EdWorkingPaper No. 21-454

# Common support violations in clustered observational studies of educational interventions

Luke Keele
University of Pennsylvania

Matthew Lenard
Harvard University

Lindsay Page
Brown University

In education settings, treatments are often non-randomly assigned to clusters, such as schools or classrooms, while outcomes are measured for students. This research design is called the clustered observational study (COS). We examine the consequences of common support violations in the COS context. Common support violations occur when the covariate distributions of treated and control units do not overlap. Such violations are likely to occur in a COS, especially with a small number of treated clusters. One common technique for dealing with common support violations is trimming treated units. We demonstrate how this practice can yield nonsensical results in some COSs. More specifically, we show how trimming the data can result in an uninterpretable estimand. We use data on Catholic schools to illustrate concepts throughout.

VERSION: August 2021

Suggested citation: Keele, Luke, Matthew Lenard, and Lindsay Page. (2021). Common support violations in clustered observational studies of educational interventions. (EdWorkingPaper: 21-454). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/7crh-th90

# Common Support Violations in Clustered Observational Studies of Educational Interventions[*]

Luke Keele[†] Matthew Lenard[‡] Lindsay Page[§]

August 23, 2021

**Abstract**

In education settings, treatments are often non-randomly assigned to clusters, such as schools or classrooms, while outcomes are measured for students. This research design is called the clustered observational study (COS). We examine the consequences of common support violations in the COS context. Common support violations occur when the covariate distributions of treated and control units do not overlap. Such violations are likely to occur in a COS, especially with a small number of treated clusters. One common technique for dealing with common support violations is trimming treated units. We demonstrate how this practice can yield nonsensical results in some COSs. More specifically, we show how trimming the data can result in an uninterpretable estimand. We use data on Catholic schools to illustrate concepts throughout.

Keywords: Causal Inference; Clustered Observational Studies; Hierarchical/Multilevel Data; Common Support

[*] We gratefully acknowledge funding support for this work from the Spencer Foundation. The opinions expressed here do not necessarily reflect those of the Spencer Foundation. All errors are our own.

[†] University of Pennsylvania, Philadelphia, PA, Email: luke.keele@gmail.com.

[‡] Harvard University, Cambridge, MA, Email: mlenard@g.harvard.edu.

[§] Brown University, Providence, RI, Email: lindsay_page@brown.edu.

# 1 Introduction and Case Study

One common study design in education research is the clustered observational study (COS) (Page et al. 2020). In a COS, an intervention is allocated non-randomly to intact groups rather than to individuals. Given that treatment is non-randomly allocated, differences in outcomes between those who are and are not treated may reflect pretreatment differences rather than the effect of the treatment itself (Cochran 1965; Rubin 1974). Such pretreatment differences may be measurable and thus constitute overt biases. Alternatively, these differences may be unmeasured and form hidden biases. In a COS, investigators use specialized methods such as multilevel matching and/or multilevel regression models to adjust for observed confounders and remove overt bias.

In any observational study, overlap or "common support" in the distribution of baseline covariates is a key consideration. That is, investigators must ensure that treated and control units overlap in terms of their covariate distributions. One strategy to ensure common support is to prune observations where the empirical density of the control units and that for the treated units do not overlap. Areas outside of common support are particularly problematic, since they require extrapolation and therefore can generate considerable model dependence. Typically, extrapolation occurs due to the fact that there are control units that are far from the treated units in terms of the observed covariates. In such a situation, a statistical model applied to the data will extrapolate over the covariate space where we lack both treated and control units. This can lead to treatment effect estimates that are sensitive to the control units that are far from the treated units; results will be sensitive to changes in the model, since the estimates will depend on the model and not the data (King and Zeng 2006).

In short, when common support violations occur, treatment effect estimates will depend on modeling assumptions rather than the data. In this paper, we review how issues of overlap can be particularly acute in the COS design. We demonstrate how overlap issues can lead to a specific conundrum: a valid causal effect that may be useless. This is because the process of solving the problem of overlap may result in a causal effect estimate that is far from what the original study intended.

We demonstrate this issue with an investigation of the causal effect of Catholic schools. The data we use are well-known and have been used by many to investigate the question of whether Catholic schools produce better academic outcomes for students. Early evidence suggested that Catholic schools were more effective than public schools in terms of higher test scores despite spending considerably less money per pupil (Coleman et al. 1982; Hoffer et al. 1985; Coleman and Hoffer 1987). Later research challenged these findings and argued that Catholic school effectiveness was little different from public school effectiveness and that observed achievement differences between these two types of schools stemmed more from their serving different populations of students (Alexander and Pallas 1983, 1985; Goldberger and Cain 1982; Noell 1982; Willms 1985). Studies that focus on various aspects of Catholic school effectiveness have been ongoing for nearly two decades (Bryk et al. 1993; Neal 1997; Grogger and Neal 2000; Altonji et al. 2005; Reardon et al. 2009). The question of Catholic school effectiveness has also spurred considerable methodological debate and innovation regarding how to make such comparisons most appropriately (Raudenbush and Bryk 2002; Morgan 2001; Morgan and Harding 2006; Morgan and Todd 2008). This papers contributes to this debate by demonstrating how fragile estimates of the Catholic school effect can be.

