AERA Open
July-September 2020, Vol. 6, No. 3, pp. 1–14
DOI: 10.1177/2332858420954401
Article reuse guidelines: sagepub.com/journals-permissions

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 practices 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 a large U.S. school district.

Keywords: *causal inference, hierarchical/multilevel data, observational study, optimal matching*

Introduction

The effectiveness of educational interventions often is evaluated using clustered randomized trials (CRTs) in which random assignment occurs at the group rather than individual level (Hedges & Hedberg, 2007; Raudenbush, 1997). Given the natural groupings of students within classrooms and schools as well as the practical and political challenges of 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 treated and untreated groups can be causally attributed to the intervention. However, randomized trials, even with group-level assignment, are not always feasible. 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 but through some uncontrolled process.

Although the literature on observational studies for deriving causal inferences when treatment selection occurs at the individual level is well developed (Rubin, 2007, 2008), the same is not true regarding COSs. For COSs, the literature remains underdeveloped, with no consensus on best practices. This is surprising in the context of educational research, where treatment selection often occurs at the group level. In this article, we outline the key considerations and steps for designing and conducting a COS. We highlight differences between COSs and observational studies with individual-level treatment selection, 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 standard principles must be altered to handle clustered treatment assignment.

Our framework employs the counterfactual model for causal inference. We begin by advocating that investigators design COSs following the principle of target trial emulation—that is, according to the cluster randomized trial that they ideally would have conducted. Although the concept of target trial emulation is not new, we highlight considerations unique to the COS context and associated hierarchical data.

Next, we discuss the importance of understanding the process through which sites were selected into treatment. 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 with non-random selection into treatment. We introduce notation, articulate assumptions necessary for causal inference in the context of a COS, and argue for the central role of analyses to understand the potential sensitivity of conclusions to an unobserved confounder. Next, we highlight possible approaches to statistical analysis. In this section, we discuss a new form of matching designed specifically for COSs. In

sum, our article aims 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 critical design aspects of COSs.

We illustrate the study design process and related concepts with an evaluation of myON, a summer reading intervention in Wake County, North Carolina. In particular, the COS design process requires gathering information on how the treatment was assigned and structuring the analysis to reflect this assignment process. We find no evidence that access to the myON tool improved student outcomes, but this does not diminish the value of the example. As with CRTs, when conducting a COS, all of the key elements of study design should occur prior to examining impacts. Furthermore, careful study design should lead policy makers to place more stock in results, even if estimated effects are not educationally meaningful.

Research Design Principles for Clustered Observational Studies

Here, we outline key considerations for designing COSs. We begin by discussing target trial emulation.

Target Trial Emulation

Target trial emulation calls for applying design principles from randomized trials to the analysis of observational data (Hernán & Robins, 2016). Under the target-trial approach, the investigator ties the design and analysis of the observational study to the experimental trial it emulates, 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 observational studies, investigators typically assume that any differences between treated and control groups are observable—that no unobserved differences exist—and that covariate adjustment can handle observable differences. We discuss identification assumptions and covariate adjustment further below.

The purpose of target trial emulation is to improve the quality of observational studies through the application of trial design principles. In an experimental study, the sample and study design are clearly delineated to enable randomization. In contrast, observational studies, particularly those formulated after program implementation, often necessitate investigation to inform decisions about and articulation of sample construction and study design. Imagining the hypothetical experiment that would generate observational data under study (Cochran & Rubin, 1973; Rubin, 2008) initially seems simple. However, this can be challenging in practice since we might conceive of several different hypothetical experiments that generate a given dataset. Here, we outline two CRT study designs common in educational interventions, corresponding to situations where (1) whole groups are assigned to a given treatment and (2) subsets of whole groups are assigned to a given treatment according to qualifying criteria.

Design 1: Clustered Treatment Assignment. Design 1 handles cases where complete clusters (e.g., whole classrooms, schools) are selected for treatment, and all units within a cluster either do or do not receive treatment. Under Design 1, we seek to mimic a CRT in which treatment assignment occurs at the cluster level, and all units within selected clusters receive (or are intended to receive) treatment. Under the COS analogue, cluster-level covariates are critical, given that the assignment occurs at this level and is presumed to have been made based on cluster-level characteristics alone. This design would be appropriate for assessing the impact of school-wide reform efforts, such as Success for All (Borman et al., 2007).

Design 2: Clustered Treatment Assignment for Student Subsets. CRTs often assume that the data include all units in each cluster or a random sub-sample of all units, such that the selected units are representative of the cluster as a whole (Donner & Klar, 2004; Torgerson, 2001). 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 nonrandom, student-level selection into the treatment within clusters; clusters are assigned to treatment, but within selected clusters only some units receive treatment. This might occur, for example, if an intervention targeted students who are struggling academically. In such cases, the causal estimand is a group-level contrast for the subset of students within their schools who are at risk for treatment.

The critical distinction from Design 1 is that under Design 2, final treatment assignment of an individual depends on school- and student-level characteristics. Selection of units for treatment within a cluster is analogous to nonrandom attrition. In a CRT, the investigator would need to correct for this selection bias. The same is true in a COS. That is, if the treatment is applied only to a subset of students within selected clusters, the analyst may need to model a second selection mechanism. This implies that in a COS analogue to Design 2, 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 aligns.

Motivating Application: A Summer School Reading Intervention

In summer 2013, the Wake County Public School System (hereafter, WCPSS) selected myON, a computer-aided instruction program for implementation at selected summer

school sites with the goal of boosting summer school attendees' reading comprehension. myON is a web-based software product that provides students with access to books and suggests titles based on their preferences and reading ability. Students at selected sites used the program for up to thirty minutes during the daily summer-school literacy block and could continue using it at home with a device and internet connection. At the time of its launch in WCPSS, the developers claimed that students using myON would improve comprehension through access to digital books that include "multimedia supports, real-time reporting and assessments and embedded close reading tools" (Capstone Digital, 2015). Given their prevalence and cost, rigorous, independent assessment of such curricular supplements is critical to sound investment decisions by educational agencies.

