ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 20-234

The Design of Clustered Observational Studies in Education

Lindsay C. Page
University of Pittsburgh

Matthew A. Lenard
Harvard University

Luke Keele
University of Pennsylvania

Clustered observational studies (COSs) are a critical analytic tool for educational effectiveness research. We present a design framework for the development and critique of COSs. The framework is built on the counterfactual model for causal inference and promotes the concept of designing COSs that emulate the targeted randomized trial that would have been conducted were it feasible. We emphasize the key role of understanding the assignment mechanism to study design. We review methods for statistical adjustment and highlight a recently developed form of matching designed specifically for COSs. We review how regression models can be profitably combined with matching and note best practice for estimates of statistical uncertainty. Finally, we review how sensitivity analyses can determine whether conclusions are sensitive to bias from potential unobserved confounders. We demonstrate concepts with an evaluation of a summer school reading intervention in Wake County, North Carolina.

VERSION: May 2020

Suggested citation: Page, Lindsay C., Matthew A. Lenard, and Luke Keele. (2020). The Design of Clustered Observational Studies in Education. (EdWorkingPaper: 20-234). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/y9ek-px57

The Design of Clustered Observational Studies in Education*

Lindsay C. Page† Matthew A. Lenard‡ Luke Keele§

May 20, 2020

Abstract

Clustered observational studies (COSs) are a critical analytic tool for educational effectiveness research. We present a design framework for the development and critique of COSs. The framework is built on the counterfactual model for causal inference and promotes the concept of designing COSs that emulate the targeted randomized trial that would have been conducted were it feasible. We emphasize the key role of understanding the assignment mechanism to study design. We review methods for statistical adjustment and highlight a recently developed form of matching designed specifically for COSs. We review how regression models can be profitably combined with matching and note best practice for estimates of statistical uncertainty. Finally, we review how sensitivity analyses can determine whether conclusions are sensitive to bias from potential unobserved confounders. We demonstrate concepts with an evaluation of a summer school reading intervention in Wake County, North Carolina.

Keywords: Causal Inference; Hierarchical/Multilevel Data; Observational Study; Optimal Matching

*For comments and suggestions, we thank Brooks Bowden, Michael Gottfried, Matthew Kraft and Luke Miratrix. We are grateful to current and former Wake County Public School System staff, especially Martina Lowry, Timothy Marshmon, Brad McMillan, Colleen Paeplow, Melanie Rhoads, and Sonya Stephens. 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 Pittsburgh, Pittsburgh, PA, Email: lpage@pitt.edu

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

§University of Pennsylvania, Philadelphia, PA, Email: luke.keele@uphs.upenn.edu

1 Introduction

The effectiveness of educational interventions is often evaluated using clustered randomized trials (CRTs) in which random assignment occurs at the level of the group rather than at the level of the individual student (Raudenbush 1997; Hedges and Hedberg 2007). Given the natural groupings of students within classrooms and schools as well as the practical and political challenges associated with individual-level random assignment in educational settings, CRTs are a common tool for drawing causal inferences about educational policies, practices, and innovations.

When an intervention is randomly assigned, differences in outcomes between groups which are and are not treated can be causally attributed to the intervention. However, randomized trials, even with assignment at the group level, are not always feasible for political, cost, ethical or other reasons. In such cases, researchers must turn to observational analyses. One alternative to a CRT is a clustered observational study (COS). In a COS, treatment assignment occurs at the group level; for example, whole schools are selected for treatment, but treatment assignment occurs through some uncontrolled process.

Whereas the literature on observational studies for deriving causal inferences when treatment selection occurs at the individual level is robust and well developed (Rubin 2007, 2008), the same is not true with regard to COSs. In fact, the literature on COSs remains underdeveloped, with no consensus on best practices. This is particularly surprising in the sphere of educational research, where treatment selection often occurs at the cluster level. In this paper, our aim is to outline comprehensively the key considerations and steps in the design and conduct of a COS. We highlight important ways in which COSs are different from observational studies where treatment is assigned at the individual level, and we propose a framework for the *design* of COSs. In doing so, we review aspects of study design for observational studies and highlight how the analyst must alter standard principles to handle clustered treatment assignment.

Our framework is built on the counterfactual model for causal inference. We begin our discussion

in the following section by advocating that COSs should be designed following the principle of target trial emulation, meaning that they are designed according to the cluster randomized trial that the analyst would have ideally carried out. Although the idea of target trial emulation is not new, we highlight considerations that are unique to the context of a COS and associated hierarchical data.

Next, we discuss the importance of understanding the process through which sites were selected into the treatment condition under study. We discuss why cluster—rather than individual-level—treatment assignment is often preferable for deriving causal inferences, as it can protect against selection bias even when selection for treatment is non-random. We then introduce notation, articulate assumptions that must hold for making causal inferences in the context of a COS, and argue for the central role of sensitivity analysis to understand whether conclusions might be sensitive to the presence of an unobserved confounder. Next, we highlight possible approaches to and general considerations regarding statistical analysis. In this section of the paper, we discuss a new form of matching designed explicitly for COSs. In sum, the goal of our paper it to make a methodological contribution with respect to design rather than analysis. That is, we are not introducing a new statistical model—instead, we are introducing readers to critical aspects of COSs related to their design.

We illustrate concepts with an evaluation of a summer reading intervention in Wake County, North Carolina. We introduce this application in more detail below, and use it throughout as an example of the ways in which the study design process is important. In particular, much of the design process for a COS requires that the investigator gather information on how the treatment was assigned and structure the analysis to reflect this assignment process. Although we find no evidence that this reading intervention led to improved student outcomes, all of the key elements of study design occur prior to the examination of study impacts. Further, careful study design should lead the policymaker to place more stock in the results, even if the magnitudes are not educationally meaningful.

2 Research Design Principles for Clustered Observational Studies

Here, we outline key considerations in the design of clustered observational studies. We begin by discussing the concept of target trial emulation.

2.1 Target Trial Emulation

Target trial emulation calls for using design principles from randomized trials and applying them to the analysis of observational data (Hernán and Robins 2016). Under the target trial approach, the investigator explicitly ties the design and analysis of the observational study to the experimental trial it is emulating, and causal estimands of interest are derived from the hypothetical target trial. Whether the causal effect from this target trial can be estimated consistently using observational data depends on certain assumptions, known as identification assumptions. In the case of observational studies, investigators typically assume that any differences between treated and control groups are observable—that there are no unobserved differences between the two groups—and that any observable differences can be handled through covariate adjustment. We return to the concepts of identification assumptions and covariate adjustment in more detail below.

The purpose of target trial emulation is to improve the quality of observational studies through the application of trial design principles. For example, in an experimental study, the sample and study design are clearly delineated to enable randomization. In contrast, observational studies, particularly those conducted after program implementation, often necessitate some level of investigation to inform decisions about and articulation of sample construction and study design. Imagining the hypothetical randomized experiment that would generate observational data under study (Rubin 2008; Cochran and Rubin 1973) appears simple at first. However, this consideration can be a difficult exercise in practice since the analyst might conceive of several different hypothetical experiments that generate a specific dataset. Here, we outline two possible target cluster

randomized trials that are common in the context of educational interventions. Specifically, we introduce two study designs that correspond to situations where (1) whole groups are assigned to a given treatment and (2) subsets of larger groups are assigned to a given treatment according to certain qualifying characteristics.

2.1.1 Design 1: Clustered Treatment Assignment

Design 1 pertains to circumstances where complete clusters (e.g., whole classrooms or schools) are selected for treatment, and all units within a given cluster either receive or do not receive treatment. Under Design 1, we seek to mimic a clustered randomized trial in which treatment assignment occurs at the cluster level, and all units within selected clusters receive (or at least are intended to receive) treatment. Under the COS analogue for this design, cluster-level covariates are critical, given that the assignment mechanism is at the cluster level, and assignment is presumed to have been made on the basis of cluster-level characteristics alone. Such a design would be appropriate for assessing the impact of any intervention that is applied to entire schools or schoolwide reform efforts, such as the Success for All reading program (Borman et al. 2007).

2.1.2 Design 2: Clustered Treatment Assignment for Student Subsets

CRTs often rely on the assumption that the data are either based on all units in the cluster or a random sub-sample of the cluster, such that the selected units are representative of the cluster as a whole (Torgerson 2001; Donner and Klar 2004). However, educational interventions are often allocated in a purposeful, targeted (e.g., non-random) fashion within clusters. Under Design 2, the target trial is a CRT with non-random, student-level selection into the treatment within schools. That is, clusters are assigned to treatment and control, but within the selected clusters only some units are targeted for treatment. This would be the case if the intervention were designed for students who were struggling academically, for example. As such, the causal estimand is a group-level contrast for a set of students within the school who are at risk for the treatment.