The paper proceeds as follows. Section 2 reviews the COS design. Section 3 discusses statistical adjustment strategies for common support violations. Here, we focus on how enforcing common support can result in poorly defined causal estimands. We outline how the tension between common support and well-defined estimands can be especially acute in the COS context. In Section 4 we present our analysis of the Catholic school data to demonstrate how to detect and correct for common support violations. In doing so, we illustrate how enforcing common support can result in a causal estimand that is of little interest. Section 5 concludes.

## 2 Clustered Observational Studies: A Review

We begin with a review of the COS design. See Page et al. (2020) for a detailed treatment of this topic. A COS is the observational study counterpart to the clustered randomized trial (CRT). Just as in a CRT, treatment assignment occurs at the group level—in educational contexts, typically the school or classroom level—while analytical interest is focused on unit outcomes such as individual student test scores. In contrast to a CRT, however, groups in a COS are purposefully selected into treatment and control regimes. Such non-random selection means that outcomes may reflect pretreatment differences in treated and control groups rather than causal effects of treatment (Cochran 1965; Rubin 1974). Pretreatment differences in treated and control groups may be either those that are measurable, thus forming overt biases, or those that are unmeasured and leading to hidden biases. In an observational study, analysts use pretreatment covariates and a statistical adjustment strategy to remove overt biases in the hopes of consistently estimating treatment effects. Next, we formalize these ideas by introducing some notation. Throughout, we treat the school as the relevant group-level unit.

Formally, there are $J$ schools. We write $Z_j = 1$ if the school is treated and $Z_j = 0$ if the school receives the control. We define causal effects using the potential outcomes framework (Neyman 1923; Rubin 1974). Prior to treatment, each student has two potential responses: $(y_{Tij}, y_{Cij})$, where for the $i$th subject $y_{Tij}$ is observed under $Z_j = 1$, and $y_{Cij}$ is observed under $Z_j = 0$. After treatment, we only observe each individual's actual response, which is a function of potential outcomes and treatment assignment: $Y_{ij} = Z_j y_{Tij} + (1 - Z_j) y_{Cij}$. Each school $j$ contains $n_j > 1$ students, $i = 1, \ldots, n_j$. We represent pretreatment covariates with the matrix $\mathbf{x}_{ij}$ which typically includes characteristics at both the individual and group levels. For example, $\mathbf{x}_{ij}$ might contain measures of student sex and race/ethnicity as well as the percentage of students in the school who are proficient in reading based on standardized test scores. Finally, $u_{ij}$ is an unobserved, binary confounder.

In a COS, the investigator might focus on the average causal effect of the form $E[y_{Tij} - y_{Cij}]$ or the average causal effect for the treated (ATT) of the form $E[y_{Tij} - y_{Cij}|Z_j = 1]$. These contrasts are referred to as estimands, since they refer to counterfactual quantities. Our notation and estimand is consistent with the COS template, where the design focuses on a treatment that is assigned at the school level. In the context of our Catholic school study, the implication is that we have a population of intact schools, and some of these schools are selected to be Catholic schools while others are public schools. The COS template is not an entirely natural fit, since such a treatment assignment process is unlikely to occur in the course of school formation or administration. Nevertheless, imagining such a process is a useful thought experiment to encourage clarity regarding the causal effect(s) of interest. That is, we would argue that we are primarily interested in how the school-level "intervention" of being a Catholic school affects student-level test scores.

Of course, one might alternatively conceive of the Catholic school assignment process as one where students (or their families) select into attending a Catholic school. The difficulty is that in this context the assignment mechanism is not explicit. Unlike a school reform that is applied to a set of existing schools, the Catholic school assignment mechanism is necessarily ambiguous. However, the COS template is plausible for this research question and helps to emphasize the need to consider overlap in covariate distributions at both the individual and group levels.