The study sample includes 3,434 summer school students from 49 different WCPSS elementary schools who attended summer school at one of 19 sites. Due to technical constraints, only some summer school sites used myON. As such, all students in a school were exposed to the software if they attended summer school at a selected site. In a COS designed to study the effects of myON, Design 2 is the relevant target trial, because myON was assigned to schools but only students required to attend summer school were exposed to the treatment. Therefore, we are interested in contrasting outcomes for groups of summer school students who were and were not exposed to myON.

Our key outcome is student-level reading performance, measured using Curriculum Associates' i-Ready Reading Assessments. Students sat for assessments in reading and mathematics at the beginning and end of summer school and results were reported in scale scores with a possible range of 0 to 800 (Curriculum Associates, 2015). Students also received a reading Lexile score used for selecting an initial bundle of digital books within myON (MetaMetrics, 2012).

Notation

Target trial emulation applies to study design, broadly, as well as analytic notation, specifically. Here, we introduce notation applicable to 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. Generally, for applications with students nested within schools, each school j contains $n_j > 1$ students. 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 summer school site level) for reasons discussed below. For the jth school receiving treatment, we write $Z_j = 1$, and if assigned to control, 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. In the myON data, for example, $\mathbf{x}_{ji}$ contains a measure of student i's gender. It also includes the percentage of students in school j who are proficient in reading; this proficiency measure takes the same value for all students in school j. Each student i in school j is described by both observed covariates and possibly an unobserved covariate u_{ji}. We refer to data of this form as multilevel data, since we have information on both units and the clusters within which the units are nested. In a CRT, we assess balance on observed pretreatment characteristics, $\mathbf{x}_{ji}$, at the time of randomization, and given the properties of randomization, we assume balance on the unobserved covariate u_{ji}.

We define causal effects using the potential outcomes framework (Neyman, 1923, 1990; 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 under $Z_j = 0$. This notation is the same for Designs 1 and 2. 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 otherwise. Writing potential outcomes this way allows for arbitrary patterns of interference among students in the same school but not across schools. The observed outcomes 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, the target counterfactual quantity of interest. In a COS in an educational setting, one reasonable estimand is the following student-level contrast: $y_{Tji} - y_{Cji}$. In our case-study context, this is the change in test scores for student i caused by school-level assignment to myON. Assuming a relevant superpopulation, we could focus on the average causal effect: $E[y_{Tji} - y_{Cji}]$; or the average causal effect for the treated: $E[y_{Tji} - y_{Cji} \mid Z_j = 1]$. With either focal estimand, the expectation is taken with regard to the superpopulation. Of course, these counterfactual quantities are estimable with data only under a set of assumptions.

Assumptions

The first key assumption is the Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1986). The notation above implies SUTVA. Here, we elaborate on what SUTVA indicates 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, we assume that while some variation may exist in the process through which students are exposed to myON, any variation in process corresponds to the same potential outcomes.

SUTVA's second component assumes that the treatment for one student does not spill over to any control student. A benefit of clustered (rather than individual) treatment assignment is that it increases the plausibility of this component of SUTVA. In the COS (or CRT) context, spillover violating SUTVA would need to occur across treated and control schools, for example, if a treatment school student gave her myON account information to a control school student who subsequently used it. Although possible, this seems unlikely to be prevalent. Generally, judging the plausibility of the no-spillover assumption requires qualitative implementation information. In the myON context, we assume that SUTVA holds.

Are SUTVA violations a concern under Design 2? Since only a subset of students within schools are treated, spillovers might occur from treated to untreated students within treated schools. Although possible, violating SUTVA is not a concern. Why? The causal effect of interest is between treated and control schools. Therefore, the relevant spillover is between treated and controls schools even if only some students are treated within a school. When Design 2 is the target trial, one might be substantively interested in such within-school spillover, but this is a different causal question. The analysis of treatment effects under interference is the focus of recent methodological work, for example, Aronow and Samii (2017) and Basse and Feller (2018). In the myON context, such within-school spillover is unlikely, as untreated students in treated schools do not attend summer school.

The next key assumption is the "selection-on-observables" assumption which pertains to the treatment assignment process. This assumption has two parts. First, we assume that there is some set of covariates such that treatment assignment is random conditional on these covariates (Barnow et al., 1980). Formally,

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

That is, after conditioning on observed characteristics, $\mathbf{x}_{ji}$, a given school's probability of assignment to treatment is related neither to the potential outcomes of its students $(\mathbf{y}_{Tji}, \mathbf{y}_{Cji})$ nor to unobservables $(\mathbf{u}_{ji})$. We assume there are no unobservable differences between the treated and control groups. This assumption requires investigators to ask: How could two schools that are identical on all meaningful background characteristics nonetheless receive different treatments? Critically, the selection-on-observables assumption is nonrefutable; it cannot be verified with observed data (Manski, 2007). Therefore, any COS should include sensitivity analyses to consider the sensitivity of results to a possible unobserved confounder. We discuss sensitivity analysis in Section 3.4. Although this assumption may seem implausible, in many examples, treatment assignment depends on observed data only (Dehejia & Wahba, 1999; Fralick et al., 2018; Hernán & Robins, 2016; Keele et al., 2020; Wong et al., 2017).

The second part of the selection-on-observables assumption pertains to "common support." Formally, we assume that all clusters have some probability of being treated or untreated such that $0 < \pi_j < 1$. That is, for no cluster is treatment either guaranteed or prohibited. In practice, large pretreatment covariate imbalances between treated and untreated clusters and/or units is a telltale signal of problems with common support. Such imbalances often arise when treated units differ substantially from control units. When this occurs, trimming may be necessary (e.g., removing some observations from the analytic sample), either at the unit or cluster level, to enforce common support and improve balance.