The critical distinction between Designs 1 and 2 is that under Design 2, final treatment assignment of an individual depends on both school- and student-level characteristics. The selection of units for treatment within the cluster is analogous to non-random attrition. In a CRT, the investigator would need to correct for this selection bias. The same is true in a COS. That is, while schools or classrooms might be selected for treatment, if the treatment is only applied to a subset of students within those clusters, the analyst may need to model a second selection mechanism. This implies that in the context of a COS, our procedures for covariate adjustment must account for data at both the school and student levels. Next, we introduce our motivating example and consider the target CRT with which it most closely aligns.

2.2 Motivating Application: A Summer School Reading Intervention

In 2012, the North Carolina state legislature required that students who did not meet state standards at the end of third grade participate in summer reading instruction or risk grade retention.[1] In summer 2013, the Wake County Public School System (hereafter, Wake County) selected myON, a type of computer-aided instruction (CAI) for implementation at selected summer school sites. The goal of introducing myON was to boost reading comprehension among summer school attendees, the majority of whom were from low-income backgrounds. myON, a product of Renaissance Learning, is a web-based product that serves primarily as an electronic reading device. The software provides students with access to a library of thousands of books and suggests titles to students based on their preferences and reading ability. Students at sites selected to implement myON used the program for up to one-half hour during the daily summer school literacy block and could continue using the program at home if they had a device and internet connection. At the time of its launch in Wake County, the developers claimed that students using myON would improve comprehension through access to more than 10,000 digital books that include "multimedia supports, real-time reporting and assessments and embedded close reading tools" (Corp 2015). Given the prevalence and cost of such supplementary curricular programs, rigorous, independent assessment of these tools is critical to sound investment

[1] North Carolina General Assembly, § 115C-83, Part 1A. North Carolina Read to Achieve Program.

decisions by school districts and other educational agencies.

The study sample includes 3,434 summer school students from 49 different Wake County elementary schools who attended summer school at one of 19 different sites. Due to technical constraints, some summer school sites used myON, while other sites did not. As such, all students in a school were exposed to the myON treatment if they attended summer school at a selected site. In a COS designed to study the effects of myON, Design 2 would appear to be the most relevant target trial. This is because myON was assigned to schools but only students required to attend summer school by virtue of their performance on the state summative reading exam score were exposed to the treatment. Therefore, we are most interested in contrasting outcomes for groups of summer school students who were and were not exposed to myON. In any observational study, a key step in the design process is to understand the process through which individuals or groups were selected into treatment. We next discuss the general benefit of cluster-level selection with the goal of deriving causal inferences and more explicitly investigate the process by which certain summer school sites were selected to receive myON.

2.3 Notation

The principle of target trial emulation applies to study design, broadly, as well as analytic notation, more specifically. Here, we apply this principle to structure our notation which is applicable to both CRTs and COSs. A defining feature of a clustered study is that individual units (e.g., students) are organized within clusters (e.g., schools) and assigned to a treatment or control condition at the cluster level rather than the individual unit level. Generally, for applications in which students are nested within schools, each school j contains $n_j > 1$ students, and we enumerate these students $i = 1, \ldots, n_j$. In the myON application, we take treatment assignment as occurring at the school level (rather than the summer school site level) for reasons discussed below. For the j^{th} school that receives the treatment, we write $Z_j = 1$. If the school is assigned to control, and students are not given myON readers, we write $Z_j = 0$. For each student within each school, we typically have observed, pretreatment covariates, $\mathbf{x}_{ji}$, including variables measured at

the student level and variables measured at the school level. For example, in the myON data, $\mathbf{x}_{ji}$ contains a measure of student gender for each student i. It also includes the percentage of students in school j who are proficient in reading based on test scores; this proficiency measure takes the same value for all students in the same school. Each student i in school j is described by both observed covariates and possibly an unobserved covariate u_{ji}. Note that data of this form is often referred to as multilevel data, since we have data on both units and the clusters within which the units are nested. In the context of a CRT, we are able to assess balance on observed pretreatment characteristics, $\mathbf{x}_{ji}$, at the time of randomization. Moreover, due to the properties of randomization, we can assume balance on the unobserved covariate u_{ji} due to the design.

Next, we define causal effects using the potential outcomes framework (Neyman 1923; Rubin 1974). Prior to treatment, each student has two potential responses: (y_{Tji}, y_{Cji}), where y_{Tji} is observed for student i in school j under $Z_j = 1$, and y_{Cji} is observed for this student under $Z_j = 0$. Note that this notation is the same across the two possible designs that we consider. In the myON application, y_{Tji} is the reading test score that student i in school j would exhibit if her school were assigned to implement myON, and y_{Cji} is the test score she would exhibit if her school were not assigned to implement myON. Writing the potential outcomes this way allows for arbitrary patterns of interference among students in the same school but not across schools. The outcomes that we actually observe are a function of potential outcomes and cluster-level treatment assignment:

$$Y_{ji}^{obs} = Z_j y_{Tji} + (1 - Z_j) y_{Cji}.$$

With potential outcomes defined, we can define the causal estimand, which is the target counterfactual quantity of interest in the study. In a COS in an educational setting, one reasonable estimand is the following student-level contrast: $y_{Tji} - y_{Cji}$. In the context of myON, this is the change in test scores for student i caused by school-level assignment to the reading program. If we assume the existence of an appropriate superpopulation, it would be natural to focus on the

average causal effect of the form $E[y_{Tji} - y_{Cji}]$ or the average causal effect for the treated of the form $E[y_{Tji} - y_{Cji}|Z_j = 1]$. With either of these focal estimands, the expectation is taken with respect to the superpopulation. Of course, given the counterfactual nature of these quantities, they are estimable with data only under a set of assumptions.

2.4 Assumptions

The first key assumption in a COS is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1986). SUTVA is assumed to hold implicitly in the notation we outlined above. This is true for both Designs 1 and 2. Here, we elaborate on what SUTVA implies in a COS. SUTVA includes two components: (1) the treatment levels of Z_j (1 and 0) adequately represent all possible versions of the treatment, and (2) one student's outcomes are not affected by other students' exposures. Under the first component of SUTVA and in the context of the myON intervention, we must assume that while there may be some variation in the process that leads to students receiving exposure to the myON program, the variation in this process corresponds to the same potential outcomes.

The second component of SUTVA assumes that the treatment for one student doesn't spillover to any control student. A benefit of clustered treatment assignment (as opposed to assignment at the individual level) is that it increases the plausibility of the second component of SUTVA. In the context of a COS, just like in a CRT, spillover that would violate SUTVA would need to occur across treated and control schools. For example, SUTVA would be violated if a student in a treatment school gave her myON account information to a control school student who subsequently used the tool. Although possible, this seems unlikely on a large scale. In general, judging the plausibility of the no-spillover assumption requires that the investigator gather qualitative information about the intervention's implementation. Throughout, we assume that SUTVA holds.

One might wonder whether SUTVA violations would be a concern under Design 2. That is, since only a fraction of the students within a school are treated, there might be spillovers from treated

to untreated students in the treated school. Although such spillovers are possible, and even likely in some cases, a SUTVA violation is still not a concern. Why? The causal effect of interest is between treated and control schools. For that causal effect, the spillover that is relevant is between treated and controls schools even if only a subset of students are treated within a school. When Design 2 is the target trial, one might also be interested in how the treatment spills over within a school, but that is a different causal question. The analysis of treatment effects under interference is the focus of much recent methodological work. See Aronow et al. (2017) and Basse and Feller (2018) for examples. In our context, we may be interested in whether students not assigned to the myON intervention improve their reading skills. Here, that is unlikely, since untreated students in treated schools are not attending summer school. However, even if such spillover did occur, it is irrelevant to whether SUTVA is reasonable to assume.

The next key assumption focuses on the process of treatment assignment. In a CRT, because of random assignment, the treatment assignment probabilities do not depend on potential outcomes, baseline covariates, or unobservables. We can write this assumption formally as:

$$\pi_j = Pr(Z_j = 1|\mathbf{y}_{Tji}, \mathbf{y}_{Cji}, \mathbf{x}_{ji}, \mathbf{u}_{ji}) = Pr(Z_j = 1).$$

Here, $\mathbf{y}_{Tji}$, $\mathbf{y}_{Cji}$, $\mathbf{x}_{ji}$, and $\mathbf{u}_{ji}$ represent vector versions of these terms defined above. In a CRT, we can assert that this assumption holds by design due to the properties of randomization. Critically, randomization ensures that, in expectation, treatment assignment does not depend on unobservable quantities such as $\mathbf{u}_{ji}$. Under this assumption, the analyst can estimate the average causal effect of offering the myON tool (the intent-to-treat effect) using a straightforward difference-in-means estimator.