Due to the fact that causal estimands are based on counterfactual quantities, we require a set of assumptions that allow us to identify these counterfactual quantities using observed data. First, we assume that the *Stable Unit Treatment Value Assumption* (SUTVA) holds (Rubin 1986). SUTVA has two parts: 1) the treatment levels of $Z_j$ (1 and 0) adequately represent all versions of the treatment, and 2) a student's outcomes are not affected by other students' exposures. Generally, clustered treatment assignment is assumed to reduce the likelihood of treatment spillovers as they must occur across schools. However, we allow arbitrary patterns of interference among students within the same school, given treatment selection is at the school level.

Next, we assume that treatment assignment depends on observed covariates only. Formally, we must assume that:

$$\pi_j = Pr(Z_j = 1|\mathbf{y}_{Tj}, \mathbf{y}_{Cj}, \mathbf{x}_{ij}, \mathbf{u}_{ij}) = Pr(Z_j = 1|\mathbf{x}_{ij}).$$

This states that the probability of being treated ($\pi_j$) could depend on potential outcomes, observed data, and unobservables, but here we assume that this probability only depends on observed data. In sum, the investigator asserts that there are no unobservable differences between the treated and control groups. Stated another way, we assume there are no unobserved confounders. This assumption is nonrefutable, since it cannot be tested with observed data (Manski 2007). When using a COS design, a key analytical step is testing the sensitivity of results to potential violations to this assumption (Hansen et al. 2014; Page et al. 2020).

We also assume that all schools have some probability of being treated such that $0 < \pi_j < 1$. This assumption is also referred to as the overlap or common support assumption, and it is this assumption on which we focus in this paper. In a COS, violations to this assumption result in large pre-treatment covariate imbalances which may occur due to treated clusters that are very dissimilar from any control clusters. Under such circumstances, the treated and control clusters are said to lack common support, since the covariate distributions do not overlap. Unlike the assumption of no unobserved confounders, a lack of common support can be verified with the data.

In the COS design, statistical adjustment requires specialized methods such as multilevel regression. Recently, a specialized form of matching—known as multilevel matching—has been developed for COS designs (Keele and Zubizarreta 2017; Pimentel et al. 2018). Both multilevel regression and multilevel matching are designed to remove overt bias due to differences in the covariate distributions of the treated and control group. Recently, Keele et al. (2021) demonstrated that multilevel matching combined with outcome modeling is the preferred analytic strategy for COS designs. Next, we outline a specific conundrum that can arise that is particularly acute in the COS design.

# 3 The Tension Between Common Support and Well-Defined Estimands

When the empirical covariate distributions of the treated and control units do not overlap, the primary solution is to prune observations until overlap is achieved. Retaining observations that are outside the area of common support can be particularly problematic since they require extrapolation, which can generate considerable model dependence. That is, treatment effect estimates will depend on modeling assumptions. Indeed, the farther the extrapolation is from the data, the more severe the model dependence can become. That is, if we must depend on extrapolation, our inferences depend on our model rather than the data, since relevant empirical observations do not exist in the data. The literature in statistics has developed a number of methods to handle this challenge by trimming observations to ensure overlap. For example, Crump et al. (2009) recommend discarding all units with estimated propensity scores that fall outside a set of pre-specified thresholds. They provide a set of guidelines for pre-specification of these thresholds. Alternatively, optimal subset or cardinality matching were also developed to enforce overlap (Rosenbaum 2012; Zubizarreta et al. 2014). These methods of matching not only pair units but also seek to retain the largest set of treated and control units for which common support holds by discarding the set of treated units that do not overlap with the control units.

These techniques, designed in the context of individual-level treatment assignment, have been adapted to the COS context. For example, multilevel matching can be used with a form of optimal subset matching, which allows investigators to trim treated groups to enforce overlap (Pimentel et al. 2018). Here, trimming can occur in two ways; one can trim either treated schools or treated students. However, common support violations tend to arise in school-level covariates due the fact that treatment selection tends to depend on school-level characteristics. While student-level trimming can occur in multilevel matching, its need arises more incidentally due to differences in sample sizes. As we show below, in the Catholic school data, the common support violations all occur at the school level.

In any context, trimming treated units to ensure overlap is not without consequence. Specifically, discarding treated units changes the causal estimand. In our application, the ATT is a natural target estimand, as it aligns with our interest in the effect of Catholic schools among those who attend them. However, if trimming is necessary to ensure overlap, the estimand is no longer the ATT, but some more local version based on whichever treated units are retained in the analysis. In short, if we trim any Catholic schools, the estimand would no longer be *the* Catholic school effect, but a more local Catholic school effect.

