third equation, A -teachers do not show up because they always have $D_i = 1$ so that $Y_i(1 - D_i) = 0$ and $(1 - D_i) = 0$.⁵ Because of the monotonicity assumptions, there are no D -type teachers either. Therefore, only C - and N -teachers are left. C -teachers account for a fraction π_C of the population. Since in this case $Z_i = 1$, for them, we observe $D_i = 0$ and, therefore, $Y_i(0)$. N -teachers are a share π_N of the population, as for them, D_i is always equal to 0, and we observe $Y_i(0)$.

Similarly, in the second and fourth line, we do not see A - and C -teachers, as for them $D_i = 1$ when $Z_i = 0$. Consequently, after ruling out the presence of D -teachers by monotonicity, only N -teachers are left.

⁵A consequence of random assignment of Z_i and of the exclusion restriction is that conditional on the compliance types defined above, potential outcomes are independent of Z_i , that is, $\{Y_i(1), Y_i(0)\} \perp Z_i | \{D_i(0), D_i(1)\}$. In fact, conditional on a given compliance type, there is a one-to-one mapping between Z_i and D_i , and therefore, knowledge of Z_i implies knowledge of D_i .

By rearranging the equations above, it is easy to obtain:

$$E(Y_i(1)|C) = \frac{E(Y_i D_i | Z_i = 1) - E(Y_i D_i | Z_i = 0)}{E(D_i | Z_i = 1) - E(D_i | Z_i = 0)}, \quad (3.2)$$

and

$$E(Y_i(0)|C) = \frac{E(Y_i(1 - D_i) | Z_i = 1) - E(Y_i(1 - D_i) | Z_i = 0)}{E((1 - D_i) | Z_i = 1) - E((1 - D_i) | Z_i = 0)}. \quad (3.3)$$

The difference between the last two expressions yields:

$$E(Y_i(1) - Y_i(0)|C) = \frac{E(Y_i | Z_i = 1) - E(Y_i | Z_i = 0)}{E(D_i | Z_i = 1) - E(D_i | Z_i = 0)}, \quad (3.4)$$

which represents the average causal effect of manipulation for classes with teachers who graded honestly because of school monitoring (that is, classes with C -teachers). Intuitively, this happens because—in the absence of D -teachers—this is the only group of teachers for whom the presence/absence of monitors generates variation in manipulation. Borrowing the definition by Angrist and Imbens (1994), the parameter on the left-hand side of (3.4) is the *local average treatment effect* (LATE). The word “local” here is motivated by causal conclusions only licensed for a subset of classes in the population.

Importantly, the expression on the right-hand side of Eq. 3.4 involves only the variables observed so that the causal parameter can be identified from the data. Standard econometric results imply that LATE is estimated by the coefficient on D_i in a two-stage least squares (TSLS) regression of Y_i on D_i , using Z_i to instrument for D_i .⁶ Table 3.4 reports the estimates of the LATE parameter in our running example and reveals that manipulation causally increased scores of students assigned to complying dishonest teachers. For example, Column (1) of Panel (A) tells us that score manipulation increases math results in classes with C -teachers by 3.827 standard deviations. This causal effect is much larger than the naïve comparison of scores by treatment status reported in Column (1) of Panel A in Table 3.2. Why is it the case? As illustrated in Sect. 3.2.3, the naïve comparison is equal to a causal effect plus selection bias. In this case, selection bias corresponds with the difference in average score of manipulators and non-manipulators if manipulation was not possible at all. As we have argued, manipulation is less likely to occur in classes with higher average true scores. So, selection bias is likely to be negative, that is, $E(Y_i(0)|D_i = 1) < E(Y_i(0)|D_i = 0)$.

⁶A similar result applies to the expressions in (3.2) and (3.3) when TSLS regressions of $Y_i D_i$ on D_i and of $Y_i(1 - D_i)$ on $(1 - D_i)$, respectively, are considered.

Table 3.4 Local average treatment effect of score manipulation on test scores

	Test scores		
	Italy	North/Center	South
	(1)	(2)	(3)
	A. Math		
Score manipulation (D_{igkt})	3.827 ^a	7.393 ^a	2.886 ^a
	(0.188)	(0.804)	(0.158)
Means	0.007	−0.074	0.141
(sd)	(0.637)	(0.502)	(0.796)
N	139,996	87,491	52,505
	B. Language		
Score manipulation (D_{igkt})	3.279 ^a	4.523 ^a	2.786 ^a
	(0.180)	(0.456)	(0.178)
Means	0.01	−0.005	0.035
(sd)	(0.523)	(0.428)	(0.649)
N	140,003	87,493	52,510

All models control for a quadratic in grade enrollment, segment dummies, and their interactions. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and proportions of missing values in these variables. All regressions include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). ^a $p < 0.01$, ^b $p < 0.05$, ^c $p < 0.1$

3.5.5 External Validity of Causal Conclusions

Causal conclusions can be drawn for classes with exams graded by C -teachers, and TSLS yield internally valid estimates of $E(Y_i(1) - Y_i(0) | C)$. However, we have that $E(Y_i(1) - Y_i(0) | C) \neq E(Y_i(1) - Y_i(0))$ in general. It follows that the ability to extend causal conclusions to all classes—that is, the external validity of $E(Y_i(1) - Y_i(0) | C)$ —is precluded in general. Using the expressions derived in the previous section, we can write:

$$\pi_c = E(D_i | Z_i = 0) - E(D_i | Z_i = 1), \quad (3.5)$$

so that the data is informative about the size of the population for whom this design can provide evidence about a causal effect. This is already a starting point to understand the extent of the external validity problem of causal estimates obtained by LATE. In the case of INVALSI data, the value of π_c is equal to 2.9% for math and 2.5% for language. This can be seen from Column (4) of Table 3.3, which reports the coefficient of Z_i in the first-stage regression of D_i on Z_i using data for all classes

in the country. This is equal to the opposite of π_C .⁷ In the South, the share of *C*-teachers grows to 6.2% for math and 4.7% for language, as can be seen from Column (6) of the same table.