Trimming is not without consequence, however, as it changes the causal estimand. After trimming, the causal estimand describes the causal effect for the population 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 to estimate treatment effects, as we discuss below.

Explicating the Assignment Process

The modern literature on observational studies emphasizes the role of the treatment-assignment mechanism (Rubin, 2008). Indeed, the assignment mechanism is critical to COS design. Since an observational study's key assumption pertains to whether treatment assignment is based on observed data, understanding how treatment assignment operates is critical. Next, we review important aspects of clustered treatment assignment, including the advantages that cluster-level treatment assignment affords in observational studies.

In any observational study, the investigator should understand and explicate the treatment assignment process. We recommend the following steps. First, understand whether the assignment process can be described as a "natural experiment." Nonexperimental, but haphazard or arbitrary assignment is often characterized as a natural experiment—the hope being that natural circumstances give rise to an arguably random assignment process (Murnane & Willett, 2010). Although haphazard treatment assignment can require considerable judgment and contextual knowledge to justify, the goal is to reduce the bias associated with treatment self-selection. For many natural experiments, analysts still rely on covariate adjustment. When covariate controls are introduced, the analyst is still relying, at least in part, on the selection-on-observables assumption. In this way, observational studies and natural experiments are related. In fact, the

principles we outline for COSs apply to natural experiments.

Second, if a study in not a natural experiment, the investigator should identify the decision-makers responsible for treatment allocation; any factors used to determine assignment (Rubin, 2008); and whether the assignment process is such that decision makers controlled treatment allocation for others within some defined population. In education applications with grouped treatments, this third feature is common and preferable, the advantage being that the selection process is more likely to be based on observed information. Although self-selected treatment assignment may reflect observed factors, it is more likely influenced, at least partially, by unobserved factors, such as a child's motivation or family's expectation regarding the treatment's benefits.

For COSs in educational settings, outside decision makers who are not directly exposed to the treatment often control treatment assignment. For example, district officials often make decisions about selecting schools for treatments. This selection structure offers a key advantage. In general, it should be possible to identify who participated in the treatment-assignment process and the factors used in decision making. Qualitative information typically is critical in this process. For example, WCPSS centrally allocated myON to selected summer school sites based on factors, including internet bandwidth, computer access, and regional distribution.[1] Thus, all summer school students who attended an elementary school close to a selected summer school site used myON. Schools had no input into program allocation. Thus, treatment assignment primarily was a function of school-level data available to district administrators, rather than, for example, teachers' interest in myON. Thus, the selection-on-observables assumption appears reasonable in this context.

Third, understanding the assignment process allows the investigator to identify the study's target trial analogue. In our application, beyond the selection of schools, a secondary, student-level selection process occurred, whereby students were identified for summer school based on their standardized test performance, per state policy. Thus, we must consider this second assignment mechanism. Because student-level selection was governed by state guidelines, student populations should not differ systematically across treatment and comparison schools. 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 illustrated below, our data follow this pattern.

Next, Hansen et al. (2014) demonstrate the advantages of cluster-level treatment assignment in observational studies. Specifically, group-level treatment assignment can reduce the potential for selection bias. For technical details, we refer readers to Hansen et al. (2014) but here convey the intuition. myON is a commercial product; one might imagine a salesperson motivated to bias evidence in favor of the product. The most effective way to do so would be to form a treatment group of individually-selected, higher-performing students who would exhibit stronger reading performance regardless of whether they used myON. If the salesperson is required to select entire schools for myON, however, the mix of students within schools will make it more difficult to guarantee better outcomes under myON. By selecting intact groups, the salesperson is less able to target high performers who would bias results in myON's favor. Therefore, group-level assignment helps to limit bias from purposeful treatment selection. A limitation is that the analyst cannot quantify how much bias is eliminated.

Statistical Analysis

Whatever the advantages of COSs, they remain observational studies. Thus, treated and control groups commonly will differ on baseline covariates, and the analyst will need to use a method of statistical adjustment to remove overt bias and increase comparability. Here, we highlight conventional and more modern approaches to statistical adjustment for COSs.

Statistical Adjustment Methods

In education, random-effects regression models are frequently used for statistical adjustment. In a COS that relies on the selection-on-observables assumption, covariates are added to the model to remove overt biases—observable differences between the treated and untreated clusters. A limitation of relying on regression-based strategies in a COS is that they can elide over any lack of actual overlap between treatment and comparison schools in covariate distributions. Areas outside of common support can be particularly problematic, since they require extrapolation and, in turn, results may suffer from model dependence. That is, conclusions may depend on the regression model's functional form.

It is not that regression-based analysis is not useful for COSs. Rather than turning directly to covariate-controlled regressions for assessing treatment effects, we advocate first taking steps to ensure balance and common support. Then, having obtained an analytic sample where balance and common support hold, regression can be used for treatment effect estimation. We discuss this further below.

Propensity score methods are one alternative to regression modeling. In a COS, the statistical adjustment strategy needs to account for the data's multilevel structure. With propensity score methods, this is done by estimating the propensity score using, for example, a random effects logistic regression model (Arpino & Mealli 2011; Hong & Raudenbush, 2006; Li et al., 2013). However, multilevel models often fail to converge when used to estimate propensity scores (Zubizarreta & Keele, 2016). Therefore, although

propensity score methods are a reasonable strategy when the treatment is allocated at the individual level, the same is not always true with cluster-level assignment. When model convergence issues hamper fitting propensity score models with hierarchical data, little can be done.