Formally, let's assume we exclude schools with covariates outside set $\mathbb{A}$, a subset of school-level covariates. $\mathbb{1}_{\mathbf{x}_{ij}\in\mathbb{A}}$ is an indicator function for the event that $\mathbf{x}_{ij}$ is an element of the set $\mathbb{A}$. For example, let $\mathbb{1}_{\mathbf{x}_{ij}\in\mathbb{A}}$ be an indicator that a school is co-ed, thus excluding all single-sex schools from the sample. The estimand is now $\mathbb{1}_{\mathbf{x}_{ij}\in\mathbb{A}}E\left[y_{Tij} - y_{Cij} \mid Z_j = 1\right]$, a subsample treatment effect defined by the school-level covariates in $\mathbb{A}$, as the estimand now depends on school-level covariates. Generally changing the estimand in this way is viewed as unproblematic if the goal is to estimate the effect of a treatment in cases where the data do not represent a well-defined population (Rosenbaum 2012). As such, discarding treated units is a reasonable decision in an observational study when interest is in a marginal population that might or might not receive the treatment of interest rather than a known, a priori well-defined population.

In COS designs, however, the treated units are often well-defined populations. For example, in a COS, it is common for the intervention to be applied to a well-defined set of schools, as in the population of Catholic schools throughout the US. In such instances, discarding treated units may result in a study with a causal effect that tells us less than the initial effect of interest. To further complicate matters, overlap issues are likely to be more common in COS designs. Why? Often sample sizes are simply more limited in a COS. In the context of education, this is because many school-level interventions are implemented within a single district, which this limits the number of schools in a given study. Even very large school districts may have only few hundred schools. When the pool of control units is limited, issues of overlap tend to be more acute. Next, we demonstrate this tension using the Catholic school data.

## 4 Empirical Analysis

In our analysis, we use data that are a public release of the 1982 High School and Beyond (HS&B) survey. This public release includes records for 7,185 high school students from 160 schools. Of these schools, 70 are Catholic schools and are thus considered treated in this application, while the remainder are public high schools and thus serve as a reservoir of controls from which we will identify matched comparisons. In the data, we observe some measures at the student level and other measures at the school level. The data are a subset of the data used in a pioneering article on the use of multilevel regression with education data by Lee and Bryk (1989). A version of these data are also used in Raudenbush and Bryk (2002).

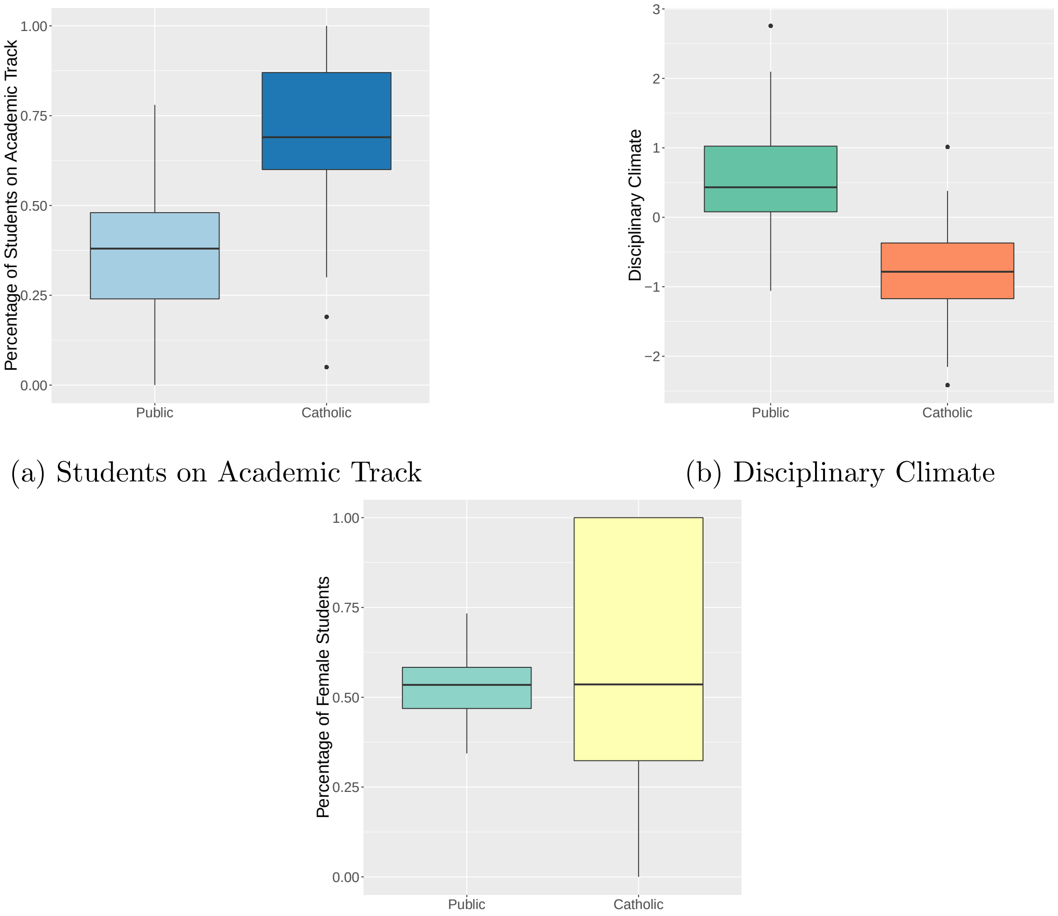
The data we use here contain three student-level covariates. The first measure is an indicator for whether or not the student is female; the second measure is an indicator for whether a student belongs to a particular racial/ethnic group, and the final measure is a scale for socioeconomic status (SES). Three of the school-level measures are simply school-level averages of these student-level measures. That is, we have a measure of the percentage of students in the school that is female, the percentage that is minority, and the average school-level SES. Three additional school-level measures are also available: total enrollment; the percentage of students on an academic track; and a measure of disciplinary climate. The disciplinary climate variable is a composite measure created from a factor score on measures of the number of attacks on teachers, fights, and other disciplinary incidents. Our outcome is interest is student performance on a standardized test of mathematics achievement in the second year of high school. Replication files for our analysis can be found at X.