In a COS, we are unable to invoke this assumption. Instead, we must apply a set of assumptions that we describe as "selection on observables." There are two parts to this assumption. First, under the selection on observables assumption, the analyst asserts that there is some set of covariates such that treatment assignment is random conditional on these covariates (Barnow

et al. 1980). Formally:

$$\pi_j = Pr(Z_j = 1|\mathbf{y}_{Tji}, \mathbf{y}_{Cji}, \mathbf{x}_{ji}, \mathbf{u}_{ji}) = Pr(Z_j = 1|\mathbf{x}_{ji}).$$

Said another way, after conditioning on observed characteristics, $\mathbf{x}_{ji}$, a given school's probability of assignment to the treatment is related neither to the potential outcomes of its students $(\mathbf{y}_{Tji}, \mathbf{y}_{Cji})$ nor to unobservables $(\mathbf{u}_{ji})$. That is, we assume there are no unobservable differences between the treated and control groups. Different fields refer to this assumption with different nomenclature, including "conditional ignorability" and "no omitted variables." This assumption requires investigators to ask themselves: How could it be that two schools that are identical on all meaningful background characteristics nonetheless receive different treatments? Critically, the selection on observables assumption is nonrefutable in that it cannot be verified with observed data (Manski 2007). Therefore, we advocate that sensitivity analysis should be a component of any COS. A sensitivity analysis allows the analyst to consider the sensitivity of results to the possibility of an unobserved confounder. We discuss sensitivity analysis in Section 3.5, below. Although this assumption often may appear implausible, there are many examples where treatment assignment has been shown to only depend on observed data (Dehejia and Wahba 1999; Wong et al. 2017; Keele et al. 2019; Fralick et al. 2018; Hernán and Robins 2016). Further, as Stuart (2010) points out, conditioning on observables also helps to account for unobserved variables to the extent that observed and unobserved measures are correlated.

The second part of the selection on observables assumption is often referred to as the assumption of common support. Formally, we assume that all clusters (e.g., schools) have some probability of being either treated or untreated such that $0 < \pi_j < 1$. That is, there are no clusters for which assignment to treatment is either guaranteed or prohibited, such that all clusters have some positive probability of receiving or not receiving treatment. In practice, large pre-treatment covariate imbalances between treated and untreated clusters and/or units is a telltale signal of problems with common support. Such imbalances often arise due to treated units that are very

dissimilar from any control units. When this occurs, it may be necessary for the investigator to trim (e.g., remove some observations from the analytic sample) the data, either at the student or school level, to enforce common support and improve balance.

Trimming treated units is not without consequence, however, as it changes the causal estimand. Once units are trimmed, the causal estimand describes the causal effect for the population of units for which the effect of treatment is marginal: units that may or may not receive the treatment. Changing the estimand in this way may be unproblematic if the data do not represent a well-defined population (Rosenbaum 2012). Under these assumptions, we use one or more statistical adjustment methods, such as regression, matching, or weighting, to estimate treatment effects. We discuss such adjustment methods in Section 3.1, below.

2.5 Explicating the Assignment Process

The role of the treatment assignment mechanism is a point of emphasis in the modern literature on observational studies (Rubin 2008). We agree that the assignment mechanism is critical to the design of COSs. That is, since the key assumption in an observational study is about whether treatment assignment is based on observed data, understanding how treatment assignment operates is key. Next, we review important aspects of clustered treatment assignment, including the fact that clustering of treatment assignment is generally advantageous compared to selection at the individual level in observational studies.

Critically, investigators should clearly understand and explicate the treatment assignment process. In the process of treatment assignment explication, we recommend the following steps. First, investigators should understand whether their application can be described as a "natural experiment." Non-experimental, but haphazard or arbitrary treatment assignments are often characterized as natural experiments—the hope being that natural circumstances give rise to a setting resembling as-if randomized treatment assignment (Murnane and Willett 2010). Although haphazard treatment assignment often requires considerable judgment and contextual knowledge to justify, the goal is to at least partially reduce the bias associated with self-selection

of treatment. For many natural experiments, analysts often still rely on covariate adjustment. When such covariate controls are introduced, the analyst is still relying, at least in part, on the selection-on-observables assumption needed in any observational study. In this way, observational studies and natural experiments are related. In fact, all of the principles we outline for COSs apply to natural experiments.

If a study cannot be described as a natural experiment, the investigator should identify both the decision-makers responsible for treatment allocation and any factors used in determining treatment assignment (Rubin 2008). In particular, analysts should try to identify whether the assignment mechanism is one where a set of decision-makers controlled the treatment allocation for others within some defined population. In education applications with grouped treatments, this is common and preferable. Why is this preferable to self-selected treatments? The advantage of assignment by an outside decision-maker is that the treatment selection process is more likely to be made based on observed information. While self-selected treatment assignment may reflect observed factors, it is also more likely to be driven, at least in part, by factors unobserved by the analyst, such as a child's motivation or a family's expectation regarding the benefits of the treatment.

For COSs in educational settings, it is more likely that treatment assignment will be controlled by outside decision-makers who are not directly exposed to the treatment. For example, district officials and principals will often be the people deciding to expose teachers or students to treatments. This selection structure offers a key advantage. In any observational study, investigators should carefully describe how treatments were assigned and outline the factors used to determine treatment allocation. Qualitative information typically is critical in this process. In a COS, it should generally be possible to identify whether district officials or principals participated in the assignment of treatment and to discover what factors they used in that decision. For example, Wake County centrally allocated myON to selected summer school sites based on a mix of factors including internet bandwidth, computer access, and regional distribution.[2] Thus, all summer

[2]District personnel responsible for implementing the Title I program, which included myON, provided the

school students who attended an elementary school close to a myON summer school site used the myON program during summer school. Principals and teachers had no input into program allocation. As such, it is more likely that the treatment assignment process was a function of the school-level data available to district administrators, rather than, for example, principals' assessments of teachers' appetite for or interest in using the tool. Thus, the selection-on-observable assumption may be reasonable in this context.

Next, analysts need to understand the assignment process to select between Designs 1 and 2. For example, in our application, beyond the selection of schools for myON, a secondary, student-level selection process occurred, whereby students were identified for summer school based on their standardized test performance, as mandated by state policy. Thus, we need to also understand this second assignment mechanism. Because student-level selection was governed by district rules, student populations should not differ systematically across treatment and comparison schools. Therefore, while there may be school-level differences related to selection for myON, it would be reasonable to expect that the set of summer-school eligible students looks similar across schools or at least that the variation in student-level characteristics is not systematically related to school-level selection into myON. Taken together, we should expect imbalances in school-level but not necessarily student-level covariates when we compare baseline characteristics between treatment and comparison schools. As we illustrate in the case below, our data follow this pattern.

Next, recent work in statistics has demonstrated that treatment assignment at the group level is advantageous in observational studies (Hansen et al. 2014). In a COS, group-level treatment assignment can reduce the effects of selection bias compared to individual-level treatment selection. For technical details, we refer readers to Hansen et al. (2014), but, here, we convey the intuition using the myON application. The myON resource is a commercial product, and one might imagine a salesperson motivated to bias evidence in favor of the product. The most effective way to do so would be to select individual students into the treatment group—specifically, to

research team with documentation related to myON's site-selection process and launch during the two-year period following implementation.

form a treatment group of high-performing students who will exhibit strong reading performance regardless of whether they used myON or not. However, if the salesperson is forced to select entire schools for the intervention, the mix of students within schools will make it more difficult to guarantee more favorable outcomes under myON. By selecting classrooms or entire schools, the salesperson is less able to target high performers who would bias results in favor of myON. Therefore, selecting groups for treatment is one way to limit bias from purposeful treatment selection. Of course, the limitation is that the analyst will never fully know how much bias is eliminated. Next, we turn to a discussion of statistical analysis.

3 Statistical Analysis

Whatever the advantages of clustered observational studies (compared to observational studies with individual-level selection), they remain observational studies. As such, investigators generally will find that treated and control groups differ on baseline covariates and therefore will need to increase comparability using a method of statistical adjustment to remove overt bias. Here, we highlight conventional and more modern approaches to statistical adjustment in the context of a COS.