In our example, the size of the compliant subpopulation is relatively small. How could one extend the conclusions drawn for a possibly small share of complying dishonest teachers to the remaining classes in the population? We follow Angrist (2004) and note that the data provide information about $E(Y_i(1) | A)$ and $E(Y_i(0) | N)$ as well. These quantities can be obtained using expressions like those we derived above (see Battistin et al., 2017, for details). For example, we have that:

$$E(Y_i(1) | A) = E(Y_i | D_i = 1, Z_i = 1),$$

$$E(Y_i(0) | N) = E(Y_i | D_i = 0, Z_i = 0).$$

The first equality holds because—in the absence of *D*-teachers—only *A*-teachers manipulate scores in the presence of monitors. Similarly, only *N*-teachers report honestly without monitors.

If potential outcomes were homogeneous across types in the population, then we would have that $E(Y_i(1) | A) = E(Y_i(1) | C)$ and $E(Y_i(0) | N) = E(Y_i(0) | C)$. If these two equalities cannot be rejected from the data, we would feel more confident about extending the results obtained for classes with complying dishonest teachers to other classes in the population.⁸

In Table 3.5, we report the comparison of $E(Y_i(1) | C)$ vis-à-vis $E(Y_i(1) | A)$ and $E(Y_i(0) | C)$ vis-à-vis $E(Y_i(0) | N)$ for Southern Italy, where the problem of manipulation is more pervasive. While the data does not reject that $E(Y_i(1) | C)$ is equal to $E(Y_i(1) | A)$, the empirical evidence suggests that $E(Y_i(0) | C)$ is much smaller than $E(Y_i(0) | N)$. For instance, as reported in Panel A of Table 3.5, for math, we have that $E(Y_i(1) | C)$ and $E(Y_i(1) | A)$ are very similar and, respectively, equal to 1.426 and 1.236 standard deviations. On the other hand, while $E(Y_i(0) | C)$ is equal to -1.662 standard deviations, $E(Y_i(0) | N)$ is much higher and equal to -0.655 standard deviations. Therefore, in this case, the data advise against the generalization of the LATE of manipulation on scores outside of the population of complying dishonest teachers.

⁷The number reported in the table is the estimate of π_C with its sign flipped. This is because the expression for share of *C* – teachers π_C is in (5). The coefficient on Z_i in the regression of D_i on Z_i identifies instead $E(D_i | Z_i = 1) - E(D_i | Z_i = 0)$, that is, the opposite of π_C .

⁸Needless to say, full homogeneity of potential outcomes across types requires also that $E(Y_i(1) | N) = E(Y_i(1) | C)$ and $E(Y_i(0) | A) = E(Y_i(0) | C)$. However, the data will never reveal $E(Y_i(1) | N)$ and $E(Y_i(0) | A)$, as we never get to observe $D_i = 1$ for *N*-teachers and $D_i = 0$ for *A*-teachers. Hence, the latter two conditions cannot be tested empirically.

Table 3.5 Average potential outcomes by type: South of Italy

	Test scores		
	Complying dishonest (C)	Always dishonest (A)	Never dishonest (N)
	(1)	(2)	(3)
A. Math			
$E(Y_i(1))$	1.426 ^a	1.236 ^a	
	(0.020)	(0.119)	
$E(Y_i(0))$	-1.453 ^a		-0.527 ^a
	(0.157)		(0.104)
N	52,505	52,505	52,505
B. Language			
$E(Y_i(1))$	1.147 ^b	1.029 ^a	
	(0.018)	(0.103)	
$E(Y_i(0))$	-1.662 ^a		-0.655 ^a
	(0.176)		(0.084)
N	52,510	52,510	52,510

$E(Y_i(1)|C)$ and $E(Y_i(0)|C)$ are obtained from 2SLS regressions as detailed in the text. $E(Y_i(1)|A)$ and $E(Y_i(0)|N)$ are computed from OLS regressions that estimate $E(Y_i|D_i = 1, Z_i = 1)$ and $E(Y_i|D_i = 0, Z_i = 0)$, respectively. All models control for a quadratic in grade enrollment, segment dummies, and their interactions. The unit of observation is the class. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and proportions of missing values in these variables. All regressions include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). ^a $p < 0.01$, ^b $p < 0.05$, ^c $p < 0.1$

3.6 Causal Reasoning with Administrative Rules: The Case of Regression Discontinuity Designs

3.6.1 Larger Classes, Worse Outcomes?

The benefits of reducing student–teacher ratios on learning, educational achievement, and eventually long-term labor market outcomes have been of long-standing concern to parents, teachers, and policy-makers. Observational studies often show a negative relationship between class size and student achievement. Yet the conclusions of such studies might be subject to the problem of self-sorting of students into smaller classes.