## 4.1 Common Support

One sensible practice in data analysis for a COS is checking that treated and control units overlap in terms of the covariate distributions. Figure 1 contains boxplots that compare the distributions of three of the school-level covariates by school type. The results are illuminating on several fronts. First, Figure 1a shows the distribution for the percentage of students on an academic track. It is immediately clear that there is limited overlap in the distributions. The same is true for the disciplinary climate covariate plotted in 1b. Of particular interest is the difference across sex distributions for Catholic and public schools in Figure 1c. Notably, none of the public schools are single sex, and no public school has a student body that is either more than 70% female or less than 35% female. As might be expected, however, 38 of the 70 Catholic schools are either all female or all male and another school is nearly 98% female. These boxplots highlight the danger of not enforcing overlap in an observational study. Including the single-sex Catholic schools in our analysis would confound the Catholic school effect with the single-sex school effect. To ensure common support on this covariate, we must exclude the 38 single-sex Catholic schools, since we have no comparable control units among public schools. Per the discussion above, discarding these 38 schools changes the estimand. Now we will no longer estimate the Catholic school effect. Instead, the best we can do is focus on the co-ed Catholic school effect. Arguably, this is still a highly relevant effect, but it is notably distinct from *the* Catholic school effect.

## 4.2 Balance

Next, we examine the data before matching to understand the level of comparability across Catholic and public schools prior to matching. Table 1 contains balance statistics for the data after removing the 38 single sex Catholic schools. The table includes means as well as the standardized differences in means. A standardized difference is the difference in treated



(a) Students on Academic Track

(b) Disciplinary Climate

(c) Percentage Female

Figure 1: Boxplots of distributions for three covariates in the HS&B Catholic school data.

and control means divided by the standard deviation before matching for the unmatched sample. The literature suggests that the standardized difference should be less than 0.20 (Ho et al. 2007; Stuart 2010), though a more stringent criterion of 0.10 is suggested by others (Mamdani et al. 2005; Rosenbaum 2010). The results in Table 1 demonstrate that the Catholic and public school populations are quite similar on some covariates but very different on others. In particular, public schools have much larger enrollments, a smaller share of students on an academic track, considerably different disciplinary climates, and students from lower SES backgrounds, on average.

Table 1: Balance Before Matching

| | Catholic School Mean | Public School Mean | Std. Diff. |
|---|---|---|---|
| | Student-Level Covariates | | |
| Minority Student | 0.30 | 0.25 | 0.10 |
| Student SES | 0.15 | -0.15 | 0.39 |
| | School-Level Covariates | | |
| % Students Minority | 30 | 25 | 0.15 |
| Enrollment | 797.33 | 1309.34 | -0.94 |
| % Students on Academic Track | 71 | 37 | 1.86 |
| Disciplinary Climate Scale | -0.82 | 0.54 | -2.08 |
| School SES Average | 0.15 | -0.15 | 0.76 |

Note: "Std. Diff." refers to the standardized difference in means.