3.1 Statistical Adjustment Methods

3.1.1 Regression

In education, random-effects regression models are frequently used to analyze data from cluster randomized trials. These models handle the multilevel structure of the data by introducing error terms at both the unit and cluster levels (Murnane and Willett 2010). In the event that we were assessing the myON tool in the context of a clustered randomized trial, the basic structure of the random-effects regression model would be as follows:

$$Y_{ji} = \gamma_0 + \gamma_1 Z_j + \upsilon_j + \epsilon_{ji},$$

where the model includes separate error terms at the school (υ_j) and student (ϵ_{ji}) levels. The simplicity of this model is due to the fact that randomization ensures that school-level treatment assignment is uncorrelated with the school- and student-level error terms. Of course, we can make this model more complex by including additional baseline (e.g., pre-randomization) covariates at the unit level, the cluster level, or both. In the context of a clustered randomized trial, the primary purpose of adding such covariates is to improve the precision of our treatment effect estimate by explaining residual variation attributable to these baseline characteristics.

The same type of regression model is also used for statistical adjustment in cluster observational studies. In the context of a COS, where the analyst relies on the selection on observables assumption, covariates are added to the model to remove overt biases—the observable differences between the treated and untreated clusters. For a COS, the analyst may tend to prefer more complex rather than parsimonious model specifications in order to reduce the potential for bias from omitted variables.

A general limitation of relying on regression-based strategies alone for analyzing data from a COS is that they can elide over the fact that there is little actual overlap in the distribution of covariates in the treatment and comparison schools. Areas outside of common support can be particularly problematic, since they require extrapolation, which can generate considerable model dependence. That is, the study conclusions will depend on the functional form of the regression model. Indeed, the farther the extrapolation is from the data, the larger the model dependence can become. In short, if we must depend on extrapolation, our inferences depend on our model and not the data, since the relevant empirical observations from the data do not exist.

This is not to say that regression-based analysis is not a useful tool for conducting COSs. Rather than turning directly to covariate-controlled regressions for assessing treatment effects, we advocate first taking steps to ensure balance and common support between treatment and control groups. Then, having obtained an analytic sample where balance and common support hold, regression-based tools can be used for treatment effect estimation. We discuss this possibility in

more detail below.

3.1.2 Propensity Score Adjustment

One alternative to regression modeling is the use of propensity score methods. Here, the analyst first models selection into treatment (the propensity score) as a function of observable data. Then, the estimated propensity score is used in the analysis through matching or weighting (Rosenbaum and Rubin 1983). The fundamental idea is that after adjusting for the propensity score, actual selection into treatment is as good as random. Several papers (e.g., Lalonde (1986); Diaz and Handa (2006)) have explored the conditions under which propensity score methods are likely to reproduce experimentally-derived causal effects in the case of unit-level selection.

In a COS, the statistical adjustment strategy needs to account for the multilevel structure of the data. Under propensity score approaches, this is done by estimating the propensity score using, for example, a random effects logistic regression model (Hong and Raudenbush 2006; Arpino and Mealli 2011; Li et al. 2013). However, these models suffer from drawbacks, such as lack of model convergence, which occurs when the estimating model fails to reach the best-fitting parameter estimates. For example, Zubizarreta and Keele (2016) find that multilevel models often fail to converge when used to estimate propensity scores. Therefore, although propensity score methods are a reasonable alternative to regression methods when the focal treatment is allocated at the individual level, in practice, the same is not always true in the context of a COS. This is because issues of model convergence often hamper fitting propensity score models with hierarchical education data. When this happens, little can be done.

3.1.3 Matching

Matching serves as another method of adjustment designed to mimic a randomized trial by constructing a set of treated and control units that are highly comparable on observed, pretreatment characteristics. While some matching methods rely on estimated propensity scores, many matching algorithms match on covariates directly and, in this way, eliminate the need for propensity scores. Matching methods primarily have been developed to handle individual level treatment as-

signment, and a large body of research has articulated best practice in this context (Rosenbaum 2020). Matching studies also have been used to evaluate a host of socially-relevant interventions (Stuart 2010), and methodological research has investigated the extent to which matching studies yield impact estimates similar to those achieved through experimental design (Dehejia and Wahba 1999; Cook et al. 2008).

Just as we can use individual-level matching to mimic an individual-level randomized trial, we can conceive of matching to mimic a CRT by creating comparable treatment and comparison clusters. Despite COSs being a natural analogue to the analytic workhorse of CRTs in the context of educational research, strategies for matching with grouped treatments are much less well developed. Extant work has focused on multilevel data structures, but mostly focused on applications where clusters are relevant in some way but not for grouped treatments. For example, Steiner et al. (2013) consider matching with multilevel data structures but assume that treatment assignment occurs at the student rather than the school level. Stuart and Rubin (2008) also focus on matching with multilevel data. They advocate building a comparison group from multiple sources when a single comparison site is not a sufficient match for a given treated group (Stuart and Rubin 2008). This approach considers matching only on student-level characteristics and disregards cluster-level measures—which renders it inapplicable to a COS where school-level covariates should be critical.

More recently, authors (2016a, 2016b) have developed matching methods designed specifically for a COS. The resulting matching method mimics a CRT by creating comparable treated and control clusters and units within clusters to remove overt bias at both the individual and group levels. Furthermore, these methods rely on matching procedures that produce an "optimal" match solution, meaning that the matching algorithm selects a mapping to minimize the sum of the distances between treatment and control observations (Rosenbaum 1989).

Although other methods of statistical adjustment are available, in the context of COSs specifically, we endorse matching methods for several reasons. First, matching tends to be more robust to a

variety of data configurations—especially when treated and control covariate distributions do not have good overlap (Imbens 2015). Second, matching methods allow for covariate prioritization through which the analyst can choose to increase treated-control comparability on covariates deemed to be of critical importance from a scientific standpoint. For example, an investigator can use matching to balance baseline test scores more closely relative to other covariates such as school size. Third, the investigator can trim the sample to yield the set of units with the highest levels of comparability. While our primary goal in this paper is to consider the design of COSs, we refer the interested reader to authors (2019) for a nontechnical overview and evaluation of these matching methods that are specifically designed for a COS. See authors (2015, 2016a, 2016b) for more technical treatments of the topic. In the case study below, we illustrate how we apply multilevel matching to the myON application.

While random effects regression models alone are not our preferred method for the analysis of COSs, they can be fruitfully combined with matching. Once matching is complete, the analyst can regress the outcome on the treatment indicator using a random-intercepts model with the matched data. Using a regression model is also useful in that post-matching covariate adjustment with regression can account for imbalances that may remain after matching. That is, any covariates that are not fully balanced can also be included in the post-match regression model to further reduce bias (Imbens 2015). As such, regression models are a useful tool for analysis once matching is complete.

3.2 Overlap

As we noted above, one of the key assumptions for a COS is common support or overlap of baseline covariate distributions. When there is little overlap between treated and control covariate distributions, trimming units via matching is one method to enforce overlap. However, analysts should take care when trimming in a COS. After trimming, the causal estimand is considered to be more local, since it applies to only a subset of the treated units. In a COS, trimming even a small number of treated schools may mean that a large percentage of treated units are lost. In

other words, trimming even a small number of clusters may make the treatment effect estimate very local. When this happens, there is not a good solution, since we do not want to estimate treatment effects using treated and control observations that are not comparable.

3.3 Correcting Standard Errors for Clustered Structure of Data

Another important principle from CRTs that also applies to COSs is that the analyst needs to correct estimates of statistical uncertainty to account for the clustering of students. Failure to do so will result in standard error estimates that are, at times, grossly underestimated given that the correlation among students in the same cluster has not been properly accounted for in the estimates of statistical uncertainty (Hayes and Moulton 2009; Angrist and Pischke 2009). In general, the investigator should account for clustering at the level at which the treatment has been assigned (Abadie et al. 2017). Thus when treatments are assigned to schools, schools become the relevant cluster in terms of standard error estimates. Typically, adjustments for the standard errors are done in one of two ways both based on regression models. One correction is to use robust variance estimators that take into account intra-cluster correlations. This method uses standard errors based on a generalization of clustered standard errors developed by Liang and Zeger (1986). Alternatively, random-effects regression models also correct for within-cluster correlations.

With matching methods, statistical adjustment is separate from the estimation of treatment effects, since outcome information is not used in the match. Therefore, adjusting for clustering occurs after the match is completed. When matching methods are used, there are two ways to account for clustering when estimating the treatment effects. First, if one uses regression models with the matched data to estimate treatment effects, then regression-based corrections can be used to account for clustering. Spiess and Abadie (2019) find that if matching is done without replacement, using post-matching regression models that account for pair clustering produces valid standard errors. As such, after creating matched pairs, the analyst should also include in the outcome model a random effect for paired school clusters. This approach accounts for

both clustering within schools and within matched school pairs. The only difficulty is that this method assumes that there is a sufficiently large number of clusters for valid inferences. The second approach for adjusting for clustering is based on randomization inference, which we review next.