In many countries, class size formation depends on grade enrollment using a deterministic rule, and Italy is no exception. As discussed in Angrist et al. (2017), until 2008, class size in primary schools in Italy must be between 10 and 25. A reform in 2009 modified these limits to 15 and 27, respectively. Class formation is regulated by law, and grade enrollment above multiples of the cap to maximum size leads to the formation of a new class. To see this, consider the cap at 25 students in place until 2008. Schools enrolling up to 25 students must form one class. One additional student enrolled after 25 would force principals to form one additional

class, with an average class size of 13 students. The same idea extends to any multiple of 25 students. For example, crossing the 50-student limit is enough to form three classes instead of two and so forth. Because of the regulation in place, class size decreases sharply when enrollment moves from just below to just above multiples of 25. Angrist and Lavy (1999) called this relationship “Maimonides’ rule” after the medieval scholar and sage Moses Maimonides who commented on a similar rule in the Talmud.⁹ Exceptions to the rule in Italy are allowed in some cases. For example, a 10% deviation from the maximum (3 students) in either direction is possible at the discretion of school principals and upon the approval from the Ministry of Education. The presence of students with disabilities or special education needs is often advocated to justify non-compliance with the law. Moreover, principals can form classes smaller than 10 students in the most remote areas of the country.

By allowing actual class size to deviate from the class size mandated by law, these exceptions generate fuzziness in the relationship between actual and predicted class size. This can be seen in Fig. 3.2, where we report the average class size in the country by grade enrollment at school for second graders before 2008.¹⁰ The sawtooth-shaped solid line reports predicted class size as a function of enrollment, the Maimonides’ rule, while the dots report average actual class size by enrollment. The law predicts class size to be a non-linear and discontinuous function of enrollment. Actual class size follows predicted class size closely and more so for schools enrolling less than 75 students (which is the majority of schools in the country). In addition, discontinuities in the actual class size/enrollment relationship show up at multiples of 25 enrolled students. Given the soft nature of the rule, however, they are weaker than the sharp ones observed for predicted class size.

3.6.2 Visual Interpretation

Figure 3.3 offers a visual representation of the size of these discontinuities and is constructed using classes at schools with enrollment that falls in a $[-12, 12]$ window around the first four cutoffs shown in Fig. 3.2. Enrollment values in each window are centered to be zero at the relevant cutoff. The y-axis shows average class size conditional on the centered enrollment value shown on the x-axis. The figure also plots fitted values generated by *locally linear regression* (LLR) fits to class-level

⁹ More precisely, let f_{igkt} be the predicted class size of class i in grade g at school k in year t . We have that $f_{igkt} = \frac{r_{gkt}}{\lceil \text{int}((r_{gkt} - 1) / c_{gt}) + 1 \rceil}$, where r_{gkt} is beginning-of-the-year grade enrollment at school

k , c_{gt} is the relevant cap (25 or 27) for grade g , and $\text{int}(x)$ is the largest integer smaller than or equal to x .

¹⁰ Similar patterns hold also for the period after the 2008 reform and for fifth graders, as shown by Angrist et al. (2017).

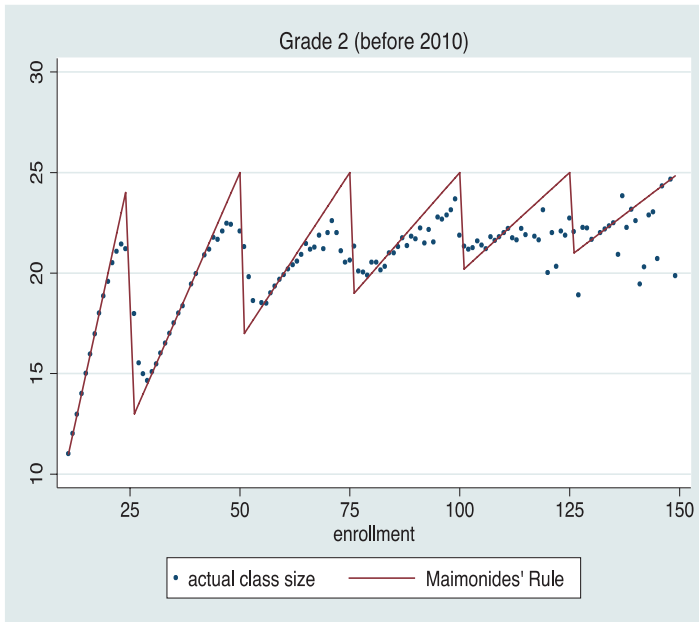


Fig. 3.2 Class size by enrollment among second-grade students in pre-reform years (Angrist et al., 2017). (It shows actual class size and class size as predicted by the Maimonides' rule in pre-reform years for second-grade students)

data, as described in Angrist et al. (2017). This representation is convenient in that one can think that small classes are those in schools with grade enrollment to the right of zero. The figure shows a clear drop at this value. Class size is minimized at about 3–4 students to the right of this value, as we would expect were Maimonides' rule to be tightly enforced.

How can we use these discontinuities in class size to assess a causal effects of class size? School enrollment may be positively correlated with test scores, for example, because larger schools are typically in urban areas, and this relationship need not be linear. However, we would be tempted to infer a causal effect of class size on test score if we observed a discontinuous change in test scores at the *exact* values of enrollment that are multiples of the maximum class size caps, where class size also discontinuously changes. This is the idea underlying the evaluation design that goes by the name of regression discontinuity (RD).

Figure 3.4 exemplifies this idea. It reports the change in average test scores as normalized enrollment moves from below to above the recentered enrollment cutoffs, separately for North and Central Italy and for the South. There is evidence of a positive discontinuity in scores as we move from below to above the cutoff in Southern Italy. Evidence of jumps for the rest of the country is instead much more limited, suggesting the possibility of causal effects of class size on learning mostly for schools in the South.