Matching. Matching provides another adjustment method designed to mimic a randomized trial by constructing a set of treated and control units that are comparable on observed, pretreatment characteristics. Matching methods primarily have been developed to handle individual-level treatment assignment, and a large literature has articulated best practice in this context (Rosenbaum, 2020). Matching studies have been used to evaluate socially-relevant interventions (Stuart, 2010), and methodological research has investigated the extent to which matching yields impact estimates similar to those achieved through experimental design (Cook et al., 2008; Dehejia & Wahba, 1999).

Just as we can use individual-level matching to mimic an individual-level RCT, 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, strategies for matching with grouped treatments are less well developed. Extant work has focused on multilevel data structures but with applications where clusters are relevant in some way but not for grouped treatments. For example, Steiner et al. (2013) consider matching with multilevel data but assume individual-level assignment. Stuart (2007) discusses group-level matching using group-level data only. Stuart and Rubin (2008) also focus on matching with multilevel data, but advocate building a comparison group from multiple sources when a single comparison site is not a sufficient match for a given treated group (Stuart & Rubin, 2008). This approach considers matching only on student-level characteristics, rendering it less relevant to COSs in which school-level covariates are critical.

Recently, Zubizarreta and Keele (2016) and Pimentel et al. (2018) have developed matching methods specifically for COSs. The resulting matching method mimics a CRT by creating comparable treated and comparison clusters and units within clusters to remove overt bias at the individual and group levels.

In the context of COSs, 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 to increase treatment-control comparability on covariates of critical importance from a scientific standpoint. For example, an investigator can opt to balance baseline test scores more closely than other covariates such as school size. Third, the investigator can trim the sample to yield the set of observations with the highest levels of comparability.

Our primary goal is to consider COS design. Nevertheless, here, we briefly summarize the mechanics of multilevel matching, as we illustrate an application in the case study below. We contrast this process with strategies for implementing standard, single-level matches. In a standard match, the user selects covariates on which to match, and these covariates are used in one of two ways. One option is for the analyst to first estimate a propensity score model and then match units on estimated propensity scores. Alternatively, the covariates may be used to generate a distance matrix—typically based on a Mahalanobis distance—which captures the multidimensional distance between each possible treatment-comparison matched pair. In a basic pair match, treated and comparison units are matched to minimize these distances. Either of these matching variants can be applied to COS data if only school-level covariates are used. However, with multilevel data available, we seek to incorporate student-level information into the school match, even when our goal is to match at the school level only.

To implement a multilevel match, the analyst must specify several parameters. First, the analyst must identify the student- and school-level covariates to be used in the matching algorithm. The analyst also must specify the design. Here, we select between Design 1 (matching only at the school level) and Design 2 (matching at the school and student levels). Next, the analyst can specify the balance priority for the school-level covariates such that the algorithm will seek to balance higher priority covariates before lower priority covariates. Based on this information, the software computes all possible treated-to-comparison *student-level* distance matrices and the size of all these possible matches. These results are then used to compute a new school-level distance matrix that is based on these student-level matches and the school-level covariates. That is, potential school-level matches are assessed based on both the similarity of school-level measures and the similarity of the students within the schools. The algorithm then uses this distance matrix to produce an "optimal" match solution, meaning that it selects a mapping to minimize the sum of the distances between treatment and comparison observations (Rosenbaum, 1989). With a match complete, balance statistics are computed. If necessary, further improvements in balance can be achieved through altering the balance priority for school covariates, adding a propensity score caliper based on school-level covariates, or dropping schools that most contribute to imbalances. We illustrate this process below and refer interested readers to Keele et al. (2020) for a nontechnical review and to Zubizarreta and Keele (2016), Pimentel et al. (2018), and Pimentel et al. (2015) for technical discussions of multilevel matching for COS designs.

Although random-effects regression models alone are not our preferred method for COS analysis, they can be fruitfully combined with matching. After matching, the analyst can regress the outcome on the treatment indicator using a random-intercepts model. Regression modeling is also

useful in that post-matching covariate adjustment with regression can handle imbalances that 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 analytic tool once matching is complete.

Overlap

As discussed above, a key assumption for a COS is common support or overlap of baseline covariate distributions. When little overlap exists between treated and control covariate distributions, trimming units via matching is one method to enforce overlap. Care should be taken, however, as after trimming, the causal estimand is more local; it applies only to a subset of all treated units. In a COS, trimming even a small number of treated schools may mean losing a large percentage of treated units. In other words, trimming even a small number of clusters may make the treatment effect estimate very local. In such cases, no simple remedy exists, since we should not estimate treatment effects using treated and control observations that are not comparable.

Inference

A key principle of inference for COSs is that the analyst must correct estimates of statistical uncertainty to account for clustering. Failure to do so will result in standard errors that are, at times, grossly underestimated given that the correlation among students in the same cluster has not been accounted for (Angrist & Pischke, 2009; Hayes & Moulton, 2009). Generally, the investigator should account for clustering at the level at which the treatment has been assigned (Abadie et al., 2017). Standard errors can be corrected using a generalization of clustered standard errors developed by Liang and Zeger (1986) or via random-effects regression modeling.

When matching methods are used, regression-based corrections can account for clustering. After matching, the analyst should include a clustering correction for schools and the paired school clusters (Abadie & Spiess, 2019), accounting for clustering both within schools and within matched school pairs. This method requires a sufficiently large number of clusters for valid inferences. 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). For example, within matched pairs, the analyst randomly reassigns treatment status and estimates a treatment effect. Doing this repeatedly allows the construction of a null distribution of treatment effects against which to evaluate the treatment effect estimated for actual assignment. The resulting inferences are valid for any sample size. However, randomization inference methods test the sharp null hypothesis which asserts a zero treatment effect for all schools and students. This differs from the more usual null hypothesis asserting an average effect of zero. In general, when sample sizes are small (e.g., 20-30 total clusters), randomization inference is useful for understanding whether inferences depend on the assumption of large sample sizes.