3.4 Randomization Inference

To account for clustering while avoiding the large sample assumptions on which regression relies, one can alternatively use randomization inference methods. Hansen et al. (2014) outline randomization inference methods for COSs that are valid when units within clusters have correlated outcomes. Randomization inference methods are often referred to as permutation methods, since inferences are based on permuting the data consistent with the implied randomization. For example, within matched pairs, the analyst randomly reassigns treatment status and then estimates a treatment effect. Doing this repeatedly allows the generation of a null distribution of treatment effects against which to evaluate the treatment effect estimated based on actual assignment. The resulting inferences are valid for any sample size. Randomization inference also allows the investigator to use rank-based statistics which are robust in the presence of heavy-tailed distributions. However, randomization inference methods test the sharp null hypothesis which asserts that the treatment effect is zero for all schools and students. This is rather different from the more usual null hypothesis which asserts that the average effect is zero. In general, when sample sizes are small (e.g., 20 to 30 total clusters), it is useful to use randomization inference methods so one can understand whether inferences are dependent on the assumption of large sample sizes. For example, the analyst could conduct hypothesis testing through both regression and randomization inference approaches and examine whether conclusions are robust to method. When p-values derived from regression-based methods are far from standard thresholds, using both methods is likely unnecessary, even with a modest number of clusters. Nevertheless, as we discuss next, randomization inference is extremely useful for conducting sensitivity analyses.

3.5 Sensitivity Analysis

We advocate that all observational studies include a sensitivity analysis. Many sensitivity analyses are based on a partial identification strategy, where bounds are placed on quantities of interest while a key assumption is relaxed. A sensitivity analysis is designed to *quantify* the degree to which a key identifying assumption must be violated in order for a researcher's original conclusion to be reversed. If a causal inference is sensitive, a slight violation of the assumption may lead to substantively different conclusions. The first sensitivity analysis explored whether it was possible for an unobserved confounder to explain the variation in lung cancer rates that remained after accounting for the association with smoking (Cornfield et al. 1959).

Although a sensitivity analysis may focus on any assumption, most sensitivity analyses focus on the selection on observables assumption (Imbens 2003; Rosenbaum 1987). In the context of a COS, the key assumption is that there are no unobserved confounders. Here, we briefly outline a form of sensitivity analysis based on randomization inference that probes the assumption of no unobserved confounders and is designed to be compatible with a matched study. Other related forms of sensitivity analysis tend to use regression models (Altonji et al. 2005).

Rosenbaum (2002, ch. 4) developed a method of bounds to understand whether the selection on observables identification assumption is sensitive to the presence of a hidden confounder. To begin, recall that under selection on observables, we assume that any two clusters that we have matched have the same underlying probability of exposure to treatment. This means that when we use randomization inference methods, the flip of the coin is fair within this pair. Of course, selection on observables is a strong assumption, and it remains possible that our matched treatment and control clusters still differ on an unobserved confounder, u_{ji}, that drives selection into treatment. Sensitivity analyses allow us to quantify how strong an influence such an unobserved confounder would need to have on treatment selection to alter our substantive conclusions.

For example, the analyst might hypothesize that, despite matching, there remains an unobserved

covariate such that selection probabilities are unequal. If that hypothesized inequality (which Rosenbaum denotes with the parameter Γ) were by a factor of two, then in our randomization inference, we would permute treatment assignment with probabilities $\frac{1}{3}$ and $\frac{2}{3}$ within each matched pair. By first considering treated clusters to be twice as likely to receive treatment and then half as likely to receive treatment, one can calculate bounds on quantities such as the treatment effect point estimate or associated p-value based on a conjectured level of confounding.

Consider bounds on the treatment effect estimate, for example. If zero were included in those bounds, it means that a failure of the key assumption for that level of confounding would reverse the study conclusions. More generally, one can vary the Γ parameter to ask what level of confounding would reverse study conclusions. We can observe at what value of Γ the upper bound on something like a p-value exceeds the conventional 0.05 threshold for each test. One way to summarize the sensitivity analysis is to determine the Γ changepoint—the value of Γ at which the estimate is no longer statistically significant at the 0.05 level. If this Γ value is relatively large, we can be confident that our inferences are insensitive to hidden bias from non-random treatment assignment related to unobserved characteristics. However, if the Γ value is small, that suggests that our inferences are vulnerable to hidden confounders.

Although our discussion here focused on the level of Γ that would negate a significant treatment effect, note that this flexible procedure also can be used to consider the level of confounding that would need to be present to mask a treatment effect from being detected. In fact, this is how we apply this tool for sensitivity analysis in the case study that follows.

4 Case Study

Next, we conduct an analysis of the myON data to demonstrate the concepts discussed above. Our data contain 3,434 summer school students from 49 schools. When introducing this application, we noted that these 49 elementary schools were grouped into 19 different summer school sites, eight of which were selected to receive myON. Therefore, one analytic decision relates to whether we define clusters as elementary schools or summer school sites.

For several reasons, we opt to treat intact elementary schools as our clusters. First, we can reasonably infer that while summer school sites were selected for myON use, this process explicitly assigns schools to treatment or control. Second, defining clusters at the elementary school level leads to a larger number of clusters for our analysis, and this helps to improve statistical power. Finally, as we advocated above, we take a matching approach to statistical adjustment. For a matching estimator, we benefit from having a greater number of treatment and comparison clusters, as this helps to increase the likelihood of obtaining good cluster-level matches. Thus, our treatment-comparison contrast is between assigning groups of students to summer school sites that use the myON software versus those that do not, under the assumption that schools were otherwise comparable. In this application, students from 20 schools (containing a total of 1,371 summer school students) used the myON reading program.

Our next step is to consider which target trial is appropriate. Based on qualitative information about the intervention, we know that while entire schools were selected for treatment, within each treated school, the intervention applied only to the subset of students required to attend summer school. While control schools were not selected for using the program, the process for identifying students for summer school was identical across all schools in the district. In theory, then, summer school students within treated and control schools should be similar. Nevertheless, the student-level selection process points to Design 2 as the more relevant target trial. Therefore, we will need to assess empirically the extent to which treated and control students are observationally similar. For example, some schools may serve a set of students who are relatively high performing, among those required to participate in summer school, whereas another school may serve summer school participants who are further down in the distribution of achievement.

Given this, we investigate balance at both the school and student levels. Table 1 contains means for the treated and control groups as well as the standardized difference before applying any statistical adjustments.[3] In Table 1 we first note that the imbalances for all the student-level

[3]The standardized difference for a given variable here is computed by taking the mean difference between treatment and comparison schools or students and dividing by the pooled standard deviation (Silber et al. 2001; Rosenbaum and Rubin 1985; Cochran and Rubin 1973). A standardized difference of less than one-tenth of a

covariates, including pre-treatment test scores, are small. We view this as indicative of the fact that the summer school selection process is uniform across treated and control schools. This is not surprising given that the summer school criteria are constant across the district.

Table 1 also contains balance statistics for school-level covariates. All of these measures were calculated by the school district and thus are based on all enrollees from the previous school year—not just the students who attended summer school. For school-level covariates, there are clear differences between treated and control schools even though students selected into summer school are more similar. We see that treated schools, on average, have higher test scores, lower staff turnover, and a lower percentage of teachers who are nonwhite. Treated schools also have a higher share of teachers who are novices (i.e., three or fewer years of experience).

We should note that when comparing the student- and school-level covariates in Table 1, mean differences of a similar magnitude translate to very different standardized mean differences at the student level and the school level. This is, in part, a function of the fact that the standard deviations used to scale mean differences are larger at the student level than at the school level. This is another way of saying that there is more variation within than across schools. In Table 1, given that, for example, there is little school-level variation in the share of English language learners students served by each school, the mean difference of just 2 percentage points translates to a standardized mean difference of -0.29. This is also a function of the fact that selection for the intervention is truly occurring at the school level, whereas the student-level selection is similarly defined across all school sites.

Next, we use matching to address these baseline imbalances through statistical adjustment. The matching process is typically iterative; the analyst performs a match, assesses the resulting balance in baseline measures, and then fine-tunes the matching procedure to further improve balance until it is deemed acceptable. Just as outcome measures are not available at the time

standard deviation is often considered an acceptable discrepancy, since we might expect discrepancies of this size from a randomized experiment (Silber et al. 2001; Rosenbaum and Rubin 1985; Cochran and Rubin 1973; Rosenbaum 2010).

Table 1: Balance on student and school level covariates before matching.