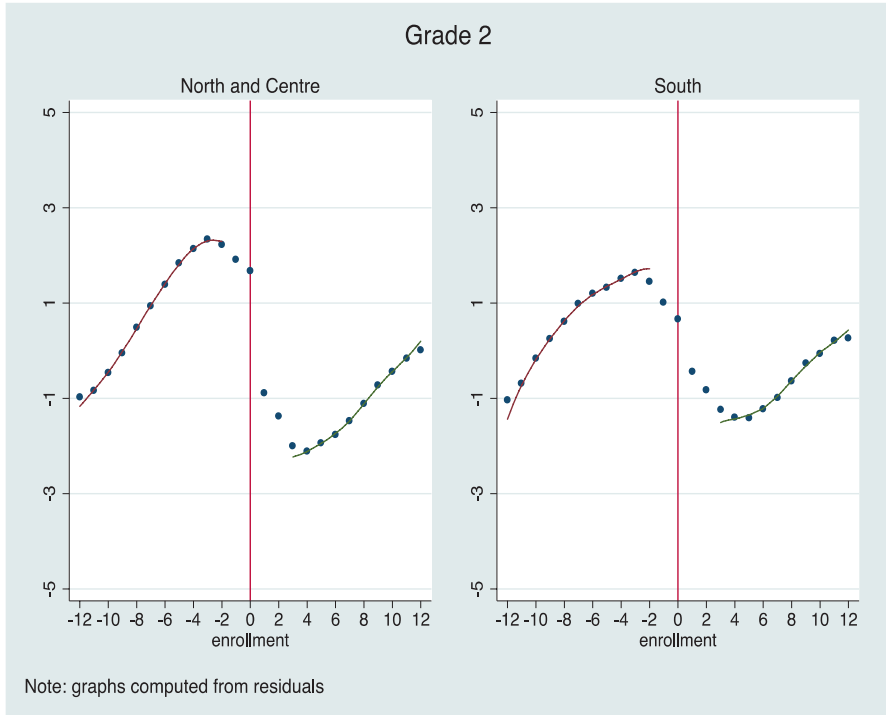


Fig. 3.3 Class size by enrollment among second-grade students, centered at the RD cutoffs (Angrist et al., 2017). (Graphs plot residuals from a regression of class size on the following controls: % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and dummies for missing values in these variables. All regressions include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). The solid line shows a one-sided LLR fit.)

The idea underlying the RD design is that the comparison of scores of classes just above and just below the enrollment cutoffs identified by the Maimonides' rule is informative of effects of class size. Still, not all classes above the cutoffs are small and not all classes below are large, because of discretion in the application of the rule. Intuitively, if compliance with the rule was perfect, then the graphical analysis would already reveal the causal effect. If compliance is not perfect, we may want to use the rule as an instrument for class size formation. Intuitively, the crucial assumption here is that the Maimonides' rule must affect performance at school only because it affects class size formation. A juxtaposition with the identification results discussed in Sect. 3.5 reveals that, in this case, the causal effect of class size on learning is identified only for schools that would form smaller classes because of compliance with the rule. We will come back to this point later in this section.

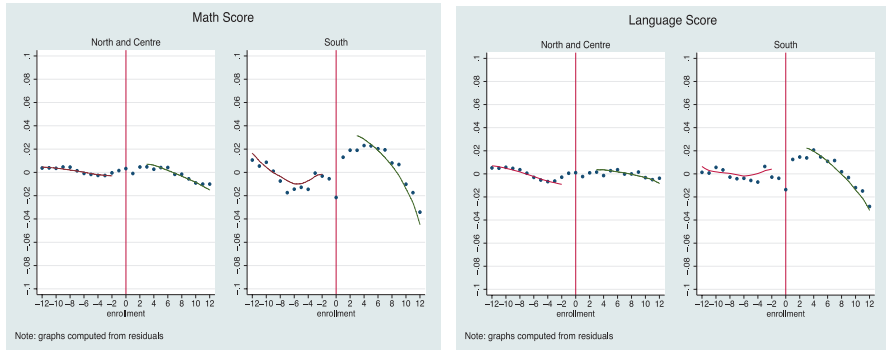


Fig. 3.4 Test scores by enrollment among second-grade students, centered at the RD cutoffs (Angrist et al., 2017), (Graphs plot residuals from a regression of test scores on the following controls: % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and proportions of missing values in these variables. All regressions additionally include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). The solid line shows a one-sided LLR fit.)

3.6.3 General Formulation of the Problem

Following our running example, the class is the statistical unit of analysis and the treatment is class size.¹¹ To ease the narrative, we distinguish between small and large classes and move to the background the possibility of a “continuous” treatment (number of students in class). Small classes will have $D_i = 1$ and large classes $D_i = 0$. In our narrative, the Maimonides’ rule predicts small classes to the right of the recentered cutoffs in Fig. 3.2. Similarly, a large class is predicted for grade enrollment at or below the cutoffs in the same figure. Potential outcomes $Y_i(1)$ and $Y_i(0)$ are the average test score that class i would get if it was small or large. Grade enrollment at school of class i is r_i . Without loss of generality and consistent with Fig. 3.3, we recentered grade enrollment at zero using a $[-12, 12]$ window around cutoffs.