Next, we seek to use matching to find a set of public schools that are comparable to the remaining set of co-ed Catholic schools. We use the `matchMulti` package in `R` to implement a match that is designed to balance both student- and school-level covariates optimally (Pimentel et al. 2018; Keele and Pimentel 2016). See Keele et al. (2021) for an introduction to this type of matching in educational applications. This multilevel matching allows for matching at either the group level or both the group and individual level. In this application, we paired schools but not students within schools, since the group-level treatment is administered to all students within a school. In the remainder of our discussion, we do not report balance results for student-level covariates, since they were well balanced under all of the matches we completed. Table 2 contains the balance results for a match without any additional balance constraints. While this first match improves on the balance from the unmatched data, there are still very large imbalances. One way to improve further balance is to prioritize balance on covariates with large imbalances. With prioritization, we can focus on balancing the covariates with the largest imbalances. In the second match, we prioritized balance on the academic track and disciplinary climate covariates. In the software, prioritization is implemented by ordering the covariates in terms of priority in a prioritization list. As the results in Table 2 show, this prioritization does little to improve balance. In a third match, we increased the prioritization on these two covariates, but again balance fails to improve.

Table 2: Balance results for five different matches seeking to find comparable public schools.

| | Unmatched | Match 1 | Match 2 | Match 3 | Match 4 | Match 5 |
|---|---|---|---|---|---|---|
| % Students Minority | -0.14 | 0.25 | 0.15 | 0.12 | 0.20 | 0.04 |
| Enrollment | -1.03 | -0.91 | -1.08 | -1.11 | -1.08 | -0.70 |
| % Students on Academic Track | 1.75 | 1.48 | 1.29 | 1.32 | 0.62 | 0.55 |
| Disciplinary Climate Scale | -1.84 | -1.40 | -1.45 | -1.51 | -1.11 | -0.97 |
| School SES Average | 1.22 | 0.85 | 0.69 | 0.66 | 0.09 | -0.20 |

Note: Cell entries are standardized differences in means.
Match 1 includes no balance refinements.
Match 2 prioritizes balance on Academic Track and Disciplinary Climate.
Match 3 refines the balance prioritization on Academic Track and Disciplinary Climate.
Match 4 Retains 20 Catholic schools.
Match 5 Retains 10 Catholic schools.

When large imbalances remain after matching, the source of the problem is typically a lack of overlap, as was evident in Figure 1. The next step in the matching process is to trim treated units. In a COS design, trimming treated units implies discarding entire schools. We adjust the algorithm so that it will discard the most dissimilar set of Catholic schools and produce the optimal balance for the subsetted data. We applied this optimal subsetting to the data next. First, from the 32 co-ed Catholic schools, we discarded 12 Catholic schools resulting in a match with 20 Catholic schools and 20 public schools (Match 4). Next, we discarded 22 Catholic schools resulting in a match with 10 Catholic schools and 10 public schools (Match 5). Table 2 contains the balance statistics for these two additional matches. The use of optimal subsetting improves balance significantly. However, even optimal subsetting is unable to reduce all the of the standardized differences to an acceptable level.

Table 3: Estimates of the effect of Catholic Schools on Math Achievement Scores

| | Point Estimate | 95% Confidence Interval | p-value |
|---|---|---|---|
| Matching Only | | | |
| Match 1 | 0.25 | [0.11, 0.39] | 0.001 |
| Match 2 | 0.22 | [0.08, 0.36] | 0.003 |
| Match 3 | 0.22 | [0.08, 0.36] | 0.003 |
| Match 4 | 0.01 | [-0.11, 0.31] | 0.343 |
| Match 5 | 0.02 | [-0.21, 0.26] | 0.818 |
| Matching and Outcome Adjusted | | | |
| Match 1 | -0.015 | [-0.18, 0.15] | 0.849 |
| Match 2 | 0.002 | [-0.16, 0.17] | 0.979 |
| Match 3 | -0.019 | [-0.19, 0.15] | 0.818 |
| Match 4 | 0.08 | [-0.14, 0.3] | 0.451 |
| Match 5 | 0.082 | [-0.43, 0.6] | 0.647 |

Note: Outcome is standardized math achievement score. Results are from a mixed model with random intercept.

To estimate the treatment effect estimate for Catholic schools, we used hierarchical linear models. Specifically, we regressed student-level math scores on the school-level treatment indicator using the matched data, and we included a random effect for the intercept. We report two sets of estimates. In the first set, we did not include school-level covariates in the model used to estimate treatment effects. In the second set, we included the school-level covariates in the model. Including these covariates in the outcome model allows for additional bias reduction (Keele et al. 2021). We estimated the treatment effect for all five matches and report results in Table 3.

One clear pattern emerges. In the first three matches, which include all 32 co-educational Catholic schools, we observe a clear Catholic school effect, such that students in Catholic schools outperform students in public schools, on average. The difficulty is that we do not know if these effects are produced by a genuine treatment effect or the fact that, as the balance tables showed, Catholic schools are very different from this set of matched public schools. That is, a considerable amount of overt bias remains after matching. We removed that bias in two different ways. First, we aimed to reduce this bias using optimal subsetting. Second, we included school-level covariates in the outcome model. Under either approach, the estimated Catholic school effect is reduced substantially and is no longer statistically significant. In sum, one we fully adjust for observed confounders, there is little evidence of a Catholic school effect.