Sensitivity Analysis

All observational studies should include a sensitivity analysis. Sensitivity analyses often 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 for an original conclusion to be reversed. If a causal inference is sensitive, a slight violation of the assumption may lead to different conclusions. Here, we outline a sensitivity analysis, based on randomization inference, that probes the selection-on-observables assumption and is compatible with a matched study (Rosenbaum, 2002, ch. 4)

To begin, recall that under selection on observables, we assume that any two matched clusters have the same underlying probability of treatment. That is, the coin flip is fair within this pair. Of course, this assumption is strong, and matched clusters may still differ on an unobserved confounder, u_{ji}, that drives treatment selection. Sensitivity analyses allow us to quantify how strong an influence such an unobserved confounder would need to have on selection to alter substantive conclusions.

For example, we might hypothesize that, despite matching, an unobserved covariate renders selection probabilities unequal. If that hypothesized inequality (which Rosenbaum denotes as 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 (and then half as likely) to receive treatment, we can calculate bounds on quantities such as the treatment effect point estimate or associated *p*-value based on a conjectured level of confounding.

Generally, one can vary the Γ parameter to ask what level of confounding would reverse study conclusions. For example, we can observe at what value of Γ the upper bound on a *p*-value exceeds the conventional 0.05 threshold. We can summarize the sensitivity analysis with the Γ changepoint—the Γ value at which a focal estimate is no longer statistically significant. If this Γ value is large, we can conclude that inferences are insensitive to hidden bias related to unobserved characteristics. If the Γ value is small, it suggests that inferences are vulnerable to hidden confounders.

Although our discussion has focused on the level of Γ that would negate a significant treatment effect, this procedure also can be used to consider the level of confounding

TABLE 1
Balance on Student- and School-Level Covariates Before Matching

Student covariates	Treated mean before	Control mean before	*Standardized 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
Non-White teachers	0.14	0.18	−0.26
Title I school	0.90	0.93	−0.11
Schools	20	29	
Summer school students	1,371	2,063	

Note. Sandardized difference for a given variable is computed as the mean difference between treatment and comparison schools or students divided by the pooled standard deviation.

that would mask a treatment effect from being detected, as we illustrate below.

Case Study

Here, we demonstrate concepts with the myON application. Our data contain 3,434 summer school students from 49 elementary schools. These 49 schools were grouped into 19 different summer school sites, eight of which received myON. Our first analytic decision relates to whether we define clusters as elementary schools or summer school sites.

For several reasons, we treat intact elementary schools as our clusters. First, we can reasonably infer that although summer school sites were selected for myON, this process explicitly assigned schools to treatment or control. Second, defining clusters at the school level leads to a larger number of clusters and improves statistical power. Finally, our statistical adjustment strategy employs optimal matching methods designed for COSs by Zubizarreta and Keele (2016) and Pimentel et al. (2018). We benefit from having a greater number of treatment and comparison clusters, as this increases the likelihood of obtaining good cluster-level matches. Thus, our treatment-comparison contrast is between assigning groups of students to summer-school sites that do or do not use myON, under the assumption that schools were otherwise comparable; 1,371 summer-school students from 20 schools used myON.

Next, we consider the appropriate target trial. While entire schools were selected for treatment, the intervention applied only to students required to attend summer school. Although control schools were not selected for myON, the summer school selection process was identical across all schools. In theory, summer school students should be similar across treated and control schools. Nevertheless, the student-level selection process points to Design 2 as the relevant target trial.

Given this, we investigate balance at the school and student levels. Table 1 contains means for the treated and control groups and standardized differences before any statistical adjustments.[2] From Table 1, the imbalances on student-level covariates, including pretreatment test scores, are small, indicating that the summer school selection process is uniform across treated and control schools.

Table 1 also contains balance statistics for school-level covariates. All school-level 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, clear differences are evident between treated and control schools. Treated schools, on average, have higher test scores, lower staff turnover, and a lower percentage of teachers who are non-White. Treated schools also have a higher share of teachers who are novices (i.e., 3 or fewer years of experience).

When comparing the student- and school-level covariates in Table 1, mean differences of a similar magnitude translate

TABLE 2
Balance on School-Level Covariates for Four Different Sets of Match Parameters

Covariate	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
Non-White teachers	−0.26	−0.38	−0.30	−0.02	0.05
Title I school	−0.11	−0.18	0.00	0.00	0.00
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.

to very different standardized mean differences at the student and school levels. This is partially because the standard deviations used to scale mean differences are larger at the student level, as there is more variation within than across schools. In Table 1, given that, for example, there is little variation in the share of English language learners (ELLs) within each school, the mean difference of 2 percentage points translates to a standardized mean difference of −0.29. This is also a function of the school-level selection process, whereas student-level selection operates similarly across schools.

Next, we use matching to address baseline imbalances. The matching process is typically iterative; the analyst performs a match, assesses the resulting balance, and then fine-tunes the matching procedure until balance is deemed acceptable. Just as outcome measures are not available at the time of randomization in an experimental study, the analyst should not examine outcomes when implementing matching. The CRT analogue to this process is conducting a randomization, assessing balance on baseline measures, and rerandomizing if baseline equivalence is not satisfied (Morgan & Rubin, 2012).

Instead of presenting results from the match with the best balance, we present a series to illustrate the iterative nature of the matching process. In doing so, we highlight additional tools for improving balance: balance prioritization, calipers, and subsetting. We refer readers to Keele et al. (2020) for further discussion of these tools. For matching, we use the R package matchMulti built specifically for matching with COS designs (Keele & Pimentel, 2016).