3.6.3.1 The Sharp RD Design

We start our discussion by assuming full compliance of school principals with the Maimonides’ rule. In other words, we pretend that all classes with r_i at or above zero are small and that all classes with r_i below zero are large. This is equivalent to

¹¹We will drop all indexes other than i in what follows. The data contains additional dimensions, but we ignore them for expositional simplicity. One dimension is grade and year. However, scores are standardized by grade and year, so we can ignore them. As a result of this normalization, we end up having repeated measurements over time for classes at the same school. Another dimension is the reform regime. We recenter enrollment to the right cutoff depending on the regulation in place, and we, therefore, abstract from this dimension.

assuming a deterministic relationship between r_i and class size, which we express using the following notation: $D_i = 1(r_i \geq 0)$. We use this *sharp* setting to write the comparison of outcomes for classes in schools with grade enrollment in a neighborhood of the Maimonides' cutoff. The notion of cutoff proximity will be exemplified by using limits from below and above zero. Accordingly, the notation $r_i^+ = 0$ in what follows should read "just above the Maimonides' cutoff"; the notation $r_i^- = 0$ is instead "just below the Maimonides' cutoff."

We have that:

$$\begin{aligned} \lim_{r \rightarrow 0^-} E(Y_i | r_i = r) &= \lim_{r \rightarrow 0^-} E(Y_i(0) | r_i = r) + \lim_{r \rightarrow 0^-} E(D_i(Y_i(1) - Y_i(0)) | r_i = r) \\ &= \lim_{r \rightarrow 0^-} E(Y_i(0) | r_i = r), \end{aligned}$$

because in classes to the left of the Maimonides' cutoff D_i is zero so that the second term vanishes. For classes with r_i above zero, we have:

$$\begin{aligned} \lim_{r \rightarrow 0^+} E(Y_i | r_i = r) &= \lim_{r \rightarrow 0^+} E(Y_i(0) | r_i = r) + \lim_{r \rightarrow 0^+} E(D_i(Y_i(1) - Y_i(0)) | r_i = r), \\ &= \lim_{r \rightarrow 0^+} E(Y_i(0) | r_i = r) + \lim_{r \rightarrow 0^+} E(Y_i(1) - Y_i(0) | r_i = r), \end{aligned}$$

because D_i is one deterministically. It follows that the outcome difference between small and large classes at the cutoff can be written as:

$$\begin{aligned} \lim_{r \rightarrow 0^+} E(Y_i | r_i = r) - \lim_{r \rightarrow 0^-} E(Y_i | r_i = r) &= \lim_{r \rightarrow 0^+} E(Y_i(1) - Y_i(0) | r_i = r) \\ &+ \lim_{r \rightarrow 0^+} E(Y_i(0) | r_i = r) - \lim_{r \rightarrow 0^-} E(Y_i(0) | r_i = r). \end{aligned}$$

The parallel with the naïve comparison discussed in Eq. 3.1 is striking: the comparison of outcomes for small ($r_i^+ = 0$) and large ($r_i^- = 0$) classes is equal to a causal effect for units just to the right of $r_i = 0$:

$$\lim_{r \rightarrow 0^+} E(Y_i(1) - Y_i(0) | r_i = r),$$

plus a selection bias term:

$$\lim_{r \rightarrow 0^+} E(Y_i(0) | r_i = r) - \lim_{r \rightarrow 0^-} E(Y_i(0) | r_i = r),$$

measuring differences in a local neighborhood of $r_i = 0$ that would have occurred even without treatment (i.e., if class size could be only large). What conditions are needed to ensure that the latter term is zero? A closer look at the two terms in the last expression reveals an idea of *continuity*. The condition:

$$\lim_{r \rightarrow 0^+} E(Y_i(0) | r_i = r) = \lim_{r \rightarrow 0^-} E(Y_i(0) | r_i = r), \quad (3.6)$$

is sufficient to eliminate selection bias and is equivalent to assuming that the relationship between the outcome $Y_i(0)$ and grade enrollment is continuous at $r_i = 0$. This is a mild regularity condition, which most likely holds in most applications, and has a very simple interpretation: our hopes to give any causal interpretation to discontinuities in school performance observed around Maimonides' cutoffs must rest on the assumption that there would have been no discontinuity in performance crossing from $r_i^- = 0$ over to $r_i^+ = 0$ had the Maimonides' rule been irrelevant for forming a small class. Assumption (3.6) combined with its counterpart for the $Y_i(1)$ outcome:

$$\lim_{r \rightarrow 0^+} E(Y_i(1) | r_i = r) = \lim_{r \rightarrow 0^-} E(Y_i(1) | r_i = r), \quad (3.7)$$

ensures:

$$\lim_{r \rightarrow 0^+} E(Y_i | r_i = r) - \lim_{r \rightarrow 0^-} E(Y_i | r_i = r) = E(Y_i(1) - Y_i(0) | r_i = 0). \quad (3.8)$$

Assumption (3.7) brings to the problem the same regularity condition in (3.6), with a similar interpretation.

The notion of continuity of potential outcomes around Maimonides' cutoffs is evocative of the properties of a full randomization of students to small and large classes in schools with grade enrollment near $r_i = 0$. For example, assumption (3.6) can be interpreted as an independence condition between $Y_i(0)$ and D_i *locally* with respect to the Maimonides' cutoff. This is the same sort of condition that we discussed in Sect. 3.4 above. It follows that the internal validity of RD estimates obtained from (3.8) hinges upon the assumption that students in schools with values of r_i near zero are as good as randomly assigned to small and large classes, as in a local randomized experiment. In Sect. 3.6.4 below, we discuss how potential violations of such condition may arise in practice and propose some tests to assess the plausibility of this assumption.

Compared to a standard randomized experiment, we pay a price in terms of external validity, as RD estimates are internally valid only around Maimonides' cutoffs. The extrapolation of this effect away from the cutoff requires further assumptions about the global shape of the potential outcome functions, that must be discussed on a case-by-case basis. We refer the interested reader to the work by Battistin and Rettore (2008), Angrist and Rokkanen (2015), Dong and Lewbel (2015), and Bertanha and Imbens (2020).