However, an additional, more subtle, issue is at play here. For the moment, let's assume that the data are balanced after we use trimming to handle the non-overlapping covariate distributions (e.g., Matches 4 and 5). If so, we could interpret the estimates from Matches 4 or 5 as valid treatment effects assuming that no unobservable factors influence the assignment of the Catholic school treatment. The difficulty is that in this application, our estimand, the causal quantity of interest, has been altered by the adjustment for the observed confounders. As we noted above, none of the estimates in Table 3 are estimates of the ATT, since we discarded the 38 single-sex Catholic schools. Thus for matches 1-3, the estimand is the ATT for co-ed Catholic schools. For matches 4 and 5, however, the causal estimand now depends on the observed covariates, since we trimmed the treated group further. For matches 4 and 5, the estimand is the ATT for the subset of co-ed Catholic schools with observed covariates like those in the Table 2. Given the highly specific form of this estimand, a legitimate question is whether one can interpret the treatment effect estimates for matches 4 and 5 as Catholic schools effects at all. The final subset of Catholic schools that look like public schools share little in common with the population of Catholic schools in these data and are likely do not resemble the broader Catholic school population. In short, in seeking to remove the common support violations, we are left with an uninterpretable estimand—one that can no longer be reasonably interpreted as any sort of Catholic school effect. A different solution to this issue would be to expand the control group to include more public schools that are comparable to Catholic schools, but this would require a new data source.

Is this particular pathology related to matching itself? The answer is no. In fact, we would argue that regression is a particularly inferior method of covariate adjustment for causal questions of this type. First, a regression model will simply extrapolate over the lack of common support, and our inference would rely solely on the modeling assumptions. Moreover, the treatment effect estimate from a regression model is also a weighted combination of treated and control units, but the weighting scheme isn't transparent (Aronow and Samii 2016; Morgan and Winship 2014). More troubling, any remaining imbalance after adjustment via regression is also not apparent to the investigator. A key benefit of matching is that it encourages clarity in describing the sample that is comparable and the extent to which imbalances remain after statistical adjustments.

# 5 Discussion

Our case study exemplifies key difficulties that arise in the analysis of clustered observational data. While common support violations are likely in any observational study, they tend to arise commonly in COSs given that the number of clusters tends to be small. We illustrate how a common technique for alleviating overlap violations, trimming the data, can lead to uninterpretable estimands. That is, once too many treated units are trimmed from the data, the estimand may no longer correspond to an interpretable quantity, and associated estimates may not generalize well to any population of interest. In our case study, our estimate likely pertained to an idiosyncratic subset of Catholic schools that differed from the broader population of Catholic schools in the data. Unfortunately, the best solution would be to obtain additional data to expand the pool of controls schools. Of course, additional data collection may be impossible in some contexts and especially when working with an extant data set.

One question is whether this problem is likely to be widespread. We have encountered this problem in related work (Keele et al. 2021). In general, we expect this problem to be more common when an analysis is restricted to a single school district or when treatment selection of schools is highly purposeful. Our results clearly demonstrate that analysts need to carefully assess whether there is overlap in covariate distributions and understand how comparable the treated and control units are after adjustments for balance have been made. Moreover, information about overlap and covariate balance should be clearly communicated to readers, so they can understand how comparable the treated and control groups are after statistical adjustment and how procedures to achieve balance influence interpretability and generalizability of results.

# References

Alexander, K. L. and Pallas, A. M. (1983), "Private Schools and Public Pol- icy: New Evidence on Cognitive Achievement in Public and Private Schools," *Sociology of Education*, 56, 170–182.

— (1985), "School Sector and Cognitive Performance: When Is a Little a Little," *Sociology of Education*, 58, 115–128.

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005), "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151–184.

Aronow, P. M. and Samii, C. (2016), "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science*, 60, 250–267.

Bryk, A. S., Lee, V. E., and Holland, P. B. (1993), *Catholic Schools and the Common Good*, New York, NY: Basic Books.

Cochran, W. G. (1965), "The Planning of Observational Studies of Human Populations," *Journal of Royal Statistical Society, Series A*, 128, 234–265.

Coleman, J. S. and Hoffer, T. (1987), *Public and Private Schools: The Impact of Communities*, New York, NY: Basic Books.

Coleman, J. S., Hoffer, T., and Kilgore, S. (1982), *High School Achievement: Public, Catholic, and Private Schools Compared*, New York, NY: Basic Books.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 96, 187–199.