Student Covariates	Treated Mean Before	Control Mean Before	Std. Difference
Reading pretest score	437.00	437.90	-0.02
Math pretest score	60.25	60.56	-0.02
Male (0/1)	0.36	0.40	-0.09
Special education (0/1)	0.47	0.43	0.09
Hispanic (0/1)	0.53	0.52	0.02
African-American (0/1)	0.22	0.22	0.00
School Covariates			
Composite proficiency	60.74	58.56	0.21
Proficient in reading	58.48	57.27	0.11
Proficient in math	60.68	58.41	0.20
Free/reduced lunch eligible	0.50	0.51	-0.10
English language learners	0.13	0.15	-0.29
Novice teachers	0.19	0.17	0.28
Staff turnover	0.11	0.12	-0.28
Nonwhite teachers	0.14	0.18	-0.26
Title 1 school	0.90	0.93	-0.11
Title 1 focus school	0.25	0.24	0.02
Schools	20	29	
Summer school students	1,371	2,063	

$Note$: Standardized difference for a given variable is computed as the mean difference between treatment and comparison schools or students and dividing by the pooled standard deviation.

of randomization in an experimental study, at no time should the analyst examine outcomes when implementing matching procedures. The CRT analogue to this process is conducting a randomization, assessing balance on baseline measures, and re-randomizing if baseline equivalence on observable characteristics is not satisfied (Morgan et al. 2012).

Instead of presenting results only from the final match that we judged to have the best balance, we present a series of matches to illustrate the iterative nature of the matching process. In doing so, we demonstrate additional tools that we use to improve balance. These tools are balance prioritization, calipers, and subsetting. For a more detailed descriptions of these tools, we refer the interested reader to authors (2019) , which as previously noted serves as a more technical companion to this paper. To perform these matches, we use the `R` package `matchMulti` which

members of our team have built specifically for matching with COS designs (authors).

The first match we present is based on the match algorithm defaults. At the defaults, no covariate is given priority in terms of balance, and no treated schools are dropped from the match. The resulting sample size includes 40 schools, with 20 treatment schools each pair-matched to a control school without replacement. The second column of Table 2 contains the results from this match. Here, some but not all of the standardized differences improve. In Match 2, we add covariate prioritization. With covariate prioritization, the investigator is able to select sets of covariates for the match to prioritize in terms of balance. Such prioritization is useful, because science and contextual understanding may motivate the investigator to prefer closer balance on some covariates over others. We add covariate prioritization by defining two covariate sets. The first set includes the three school-level test score measures. The second covariate set includes the proportion of English language learners and the proportion of nonwhite teachers. Under balance prioritization, the matching algorithm first seeks to balance the covariates in set 1, and then seeks to balance the covariates in set 2. The remaining covariates are then given the lowest priority for balance. In Match 2, balance on the test score measures is much improved, as we would expect. However, the improvements in the other two covariates we set for prioritization are minimal. In Match 2, we again have a resulting sample of 20 treatment and 20 control schools.

Next, we applied a school-level caliper to the match. The `matchMulti` package includes a function that calculates a school-level propensity score, which is the estimated probability of being selected for treatment based on the vector of baseline measures. Then we impose a caliper on this estimated propensity score as another tool to improve balance. We set the caliper to 0.20, which forbids any school-level matches that differ by more than 0.20 of a standard deviation on the estimated propensity score. We also add a third covariate balance prioritization set which includes the proportion of novice teachers and the staff turnover rate. Match 3 in Table 2 contains the results from this match. Match 3 is generally better, although balance is worse with regard to the proportion of novice teachers. Critically, this match discarded some treated schools because

Table 2: Balance on school level covariates for four different sets of match parameters.

	Unmatched	Match 1 Default Settings	Match 2 Covariate Prioritization	Match 3 School Caliper	Match 4 Optimal Subsetting
Composite proficiency	0.21	0.27	0.12	-0.01	-0.06
Proficient in reading	0.11	0.18	0.04	0.08	-0.01
Proficient in math	0.20	0.28	0.13	-0.01	-0.06
Free/reduced lunch eligible	-0.10	-0.05	-0.03	-0.03	0.14
English language learners	-0.29	-0.14	-0.14	-0.02	0.13
Novice teachers	0.28	0.12	0.21	0.30	0.15
Staff turnover	-0.28	-0.16	-0.25	0.11	0.03
Nonwhite teachers	-0.26	-0.38	-0.30	-0.02	0.05
Title 1 school	-0.11	-0.18	0.00	0.00	0.00
Title 1 focus school	0.02	0.00	0.00	0.00	0.14
Schools	49	40	40	30	32
Summer school students	3,434	2,888	2,751	1,210	1,378

$Note$: Cell entries are standardized differences.

for these schools, the caliper constraint could not be satisfied. Once the use of a caliper discards schools, it is better to use optimal subsetting of the data. This is because it is possible, with optimal subsetting, to achieve a similarly good balance without losing as many treatment sites as might be lost with a caliper strategy for improving balance.

In the context of multilevel data, optimal subsetting can be used to trim clusters, units, or both. Given sample sizes, however, trimming is typically necessary only at the school level. In applying optimal subsetting, the analyst specifies a minimum number of treated clusters (or units) that must be included. By adjusting this number downward, the analyst can drop treated schools one-by-one until balance improves. For example, if there are 20 treated schools and the optimal subset number is set to 19, the algorithm will discard the one treated school with the poorest match among all of the treated schools. In general, we recommend dropping schools one-by-one until balance is deemed acceptable.

We improve balance on the proportion of novice teachers by dropping four treated schools via optimal subsetting and rematching. Specifically, this match (Match 4) excludes the four treated

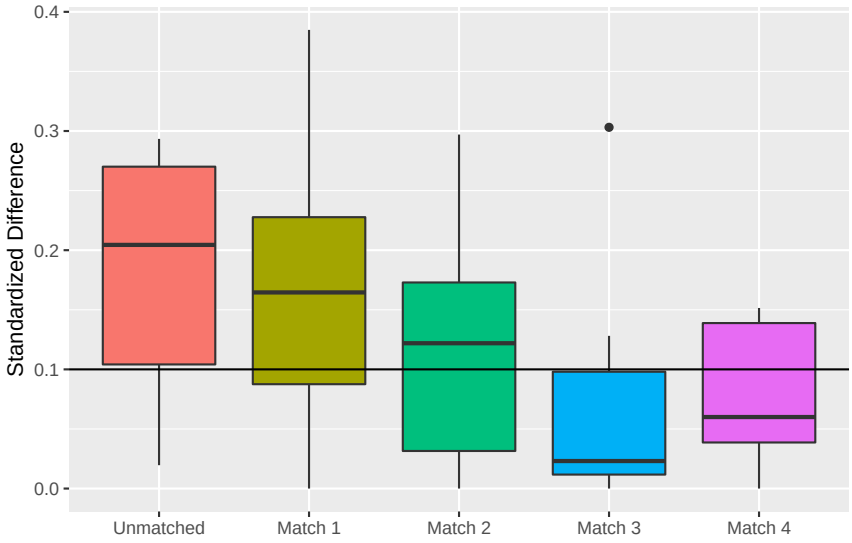

Figure 1: Boxplots of the distribution of absolute standardized differences for school level covariates.

schools with the largest covariate imbalances. This match changes the casual estimand in that subsequent treatment effect estimates will apply to a subset of treated schools—not the entire treated population. We might ask whether the estimand is *too* local, since we lost 20% of the treated school population. Is this still a population of interest? Unfortunately, that isn't a question that can be answered using statistics. In practice, we advise examining descriptive statistics for the treated population that remains in order to understand whether and how it differs from the full treated population.

Next, we plot the distribution of covariate standardized differences for each match in Figure 1. The boxplots reveal a few clear patterns. First, matching using the default settings does improve balance overall (Match 1), but a few covariates remain highly imbalanced. It is also clear that Match 3 is well-balanced with the exception of one covariate, as we saw in Table 2. This is relevant information. It tells us that the trimming removed schools with a larger proportion of novice teachers—and that the schools in Match 4 differ from the overall treated population mostly with respect to this covariate.

Finally, we briefly return to the question of balance for the student-level covariates after matching

in Table 3. As noted above, even in the unmatched data, the standardized differences between treatment and control students are quite modest, as none exceeds 0.10. In assessing the balance in student-level covariates for the first and final matches, we find that it remains roughly the same across the matches. Taken together, there is little evidence that treatment selection was a function of observed student-level covariates for students participating in summer school.

Table 3: Balance on student level covariates.