RD estimates of causal effects are obtained from the sample analogue of the expression in (3.8).¹² The simplest way to proceed is by comparing the mean sample outcomes for small and large classes within a fixed distance from the Maimonides' cutoff $r_i = 0$. The simplicity of this estimator is very appealing, but we may

¹²Lee and Lemieux (2010) provide a thorough discussion of estimation issues in RD designs. We refer the interested reader to their survey for additional details.

encounter statistical validity issues if the data are “sparse” around the Maimonides’ cutoff. In fact, we face a trade-off. On the one hand, to enhance statistical validity, we would be tempted to enlarge the width of the neighborhood around the Maimonides’ cutoff considered for estimation. On the other hand, by so doing, we would end up using also data points far away from the cutoff. If the relationship between Y_i and r_i was not flat, this could endanger the internal validity of the design.

To minimize this trade-off, researchers often rely on semi-parametric estimators. Kernel-weighted local regressions of the outcome on a low-order (linear or quadratic) polynomial in r_i estimated separately for classes to the left and to the right of r_i are the most common option (as in Fig. 3.4). By giving a larger weight to data point that are closer to the cutoff and allowing for a non-flat relationship between test scores and enrollment, this estimator permits to enlarge sample size while maintaining internal validity. A flexible parametric regression of Y_i and r_i that uses all the available data could also be an option when sample size is small, but this may raise additional issues if high-order polynomials are adopted (see Gelman & Imbens, 2019).

3.6.3.2 The Fuzzy RD Design

When compliance with the Maimonides’ rule is far from perfect, as in Italian primary schools, the sharp setting described in the previous section no longer applies. The fuzziness introduced by non-compliance can be dealt with using the class size predicted from the Maimonides’ rule as an instrumental variable for the actual class size. The key assumption underlying this approach is that the regulation on class size formation must influence standardized tests only because the regulation affects how classes are eventually formed. This is, once again, an exclusion restriction of the form discussed in Sect. 3.5.3.3, above.

A few refinements of this idea are needed in this setting because the Maimonides’ rule yields experimental-like variation only near $r_i = 0$, implying that the “as good as random” condition in Sect. 3.5.3.2 must hold only *locally* with respect to this point. Complying classes here are those turning small because of compliance with the class size regulation when grade enrollment crosses from $r_i^- = 0$ over to $r_i^+ = 0$ (see Sect. 3.5.3.1). Moreover, the first-stage condition, which ensures that the Maimonides’ rule shapes—at least in part—the way classes in Italy are eventually formed stems from the following contrast:

$$\lim_{r \rightarrow 0^+} E(D_i | r_i = r) - \lim_{r \rightarrow 0^-} E(D_i | r_i = r). \quad (3.9)$$

Eq. 3.9 compares the share of small classes just above and just below the Maimonides’ cutoff $r_i = 0$. Contrary to the case of a sharp RD, where this contrast is one because of full compliance, fuzziness arising from it makes this quantity lower than one depending on the number of complying classes. The more severe is the extent of non-compliance, the lower will be the external validity of the causal conclusions, as we discussed in Sect. 3.5.5.

The same argument used in Sect. 3.5 extends to the case considered here and can be used to write:

$$E[Y_i(1) - Y_i(0) | C, r = 0] = \frac{\lim_{r \rightarrow 0^+} E(Y_i | r = 0) - \lim_{r \rightarrow 0^-} E(Y_i | r = 0)}{\lim_{r \rightarrow 0^+} E(D_i | r = 0) - \lim_{r \rightarrow 0^-} E(D_i | r = 0)}. \quad (3.10)$$

The expression in Eq. 3.10 reveals that a causal effect is retrieved by the ratio of the discontinuities in the outcome and in the treatment probability at the Maimonides' cutoff. This expression bears strong similarities with Eq. 3.4 above, once we assign the role played by the instrumental variable to a dummy for being above the Maimonides' cutoff, $Z_i = 1(r_i \geq 0)$. In fact, Hahn et al. (2001) showed that non-compliance leads the fuzzy RD design to be informative about a local average treatment effect, strengthening this similarity. However, the parameter uncovered by the fuzzy RD is local in two senses. First, it refers only to complying classes. Second, it yields causal conclusions only about classes with a value of r_i close to 0, limiting external validity even further.

Following the analogy to the instrumental variable case, discussed in Sect. 3.5, estimation of fuzzy RD effects is usually carried out using two-stage least square (TSLS) methods. The general idea is to instrument the treatment dummy D_i with the dummy $Z_i = 1(r_i \geq 0)$. As in the sharp RD case, researchers can choose to model the relationship between test scores and enrollment using either parsimonious local regressions or flexible global polynomial regressions. In general, and unlike in the sharp RD case, a single TSLS regression is estimated using data on both sides of the cutoff but permitting the polynomial in r_i to have a different shape on each side of the cutoff. This is done by including interaction terms between the polynomial in r_i and D_i that are instrumented by interaction terms between the polynomial in r_i and Z_i .¹³

The estimated fuzzy RD effects of class size on test scores for our running example are reported in Table 3.6 and show a negative and significant effect of class size reduction for compliers at the relevant discontinuity cutoffs. For simplicity, these are obtained using continuous class size. For instance, according to the estimates reported in Column (1) of Panel A, when we consider data for the whole of Italy, we estimate that math scores would increase by an average of 0.06 standard deviations if we decreased class size by 1 unit. As revealed by Columns (2) and (3) and in accordance with Fig. 3.4, the magnitude of such effect is much larger in Southern Italy than in the rest of the country.