Goldberger, A. S. and Cain, G. G. (1982), "The Causal Analysis of Cognitive Outcomes in the Coleman, Hoffer, and Kilgore Report," *Sociology of Education*, 55, 103–122.

Grogger, J. and Neal, D. (2000), "Further Evidence on the Effects of Catholic Secondary Schooling," *Brookings-Wharton Papers on Urban Affairs*, 151–201.

Hansen, B. B., Rosenbaum, P. R., and Small, D. S. (2014), "Clustered Treatment Assignments and Sensitivity to Unmeasured Biases in Observational Studies," *Journal of the American Statistical Association*, 109, 133–144.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007), "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," *Political analysis*, 15, 199–236.

Hoffer, T., Greeley, A. M., and Coleman, J. S. (1985), "Achievement Growth in Public and Catholic Schools," *Sociology of Education*, 58, 74–97.

Keele, L. and Pimentel, S. (2016), *matchMulti: Optimal Multilevel Matching using a Network Algorithm*, r package version 1.1.5.

Keele, L. J., Lenard, M., and Page, L. (2021), "Matching Methods for Clustered Observational Studies in Education," *Journal of Educational Effectiveness*, in press., unpublished Manuscript.

Keele, L. J. and Zubizarreta, J. (2017), "Optimal Multilevel Matching in Clustered Observational Studies: A Case Study of the Effectiveness of Private Schools Under a Large-Scale Voucher System," *Journal of the American Statistical Association*, 112, 547–560.

King, G. and Zeng, L. (2006), "The dangers of extreme counterfactuals," *Political Analysis*, 14, 131–159.

Lee, V. E. and Bryk, A. S. (1989), "A Multilevel Model of The Social Distribution of High School Achievement," *Sociology of Education*, 62, 172–192.

Mamdani, M., Sykora, K., Li, P., Normand, S.-L. T., Streiner, D. L., Austin, P. C., Rochon, P. A., and Anderson, G. M. (2005), "Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding," *Bmj*, 330, 960–962.

Manski, C. F. (2007), *Identification For Prediction And Decision*, Cambridge, Mass: Harvard University Press.

Morgan, S. L. (2001), "Counterfactuals, Causal Effect Heterogeneity, and the Catholic School Effect on Learning," *Sociology of Education*, 74, 341–374.

Morgan, S. L. and Harding, D. J. (2006), "Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice," *Sociological Methods & Research*, 35, 3–60.

Morgan, S. L. and Todd, J. J. (2008), "A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects," *Sociological Methodology*, 38, 231–281.

Morgan, S. L. and Winship, C. (2014), *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, New York, NY: Cambridge University Press, 2nd ed.

Neal, D. A. (1997), "The Effects of Catholic Secondary Schooling on Educational Achievement," *Journal of Labor Economics*, 15, 98–123.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science*, 5, 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed (1990).

Noell, J. (1982), "Public and Catholic Schools: A Reanalysis of 'Public and Private' Schools," *Sociology of Education*, 55, 123–132.

Page, L. C., Lenard, M., and Keele, L. (2020), "The Design of Clustered Observational Studies in Education," *AERA Open*, 6, 1–14.

Pimentel, S. D., Page, L. C., Lenard, M., and Keele, L. J. (2018), "Optimal Multilevel Matching Using Network Flows: An Application to a Summer Reading Intervention," *Annals of Applied Statistics*, 12, 1479–1505.

Raudenbush, S. W. and Bryk, A. (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods*, Thousand Oaks, CA: Sage.

Reardon, S. F., Cheadle, J. E., and Robinson, J. P. (2009), "The Effect of Catholic Schooling on Math and Reading Development in Kindergarten Through Fifth Grade," *Journal of Educational Effectiveness*, 2, 45–87.

Rosenbaum, P. R. (2010), *Design of Observational Studies*, New York: Springer-Verlag.

— (2012), "Optimal Matching of an Optimally Chosen Subset in Observational Studies," *Journal of Computational and Graphical Statistics*, 21, 57–71.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 6, 688–701.

— (1986), "Which Ifs Have Causal Answers," *Journal of the American Statistical Association*, 81, 961–962.

Stuart, E. A. (2010), "Matching Methods for Causal Inference: A review and a look forward," *Statistical Science*, 25, 1–21.

Willms, D. J. (1985), "Catholic-School Effects on Academic Achievement: New Evidence from the High School and Beyond Follow-up Study," *Sociology of Education*, 58, 98–114.

Zubizarreta, J. R., Paredes, R. D., and Rosenbaum, P. R. (2014), "Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile," *The Annals of Applied Statistics*, 8, 204–231.