	Unmatched	Match 1	Match 4
Reading pretest score	-0.02	-0.03	0.01
Math pretest score	-0.02	-0.03	-0.01
Male (0/1)	-0.09	-0.05	-0.11
Special education (0/1)	0.09	0.06	0.11
Hispanic (0/1)	0.02	0.01	0.02
African-American (0/1)	-0.00	0.02	-0.00

4.1 Outcome Estimates

Next, we turn to assessing the effectiveness of the myON intervention to improve performance on the post-summer school standardized reading assessment in our matched sample. As outlined above, once matching is complete, we can estimate treatment effects in a number of different ways. One approach is to simply fit a regression model to the matched data with the outcome regressed on the treatment indicator. Here, we use multilevel models with a random intercept and clustering at both the school and matched school-pair level. One advantage to using regression models for treatment effect estimation is that the analyst can add baseline covariates to the model. In particular, it is useful to include any covariates that did not balance sufficiently in the match. For example, in Match 3, we were unable to reduce the standardized differences below 0.10 for school enrollment, the percentage of novice teachers, and the staff turnover rate. To more completely remove bias from the imbalance in these covariates, we can include them in the model used to estimate the treatment effect.

Table 4 reports unadjusted estimates as well as those produced from regression adjustment alone, matching alone (for Matches 3 and 4), and matching in combination with regression adjustment

(again with Matches 3 and 4). Two facts are clear from the results. First, there is little difference between the unadjusted estimate and any of the adjusted estimates. This suggests either that little self-selection is present in this application, or if selection bias is present, it is *not* a function of the observed data. Second, when selection biases are not a function of observed data, the effect of the adjustment methods will be minimal. This is also true here. Estimates based on regression alone, matching, or matching plus regression all produce similar estimates. Across all methods, effect sizes are small, and the associated 95% confidence intervals include zero—results that conflict with claims made by the creators of the myON software.[4] Finally, it is worth noting that the causal estimand for Match 4 differs from the other estimates. In Match 4, we dropped four treated schools, so the results from Match 4 do not apply to the entire treated population. However, that difference appears to be unimportant, given that substantive conclusions are similar between Match 4 and the other matches.

In this example, the estimates of the treatment effect do not vary with the type of statistical adjustment. Should we interpret this as evidence that these choices are inconsequential? Design choices—including the type of the match—should be done without reference to outcomes. Such choices may be of consequence in other applications. If we use regression alone, we still cannot be sure that our inferences are not overly dependent on the model to extrapolate between treatment and control sites that had poor overlap. The inferences we are able to derive and the confidence that we have in them have more to do with the strength of our design process and less to do with how the results change across the different strategies represented in Table 4. Our confidence can be increased further by using a sensitivity analysis.

4.2 Sensitivity Analysis

Next, we present the results from a sensitivity analysis. As noted above, we can use sensitivity analyses to identify whether it would take a weak or strong unobserved confounder to render

[4]For example, myON documentation suggests that students using the product can increase their Lexile scores by more than 20% Corp (2015). Ortlieb et al. (2014) find that while myON can potentially improve reading achievement when used in conjunction with traditional books, it has no positive impacts as a stand-alone product.

Table 4: Outcome estimates for the treatment effect of the myON reading program.

	Treatment Effect Estimate
Unadjusted	0.03
	[-0.08, 0.14]
Regression	0.05
	[-0.03, 0.13]
Match 3	0.04
	[-0.09, 0.16]
Match 3 + Regression	0.04
	[-0.12, 0.21]
Match 4	0.04
	[-0.07, 0.16]
Match 4 + Regression	0.07
	[-0.06, 0.21]

Note: Quantities in brackets are 95% confidence intervals. Outcomes are standardized test scores.

a significant treatment effect no longer significant. That is, we seek to understand whether the study results could be easily explained by bias from an unobserved confounder. In the myON analysis, however, the treatment effect estimates are small, and the confidence intervals include zero, so we are unable to reject the null hypothesis of no treatment effect. One might conclude that given the null results, there is no need to conduct a sensitivity analysis. Here, we illustrate how we can instead explore the possibility that bias from a hidden confounder *masks* an educationally meaningful effect. That is, an unobserved confounder may leave us to conclude that there is no effect when in fact such an effect exists. We can explore this possibility by using a test of equivalence with a sensitivity analysis (Rosenbaum 2008; Rosenbaum and Silber 2009; Rosenbaum 2010).

Because we did not discuss it above, we first review tests of equivalence. Under a test of equivalence, the null hypothesis asserts that the absolute value of the treatment effect is greater than some δ, a treatment effect size set by the researcher. That is, $H_{\neq}^{(\delta)} : |\tau| > \delta$ for some specified $\delta > 0$. Here, we set δ to 0.20 of a standard deviation, as 0.20 is generally considered to be a meaningful effect size in education research (Kraft 2019). Therefore, the relevant null

hypothesis for a test of equivalence is that the treatment effect, denoted τ, is greater than 0.20 or less than -0.20. Rejecting the null hypothesis provides a basis for asserting with 95% confidence that τ is between -0.20 and 0.20. That is, $|\tau| < \delta$. $H_{\neq}^{(\delta)}$ is the union of two exclusive hypotheses: $\overleftarrow{H}_0^{(\delta)} : \tau \leq -\delta$ and $\overrightarrow{H}_0^{(\delta)} : \tau \geq \delta$, and $H_{\neq}^{(\delta)}$ is rejected if both $\overleftarrow{H}_0^{(\delta)}$ and $\overrightarrow{H}_0^{(\delta)}$ are rejected (Rosenbaum and Silber 2009). We can apply the two tests without correction for multiple testing, since we test two mutually exclusive hypotheses. That is, we set $\alpha = 0.05$ for each test. Thus we can test whether the estimate from our study is different from other possible treatment effects which are represented by δ. With a test of equivalence, it is not possible to demonstrate a total absence of effect, but instead we test that our estimated effect is not as large as δ (in a positive or negative direction). Under a test of equivalence, the closer the estimated treatment effect is to zero, the smaller the p-values will be, since the estimated effects will be farther from δ.

Next, we implement a test of equivalence for the myON data for Match 4, first assuming no unobserved confounding. We test $\overleftarrow{H}_0^{(\delta)}$ and find that the one-sided p-value from this test is 0.011. We then test $\overrightarrow{H}_0^{(\delta)}$, and we find that the associated one-sided p-value is 0.027. The overall test of equivalence is then based on the larger of these two p-values. Therefore, we are able to reject the null that the estimated treatment effect we observe in this study is equivalent to an educationally significant effect.

If our study were a CRT, we could be confident that the results were not due to unobserved differences between the treated and control group. However, in a COS, we may reject the null hypothesis of equivalence due to hidden confounding. That is, the test results above are conducted under the assumption that there is no hidden bias (e.g., $\Gamma = 1$). However, using a sensitivity analysis we can explore whether and to what extent the test of equivalence is sensitive to potential biases from non-random treatment assignment. That is, we ask whether our inference of no educationally meaningful effect is sensitive to bias from a confounder.

To conduct the sensitivity analysis, we repeat the test of equivalence, but use values of Γ that

are larger than 1. When Γ is greater than 1, we obtain upper and lower bounds on the p-values derived above. We then find the Γ changepoint—the value of Γ at which the upper-bound on the p-value is no longer statistically significant at the 0.05 level. This is the amount of confounding that would need to be present for our test result to no longer be statistically significant. In the myON study, we find that when Γ is as small as 1.3, the upper bound on the p-value for the test of equivalence is 0.049. Γ is on an odds-ratio scale. This implies that if there were a binary unobserved confounder that caused the odds of treatment to differ by 30%, that could explain the result from the test of equivalence.

Is this a large or small value of Γ? To provide a benchmark, we regressed treatment status on the observed covariates using a logit model. We then calculated the odds-ratios for the covariates in this model. We can compare these odds-ratios from this model to those from the sensitivity analysis. If the odds-ratios from this model are smaller than the Γ value, then the hidden confounder would need to have an effect on the odds of treatment that is generally larger than the covariates we observe. We can interpret this as a robust result, since the effect of the unobserved confounder would need to be generally larger than the observed data. However, if the Γ value is smaller than these estimated odds-ratios, then the unobserved confounder could be similar to observed confounders. For example, if we increase composite school test scores by one-tenth of a standard deviation, that increases the odds of being treated by 1.42. Thus, a Γ value of 1.3 is smaller. This implies that an unobserved confounder could easily mask an educationally meaningful effect.

5 Discussion

Although randomized trials are considered the "gold standard" for conducting educational effectiveness research, they are not always possible for cost, political, or other reasons. Furthermore, an investigator may have questions about the efficacy of a given educational intervention only after it has been implemented in a non-random manner. In educational contexts, such non-random allocation of educational opportunities often occurs at the cluster (e.g., school or classroom) level

rather than at the individual level.