¹³Further details about estimation in the fuzzy RD design are discussed in Lee and Lemieux (2010a, b).

Table 3.6 Local average treatment effect of class size on test scores (Angrist et al., 2017)

	Test scores		
	Italy	North/Center	South
	(1)	(2)	(3)
	A. Math		
Class size	-0.0609 ^a	-0.0417 ^a	-0.1294 ^a
	(0.0196)	(0.0171)	(0.0507)
N	140,010	87,498	52,512
	B. Language		
Class size	-0.0409 ^a	-0.0215	-0.0937 ^b
	(0.0155)	(0.0136)	(0.0403)
N	140,010	87,498	52,512

The table reports 2SLS estimates using class size cutoffs as an instrument. All models control for a quadratic in grade enrollment, segment dummies, and their interactions. The unit of observation is the class. Class size coefficients show the effect of 10 students. Robust standard errors, clustered on school and grade, are shown in parentheses. Control variables include % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and dummies for missing values. All regressions include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). ^ap<0.01, ^bp<0.05, ^cp<0.1

3.6.4 Validating the Internal Validity of the Design

An underlying assumption behind the approach discussed so far is that units cannot precisely manipulate their value of the running variable. For instance, suppose that parents of pupils with above-average ability could perfectly predict enrollment by school and choose to apply only for schools where enrollment is locally above the relevant cutoffs so that their pupils would systematically end up in smaller classes.¹⁴ If this was the case, then the RD design would be invalid, as the ability composition of pupils in schools where enrollment is just above and just below the cutoff would be different.

In general, if units cannot precisely manipulate their value of the score, there should be no systematic differences between units with similar values of the score. Therefore, a test for the internal validity of an RD design is to verify whether there are discontinuities in these covariates at the cutoff. If predetermined variables that correlate with the outcome are discontinuous at the cutoff, then continuity of potential outcomes is unlikely to hold. These tests are akin to the “balancing” tests presented for the pure randomization case but are carried out locally, at the cutoff.

Table 3.7 reports results for these tests and shows precisely estimated zero effects of passing the RD cutoffs on some predetermined controls, such as the share of students present in class on the day of the test, supporting the validity of this RD design.

¹⁴For instance, Urquiola and Verhoogen (2009) show evidence of discontinuities between enrollment and household characteristics in Chilean private schools.

Table 3.7 Covariate balance for class size discontinuities (Angrist et al., 2017)

	Italy		North/Center		South	
	Control mean	Treatment difference	Control mean	Treatment difference	Control mean	Treatment difference
	(1)	(2)	(3)	(4)	(5)	(6)
% in class sitting the test	0.9392 [0.0643]	0.0000 (0.0001)	0.9345 [0.0657]	0.0001 (0.0001)	0.9471 [0.061]	0.0000 (0.0001)
% in school sitting the test	0.9386 [0.0534]	0.0001 (0.0001)	0.9339 [0.0548]	0.0001 (0.0001)	0.9464 [0.05]	0.0001 (0.0001)
% in institution sitting the test	0.9374 [0.0436]	-0.0001 (0.0001)	0.9327 [0.0426]	-0.0001 (0.0001)	0.9451 [0.0441]	-0.0000 (0.0001)
N	140,010		87,498		52,512	

Columns 1, 3, and 5 show means and standard deviations for variables listed at the left. Other columns report coefficients from regressions of each variable on predicted class size, a quadratic in grade enrollment, segment dummies and their interactions, grade and year dummies, and sampling strata controls (grade enrollment at institution, region dummies, and their interactions). Standard deviations for the control group are in square brackets; robust standard errors are in parentheses. ^ap<0.01, ^bp<0.05, ^cp<0.1

3.7 Conclusion

This chapter has discussed a selected number of approaches among the most popular in the toolbox of good empiricists interested in causal relationships. Randomization, instrumental variation, and discontinuity designs are very closely related members of the same family and, when properly implemented, are thought to yield the most credible estimates of the causal effects of public interventions.

The beauty of randomized assignment is that the composition of “treatment” and “control” groups is by design not driven by any form of selection. In this case, differences in the composition of groups due to sampling variation tend to vanish as sample size increases so that the main concern should be the one of statistical validity. External validity and general equilibrium effects may also be a concern, especially if the intervention has to be implemented in different contexts or scaled up to cover a whole country.

Instrumental variation is a good way to go when randomized assignment is not viable. It seeks sources of random variation that have affected indirectly the chance of receiving “treatment.” Clearly, a good source of variability must affect only the treatment assignment and, through this, the outcome of interest. Sources of external random variation affecting at the same time both treatment allocation and the outcome will not allow to distinguish the effect of the instrument on the outcome from the effect of the treatment on the same outcome. As we have made clear, the price to pay for the lack of randomized assignment to treatment is external validity: estimates of causal effects obtained from instrumental variation are limited to the fraction of the population changing the treatment status because of the instrument. How large and comparable this fraction is with respect to the entire population is an