Our first match is based on the match algorithm defaults. At the defaults, no covariate is given priority, and no treated schools are dropped. The resulting sample includes 40 schools, with 20 treatment schools pair-matched to a control school without replacement. Table 2 (column 2) reports on this match in which some but not all of the standardized differences improve. In Match 2, we add covariate prioritization with which we select sets of covariates to prioritize in terms of balance. Such prioritization is useful, because science and context may justify preferring closer balance on certain measures.

For covariate prioritization, we define two covariate sets. Set 1 includes the school-level test score measures. Set 2 includes the proportion of ELLs and the proportion of non-White teachers. Under balance prioritization, the matching algorithm works to balance the set 1 covariates first, followed by the set 2 covariates. The remaining covariates receive lowest priority for balance. In Match 2, balance on the test score measures is improved, however improvements for the set 2 covariates are minimal.

Next, we applied a school-level caliper. The matchMulti package includes a function that calculates a school-level propensity score, which is the estimated probability of treatment selection based on baseline measures. We can impose a caliper on this estimated propensity score as another tool to improve balance. We set the caliper to 0.20, which forbids school-level matches differing 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, which contains the results from this match, is generally better, although balance is worse on the proportion of novice teachers. Note that this match discarded some treated schools, since for these schools, the caliper constraint could not be satisfied. Once the use of a caliper discards schools, optimal subsetting is a better tool for match refinement. This is because with optimal subsetting, one can achieve similarly good balance without losing as many treatment sites as might be lost with a caliper strategy.

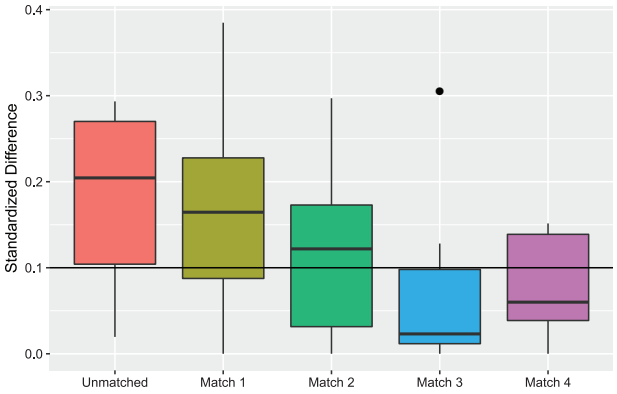

FIGURE 1. *Boxplots of the distribution of absolute standardized differences for school-level covariates.*

With 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 iteratively adjusting this number downward, treated schools are dropped one-by-one until balance improves. For example, if there are 20 treated schools and the optimal subset number is 19, the algorithm will discard the treated school with the poorest match. In general, we recommend dropping schools one-by-one until balance is acceptable.

We improve balance on the proportion of novice teachers by dropping four treated schools via optimal subsetting and rematching. The resulting match (Match 4) excludes the four treated schools with the largest covariate imbalances. For this match, the casual estimand differs; 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 schools. Is this still a population of interest? We cannot answer this question using statistics. In practice, we advise examining descriptive statistics for the remaining treated population to understand whether and how it differs from the full treated population.

Next, we plot the distribution of covariate standardized differences for each match (Figure 1) and observe clear patterns. First, the default match settings do improve balance overall (Match 1), but a few covariates remain highly imbalanced. Second, Match 3 is well-balanced with the exception of one covariate. This 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, balance in student-level covariates remains roughly the same across the matches (Table 3). Taken together, there is little evidence that treatment selection was a function of student-level characteristics for summer school participants.

TABLE 3
Balance on Student-Level Covariates

Student-level characteristic	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

TABLE 4
Outcome Estimates for the Treatment Effect of the myON Reading Program

Adjustment method	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.

Outcome Estimates

Next, we assess the effectiveness of myON for improving performance on the end-of-summer i-Ready reading assessment in our matched sample. With matching complete, we can estimate treatment effects. We use multilevel models with a random intercept, clustering at the school and matched school-pair levels and regressing the outcome on the treatment indicator.

One advantage of regression-based estimation is that the analyst can add baseline covariates to the model. It is useful to include 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 the percentage of novice teachers and the staff turnover rate. To more completely remove bias from the imbalance in these covariates, we include them in the treatment effect model.

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, little difference exists between the unadjusted and adjusted estimates. This suggests either that little self-selection is present, 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 true here. Estimates based on regression alone, matching, and 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 myON developers.[3] Finally, we note that the causal estimand for Match 4 differs from the others. In Match 4, we dropped four treated schools, so the Match 4 results do not apply to the entire treated population. However, that difference appears unimportant, given the similarity of substantive conclusions between Match 4 and the other matches.

In this example, the treatment effect estimates do not vary across statistical adjustment strategies. Is this evidence that these choices are inconsequential? Design choices—including the type of match—should be made without reference to outcomes. Such choices may be of consequence in other applications. In general, if we use regression alone, we cannot be sure that inferences are not overly dependent on the model to extrapolate between treatment and control sites with poor overlap. The inferences we derive and our confidence in them have more to do with the strength of our design process and less to do with how results may change across the different strategies. Sensitivity analyses can increase our confidence further.

Sensitivity Analysis

We can use sensitivity analyses to determine whether it would take a weak or strong unobserved confounder to render a significant treatment effect no longer significant. In the myON example, however, treatment effect estimates are small, and confidence intervals include zero, so we fail to reject the null hypothesis of no treatment effect. Given the null results, one might conclude that sensitivity analyses are not needed. Here, we illustrate how to 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 such an effect actually exists. We can explore this possibility using a test of equivalence with a sensitivity analysis (Rosenbaum, 2008, 2010; Rosenbaum & Silber, 2009).