In such instances, thoughtfully designed cluster observational studies conducted in concert with sensitivity analyses are an important tool in the education analyst's arsenal. However, the key to conducting a high-quality COS is thoughtful design. Here, we outline a set of principles for the design of clustered observational studies. We advocate that COSs be designed with their cluster randomized trial analogue as a guide. Analysts should focus on the assignment mechanism and seek to identify the factors used for treatment allocation. We further advocate the use of multilevel matching strategies to achieve treatment and control balance and common support prior to the application of regression or other analytical tools for estimating treatment effects.

The weakness of a COS, of course, is that even after thoughtful application of such matching and regression-based strategies, the analyst can never definitively know whether a critical unobserved confounder is the true driver of an impact estimate or whether such an unobserved measure may be masking true effects that remain unobserved. Nevertheless, sensitivity analyses, such as those discussed above, allow the analyst to consider how large such confounders would need to be to operate in either of these ways, and whether a confounder of such a magnitude is reasonable within the context under consideration. In short, there is much to be learned from thoughtfully designed and implemented COSs. Our goal in this paper is to establish a framework and guidelines for such work.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017), "When Should You Adjust Standard Errors for Clustering?" Tech. rep., National Bureau of Economic Research.

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.

Angrist, J. D. and Pischke, J̈.-S. (2009), *Mostly Harmless Econometrics*, Princeton, NJ: Princeton University Press.

Aronow, P. M., Samii, C., et al. (2017), "Estimating average causal effects under general interference, with application to a social network experiment," *The Annals of Applied Statistics*, 11, 1912–1947.

Arpino, B. and Mealli, F. (2011), "The specification of the propensity score in multilevel observational studies," *Computational Statistics & Data Analysis*, 55, 1770–1780.

Barnow, B., Cain, G., and Goldberger, A. (1980), "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies*, eds. Stromsdorfer, E. and Farkas, G., San Francisco, CA: Sage, vol. 5, pp. 43–59.

Basse, G. and Feller, A. (2018), "Analyzing two-stage experiments in the presence of interference," *Journal of the American Statistical Association*, 113, 41–55.

Borman, G. D., Slavin, R. E., Cheung, A. C., Chamberlain, A. M., Madden, N. A., and Chambers, B. (2007), "Final reading outcomes of the national randomized field trial of Success for All," *American Educational Research Journal*, 44, 701–731.

Cochran, W. G. and Rubin, D. B. (1973), "Controlling Bias in Observational Studies," *Sankyha-Indian Journal of Statistics, Series A*, 35, 417–446.

Cook, T. D., Shadish, W., and Wong, V. C. (2008), "Three conditions under which observational

studies produce the same results as experiments," *Journal of Policy Analysis and Management*, 27, 724–750.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959), "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions," *Journal of National Cancer Institute*, 22, 173–203.

Corp, C. (2015), "myON: A Complete Digital Literacy Program," `http://thefutureinreading.myon.com/overview/complete-literacy-program`.

Dehejia, R. and Wahba, S. (1999), "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evalulation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.

Diaz, J. J. and Handa, S. (2006), "An assessment of propensity score matching as a nonexperimental impact estimator evidence from Mexico's PROGRESA program," *Journal of human resources*, 41, 319–345.

Donner, A. and Klar, N. (2004), "Pitfalls of and Controversies in Cluster Randomization Trials," *American Journal of Public Health*, 94, 416–422, pMID: 14998805.

Fralick, M., Kesselheim, A. S., Avorn, J., and Schneeweiss, S. (2018), "Use of health care databases to support supplemental indications of approved medications," *JAMA internal medicine*, 178, 55–63.

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.

Hayes, R. and Moulton, L. (2009), *Cluster Randomised Trials*, Chapman& Hall/CRC.

Hedges, L. V. and Hedberg, E. C. (2007), "Intraclass correlation values for planning group-randomized trials in education," *Educational Evaluation and Policy Analysis*, 29, 60–87.

Hernán, M. A. and Robins, J. M. (2016), "Using big data to emulate a target trial when a randomized trial is not available," *American journal of epidemiology*, 183, 758–764.

Hong, G. and Raudenbush, S. W. (2006), "Evaluating Kindergarten Retention Policy: A Case of Study of Causal Inference for Multilevel Data," *Journal of the American Statistical Association*, 101, 901–910.

Imbens, G. W. (2003), "Sensitivity to Exogeneity Assumptions in Program Evaluation," *The American Economic Review Papers and Proceedings*, 93, 126–132.

— (2015), "Matching methods in practice: Three examples," *Journal of Human Resources*, 50, 373–419.

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., Miratrix, L., and Page, L. (2019), "Matching Methods for Clustered Observational Studies in Education," Unpublished Manuscript.

Kraft, M. A. (2019), "Interpreting effect sizes of education interventions," Edworkingpaper: 19-10, Annenberg Institute at Brown University.

Lalonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604–620.

Li, F., Zaslavsky, A. M., and Landrum, M. B. (2013), "Propensity score weighting with multilevel data," *Statistics in medicine*, 32, 3373–3387.

Liang, K.-Y. and Zeger, S. L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 13–22.

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

Morgan, K. L., Rubin, D. B., et al. (2012), "Rerandomization to improve covariate balance in experiments," *The Annals of Statistics*, 40, 1263–1282.

Murnane, R. J. and Willett, J. B. (2010), *Methods matter: Improving causal inference in educational and social science research*, Oxford University Press.

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).

Ortlieb, E., Sargent, S., and Moreland, M. (2014), "Evaluating the efficacy of using a digital reading environment to improve reading comprehension within a reading clinic," *Reading Psychology*, 35, 397–421.

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015), "Large, Sparse Optimal Matching with Refined Covariate Balance in an Observational Study of the Health Outcomes Produced by New Surgeons," *Journal of the American Statistical Association*, 110, 515–527.

Pimentel, S. D., Page, L. C., Lenard, M., and Keele, L. J. (2017), "Optimal Multilevel Matching Using Network Flows: An Application to a Summer Reading Intervention," *Annals of Applied Statistics*, In press.

Raudenbush, S. W. (1997), "Statistical analysis and optimal design for cluster randomized trials." *Psychological Methods*, 2, 173.

Rosenbaum, P. R. (1987), "Sensitivity Analysis For Certain Permutation Inferences in Matched Observational Studies," *Biometrika*, 74, 13–26.

— (1989), "Optimal Matching for Observational Studies," *Journal of the American Statistical Association*, 84, 1024–1032.

— (2002), *Observational Studies*, New York, NY: Springer, 2nd ed.

— (2008), "Testing hypotheses in order," *Biometrika*, 95, 248–252.

— (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.

— (2020), "Modern Algorithms for Matching in Observational Studies," *Annual Review of Statistics and Its Application*, 7.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The Central Role of Propensity Scores in Observational Studies for Causal Effects," *Biometrika*, 76, 41–55.

— (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods," *The American Statistician*, 39, 33–38.

Rosenbaum, P. R. and Silber, J. H. (2009), "Sensitivity Analysis for Equivalence and Difference in an Observational Study of Neonatal Intensive Care Units," *Journal of the American Statistical Association*, 104, 501–511.

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.

— (2007), "The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials," *Statistics in medicine*, 26, 20–36.

— (2008), "For Objective Causal Inference, Design Trumps Analysis," *The Annals of Applied Statistics*, 2, 808–840.

Silber, J. H., Rosenbaum, P. R., Trudeau, M. E., Even-Shoshan, O., Chen, W., Zhang, X., and Mosher, R. E. (2001), "Multivariate matching and bias reduction in the surgical outcomes study," *Medical Care*, 39, 1048–1064.

Spiess, J. and Abadie, A. (2019), "Robust Post-Matching Inference," Unpublished Manuscript.

Steiner, P., Kim, J.-S., and Thoemmes, F. (2013), "Matching strategies for observational multilevel data," in *JSM proceedings*, pp. 5020–5032.

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

Stuart, E. A. and Rubin, D. B. (2008), "Matching with multiple control groups with adjustment for group differences," *Journal of Educational and Behavioral Statistics*, 33, 279–306.

Torgerson, D. J. (2001), "Contamination in trials: is cluster randomisation the answer?" *BMJ: British Medical Journal*, 322, 355.

Wong, V. C., Valentine, J. C., and Miller-Bains, K. (2017), "Empirical performance of covariates in education observational studies," *Journal of Research on Educational Effectiveness*, 10, 207–236.

Zubizarreta, J. R. and Keele, L. (2016), "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.