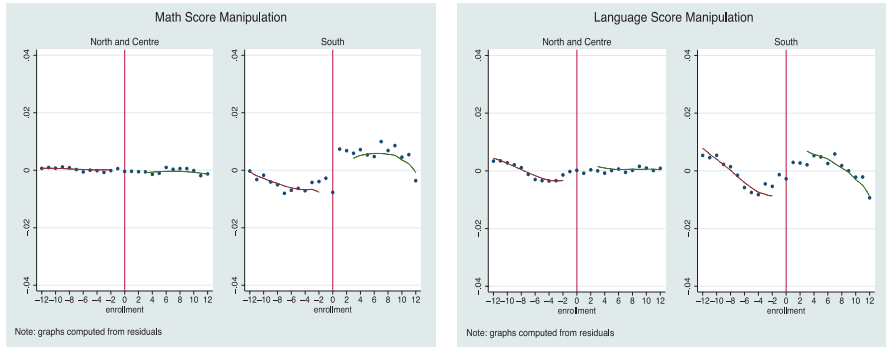


Fig. 3.5 Score manipulation by enrollment among second-grade students, centered at the RD cutoffs (Angrist et al., 2017). (Graphs plot residuals from a regression of final scores on the following controls: % female students, % immigrants, % fathers at least high school graduate, % employed mothers, % unemployed mothers, % mother NILF, grade and year dummies, and proportions of missing values in these variables. All regressions additionally include sampling strata controls (grade enrollment at institution, region dummies, and their interactions). The solid line shows a one-sided LLR fit)

empirical matter, which should be discussed on a case-by-case basis. We have discussed some test for homogeneity of potential outcomes that allow to extend validity to the whole population of interest.

Finally, the idea of regression discontinuity is most easily put across by thinking of a properly conducted randomization only locally with respect to the discontinuity cutoff. The pros are clear-cut, and the cons concern the external validity of the estimates away from the relevant discontinuity.

What else could possibly go wrong? Books and chapters like this are always written to show a path forward for the implementation of methods. The day-to-day experience as a researcher is way more intricate. For example, Figure 3.5 taken from Angrist et al. (2017) casts doubt on the validity of the assumptions used in our discussion on the effects of class size. It shows that score manipulation also changes discontinuously at $r_i = 0$ in Southern Italy, suggesting that teachers in small classes are more likely to manipulate scores. As a result, the alleged causal effect of class size on test scores in Southern Italy discussed above does not reflect more learning in smaller classes, but increased manipulation of scores in smaller classes. As discussed by Angrist et al. (2017), these findings show how class size effects can be misleading even where internal validity is probably not an issue.

This example should prompt the reader to weigh methods with a grain of salt and a proactive attitude: the most credible approach to causal inference is often a combination of different identification strategies, and its credibility must stem from the institutional context under investigation rather than clueless statistical assumptions.

Review Questions

1. Why is the naïve comparison of mean outcomes for treated and control subjects not always informative of a causal effect?

2. What are the differences between internal, external, and statistical validity of a research design?
3. How does random assignment of the treatment help to achieve internal validity?
4. Under which assumptions do natural experiments and discontinuities provide a feasible avenue to estimate causal relationships?
5. What is the price to pay in terms of validity when pursuing these empirical strategies with respect to a proper randomization?

Replication Material

Access to data and codes is available from the American Economic Association website at: <https://www.aeaweb.org/articles?id=10.1257/app.20160267>

References

- Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *Economic Journal*, 114(494), C52–C83.
- Angrist, J. D., Battistin, E., & Vuri, D. (2017). In a small moment: class size and moral hazard in the Italian mezzogiorno. *American Economic Journal: Applied Economics*, 9(4), 216–249.
- Angrist, J. D., & Imbens, G. W. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2), 467–475.
- Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, 114(2), 533–575.
- Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: an empiricist's companion*. Princeton University Press.
- Angrist, J. D., & Rokkanen, M. (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512), 1331–1344.
- Bates, M. A., & Glennerster, R. (2017). The Generalizability Puzzle. *Stanford Social Innovation Review*, Summer, 2017, 50–54.
- Battistin, E. (2016). *How manipulating test scores affects school accountability and student achievement*. IZA World of Labor.
- Battistin, E., De Nadai, M., & Vuri, D. (2017). Counting rotten apples: Student achievement and score manipulation in Italian elementary schools. *Journal of Econometrics*, 200(2), 344–362.
- Battistin, E., & Rettore, E. (2008). Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. *Journal of Econometrics*, 142(2), 715–730.
- Bertanha, M., & Imbens, G. W. (2020). External validity in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*, 38(3), 593–612.
- Bertoni, M., Brunello, G., & Rocco, L. (2013). When the cat is near, the mice won't play: The effect of external examiners in Italian schools. *Journal of Public Economics*, 104, 65–77.
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433–1436.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Rand McNally.
- Cullen, J. B., Jacob, B. A., & Levitt, S. (2006). The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica*, 74(5), 1191–1230.
- Dong, Y., & Lewbel, A. (2015). Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics*, 97(5), 1081–1092.
- Duflo, E., Glennerster, R., & Kremer, M. (2008a). Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4, 3895–3962.

- Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., & Oregon Health Study Group. (2012). The Oregon health insurance experiment: evidence from the first year. *Quarterly Journal of Economics*, 127(3), 1057–1106.
- Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, 37(3), 447–456.
- 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.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- Lee, D. S., & Lemieux, T. (2010a). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2), 281–355.
- Leigh, A. (2018). *Randomistas: How radical researchers changed our world*. Yale University Press.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
- Peters, J., Langbein, J., & Roberts, G. (2018). Generalization in the tropics—development policy, randomized controlled trials, and external validity. *The World Bank Research Observer*, 33(1), 34–64.
- Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review*, 99(1), 179–215.
- Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2), 557–598.

Suggested Readings

- Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: an empiricist's companion*. Princeton University Press.
- Duflo, E., Glennerster, R., & Kremer, M. (2008b). Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4, 3895–3962.
- Lee, D. S., & Lemieux, T. (2010b). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2), 281–355.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