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, an 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 standard deviations, as 0.20 is generally considered a meaningful effect size in education research (Kraft, 2020). Therefore, the relevant null hypothesis 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 *p*-values $\overleftarrow{H}_0^{(\delta)}$ and $\overrightarrow{H}_0^{(\delta)}$ are rejected (Rosenbaum & Silber, 2009). We can apply the two tests without correction for multiple testing, setting $\alpha = 0.05$ for each test, since we test two mutually exclusive hypotheses. Thus, we test whether our study's estimate differs from other possible treatment effects represented by δ. With a test of equivalence, we cannot 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 farther it will be from δ and the smaller the *p*-values will be.

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

Were our study a CRT, we could be confident that the results were not due to unobserved treatment-control group differences. In a COS, however, we may reject the null hypothesis of equivalence due to hidden confounding. The test above is conducted under the assumption of no hidden bias (e.g., $\Gamma = 1$). However, with sensitivity analyses we can explore whether and to what extent the test of equivalence is sensitive to potential biases from non-random treatment assignment (e.g., bias from a confounder).

To do so, we repeat the test of equivalence, but use Γ values 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 Γ value at which the upper-bound on the *p*-value is greater than 0.05. This is the level 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 implying that if there were a binary unobserved confounder that caused the odds of treatment to differ by 30%, this could explain the result from the test of equivalence.

Is this a large or small Γ value? To provide a benchmark, we regress treatment status on the observed covariates using a logit model and calculate odds ratios for the covariates to compare to those from the sensitivity analysis. If the covariate odds ratios are smaller than the Γ value, the hidden confounder would need to have an effect on the odds of treatment larger than that of the observed covariates. We would interpret this as a robust result, since the effect of the unobserved

confounder would need to be larger than that of the observed data. However, if the Γ value is smaller, then the unobserved confounder could be similar to observed confounders. In our case study, if we increase composite school test scores by one-tenth of a standard deviation, for example, that increases the odds of being treated by 1.42. Since a Γ value of 1.3 is less than 1.42, we conclude that an unobserved confounder could easily mask an educationally meaningful effect.

Discussion

Although randomized trials are considered the "gold standard" for conducting educational effectiveness research, they are not always feasible. Furthermore, an investigator may have questions about the efficacy of an intervention after it has been implemented in a nonrandom manner. In educational contexts, such nonrandom allocation often occurs at the cluster (e.g., school or classroom) rather than individual level.

In such instances, thoughtfully designed COSs together with sensitivity analyses are an important tool in the education analyst's arsenal. Thoughtful design is key to conducting a high-quality COS. We outline principles for COS design and advocate designing COSs with their CRT analogue as a guide. Analysts should focus on the assignment mechanism and identify the factors that guided treatment allocation. We further advocate multilevel matching strategies for achieving treatment-control balance and common support prior to the application of regression or other strategies to estimate treatment effects.

The weakness of a COS, of course, is that even with thoughtful design and analysis, one can never definitively know whether a critical unobserved confounder is driving an impact estimate or whether such an unobserved measure is masking true effects. Nevertheless, sensitivity analyses 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 sum, there is much to be learned from thoughtfully designed and implemented COSs.

Acknowledgments

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 (Grant # 201900074). The opinions expressed here do not necessarily reflect those of the Spencer Foundation. All errors are our own.

ORCID iD

Matthew A. Lenard https://orcid.org/0000-0003-2234-0666

Notes

1. District personnel provided documentation regarding myON's site-selection process and launch.

2. The standardized difference for a variable is computed as the mean difference between treatment and comparison schools or students, divided by the pooled standard deviation (Cochran & Rubin, 1973; Rosenbaum & Rubin, 1985; Silber et al., 2001). A standardized difference of less than one-tenth of a standard deviation is considered acceptable, since a randomized experiment might yield discrepancies of this size (Cochran & Rubin, 1973; Silber et al., 2001; Rosenbaum, 2010; Rosenbaum & Rubin, 1985).

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

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* (Working Paper No. 24003). National Bureau of Economic Research. https://doi.org/10.3386/w24003

Abadie, A., & Spiess, J. (2019). *Robust post-matching inference* [Working paper]. https://scholar.harvard.edu/files/spiess/files/robust.pdf

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics*. Princeton University Press. https://doi.org/10.1515/9781400829828

Aronow, P. M., & Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics*, *11*(4), 1912–1947. https://doi.org/10.1214/16-AOAS1005

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, *55*(4), 1770–1780. https://doi.org/10.1016/j.csda.2010.11.008

Barnow, B., Cain, G., & Goldberger, A. (1980). Issues in the analysis of selectivity bias. In E. Stromsdorfer & G. Farkas (Eds.), *Evaluation studies* (Vol. 5, pp. 43–59). Sage.

Basse, G., & Feller, A. (2018). Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, *113*(521), 41–55. https://doi.org/10.1080/01621459.2017.1323641

Borman, G. D., Slavin, R. E., Cheung, A. C., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal*, *44*(3), 701–731. https://doi.org/10.3102/0002831207306743

Capstone Digital. (2015). *myON: A complete digital literacy program*.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankyha—Indian Journal of Statistics, Series A*, *35*(4), 417–446.

Cook, T. D., Shadish, W., & Wong, V. C. (2008). Three conditions under which observational studies produce the same results as experiments. *Journal of Policy Analysis and Management*, *27*(4), 724–750. https://doi.org/10.1002/pam.20375

Curriculum Associates. (2015). *i-Ready Diagnostic & Instruction: User guide* (Version 6.0).

Dehejia, R., & Wahba, S. (1999). Causal effects in non-experimental studies: Re-evaluating the evalulation of training programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062. https://doi.org/10.1080/01621459.1999.10473858

Donner, A., & Klar, N. (2004). Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health*, *94*(3), 416–422. https://doi.org/10.2105/AJPH.94.3.416

Fralick, M., Kesselheim, A. S., Avorn, J., & Schneeweiss, S. (2018). Use of health care databases to support supplemental indications of approved medications. *JAMA Internal Medicine*, *178*(1), 55–63. https://doi.org/10.1001/jamainternmed.2017.3919

Hansen, B. B., Rosenbaum, P. R., & Small, D. S. (2014). Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *Journal of the American Statistical Association*, *109*(505), 133–144. https://doi.org/10.1080/01621459.2013.863157

Hayes, R., & Moulton, L. (2009). *Cluster randomised trials*. Chapman & Hall/CRC. https://doi.org/10.1201/9781584888178

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal Of Epidemiology*, *183*(8), 758–764. https://doi.org/10.1093/aje/kwv254

Hong, G., & 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*(475), 901–910. https://doi.org/10.1198/016214506000000447

Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, *50*(2), 373–419. https://doi.org/10.3368/jhr.50.2.373

Keele, L. J., & Pimentel, S. (2016). *matchMulti: Optimal multilevel matching using a network algorithm* (R package, Ver. 1.1.5).

Keele, L. J., Lenard, M., Miratrix, L., & Page, L. (2020). *Matching methods for clustered observational studies in education* (EdWorkingPaper: 20-235). Annenberg Institute at Brown University: https://doi.org/10.26300/r5hw-g721

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, *32*(19), 3373–3387. https://doi.org/10.1002/sim.5786

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

Manski, C. F. (2007). *Identification for prediction and decision*. Harvard University Press.

MetaMetrics. (2012). *The Lexile Framework for Reading* [Technical report].

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, *40*(2), 1263–1282. https://doi.org/10.1214/12-AOS1008

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

Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (D. M. Dabrowska & T. P. Speed, Trans.). *Statistical Science*, *5*(4), 465–472. https://doi.org/10.1214/ss/1177012031 (Original work published 1923)

Ortlieb, E., Sargent, S., & Moreland, M. (2014). Evaluating the efficacy of using a digital reading environment to improve reading comprehension within a reading clinic. *Reading Psychology*, *35*(5), 397–421. https://doi.org/10.1080/02702711.2012.683236

Pimentel, S. D., Kelz, R. R., Silber, J. H., & 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*(510), 515–527. https://doi.org/10.1080/01621459.2014.997879

Pimentel, S. D., Page, L. C., Lenard, M., & Keele, L. J. (2018). Optimal multilevel matching using network flows: An application to a summer reading intervention. *Annals of Applied Statistics*, *12*(3), 1479–1505. https://doi.org/10.1214/17-AOAS1118

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173–185. https://doi.org/10.1037/1082-989X.2.2.173

Renaissance Learning. (n.d.). *myON: A complete digital literacy program*. https://www.renaissance.com/products/myon-reader/

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*(408), 1024–1032. https://doi.org/10.1080/01621459.1989.10478868

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). Springer.

Rosenbaum, P. R. (2008). Testing hypotheses in order. *Biometrika*, *95*(1), 248–252. https://doi.org/10.1093/biomet/asm085

Rosenbaum, P. R. (2010). *Design of observational studies*. Springer-Verlag.

Rosenbaum, P. R. (2012). Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics*, *21*(1), 57–71. https://doi.org/10.1198/jcgs.2011.09219

Rosenbaum, P. R. (2020). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, *7*, 143–176. https://doi.org/10.1146/annurev-statistics-031219-041058

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods. *The American Statistician*, *39*(1), 33–38. https://doi.org/10.1080/00031305.1985.10479383

Rosenbaum, P. R., & 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*(486), 501–511. https://doi.org/10.1198/jasa.2009.0016

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (1986). Which ifs have causal answers. *Journal of the American Statistical Association*, *81*(396), 961–962. https://doi.org/10.1080/01621459.1986.10478355

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20–36. https://doi.org/10.1002/sim.2739

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, *2*(3), 808–840. https://doi.org/10.1214/08-AOAS187

Silber, J. H., Rosenbaum, P. R., Trudeau, M. E., Even-Shoshan, O., Chen, W., Zhang, X., & Mosher, R. E. (2001). Multivariate matching and bias reduction in the surgical outcomes study. *Medical Care*, *39*(10), 1048–1064. https://doi.org/10.1097/00005650-200110000-00003

Steiner, P., Kim, J.-S., & Thoemmes, F. (2013). Matching strategies for observational multilevel data. In *JSM proceedings* (pp. 5020–5032). American Statistical Association.

Stuart, E. A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher*, *36*(4), 187–198. https://doi.org/10.3102/0013189X07303396

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Stuart, E. A., & Rubin, D. B. (2008). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, *33*(3), 279–306. https://doi.org/10.3102/1076998607306078

Torgerson, D. J. (2001). Contamination in trials: Is cluster randomisation the answer? *British Medical Journal*, *322*(7282), 355–357. https://doi.org/10.1136/bmj.322.7282.355

Wong, V. C., Valentine, J. C., & Miller-Bains, K. (2017). Empirical performance of covariates in education observational studies. *Journal of Research on Educational Effectiveness*, *10*(1), 207–236. https://doi.org/10.1080/19345747.2016.1164781

Zubizarreta, J. R., & 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*(518), 547–560. https://doi.org/10.1080/01621459.2016.1240683

Authors

LINDSAY C. PAGE is an associate professor of research methodology and a research scientist in the Learning Research and Development Center at the University of Pittsburgh. Her research focuses on quantitative methods and their application to questions regarding the effectiveness of educational policies and programs across the preschool to postsecondary spectrum.

MATTHEW A. LENARD is a PhD student at the Harvard Graduate School of Education. His research focuses on the economics of education, teacher labor markets, and program and policy evaluation.

LUKE KEELE is an associate professor at the University of Pennsylvania. His research focuses on applied statistical methods for causal inference including matching, instrumental variables, and differences-in-differences.