Submitted to the Annals of Applied Statistics

OPTIMAL MULTILEVEL MATCHING USING NETWORK FLOWS: AN APPLICATION TO A SUMMER READING INTERVENTION*

By Samuel D. Pimentel, Lindsay C. Page, Matthew Lenard and Luke Keele

University of California, Berkeley, University of Pittsburgh, Wake County Public Schools, and Georgetown University

Many observational studies of causal effects occur in settings with clustered treatment assignment. In studies of this type, treatment is applied to entire clusters of units. For example, an educational intervention might be administered to all the students in a school. We develop a matching algorithm for multilevel data based on a network flow algorithm. Earlier work on multilevel matching relied on integer programming, which allows for balance targeting on specific covariates but can be slow with larger data sets. Although we cannot directly specify minimal levels of balance for individual covariates, our algorithm is fast and scales easily to larger data sets. We apply this algorithm to assess a school-based intervention through which students in treated schools were exposed to a new reading program during summer school. In one variant of the algorithm, where we match both schools and students, we change the causal estimand through optimal subset matching to better maintain common support. In a second variant, we relax the common support assumption to preserve the causal estimand by only matching on schools. We find that the summer intervention does not appear to increase reading test scores. In a sensitivity analysis, however, we determine that an unobserved confounder could easily mask a larger treatment effect.

1. Introduction.

1.1. *Summer Learning Loss.* Summer learning loss, also known as the "summer slide" or "summer setback" occurs when students educated on the traditional school calendar experience a decline in academic skills during the summer when school is not in session (Borman, Benson and Overman 2005; Cooper et al. 2000; Entwisle and Alexander 1992). Summer learning loss is a well-documented phenomenon. Estimates of the average summer learning loss range from roughly one-tenth to one-third of a standard deviation (SD),

*For comments and suggestions, we thank Luke Miratrix and Winston Lin.

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

depending on methodology, subgroup, and academic subject (Borman and Dowling 2006; Quinn 2015; Rambo-Hernandez and McCoach 2015; Skibbe et al. 2012; Zvoch and Stevens 2015). Cooper et al. (1996) find that summer loss is particularly prevalent in math computation and spelling, and estimate an overall loss of 0.14 SDs in math and 0.05 SDs in reading. Since the 1950s, summer school has been a popular strategy to "keep the faucet on" as well as to remediate those who fall behind during the traditional school year (Cooper et al. 2000). Though more recent estimates are not readily available, approximately nine percent of public school students participated in summer school in 2000 (Wirt et al. 2000).

This study investigates the effectiveness of a summer school reading intervention in Wake County, North Carolina. North Carolina state legislation required that students who did not meet district standards at the end of 3rd grade were required to attend summer reading camps or risk retention. In summer 2013, the Wake County Public School System (WCPSS) selected myON, a product of Capstone Digital, for implementation at Title I summer school sites in an effort to boost reading comprehension among the majority low-SES attendees. myON is a form of internet-based software designed to serve primarily as an electronic reading device. The software provides students with access to a library of books and suggests titles to students based on topic interests and reading ability. Students at myON sites used the program for up to one-half hour during the daily literacy block and could continue using the program at home if they had a device and internet connection. The developers of myON claim that students using the software will 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).

Not all summer school students in the Wake County school system were given access to the myON reading program. Summer school sessions were held at designated sites, such that students from multiple schools attend summer school at central locations. The myON program was used by teachers at eight of the 19 summer school sites. Summer school sites were selected for myON usage based on a mix of factors including internet bandwidth, computer access, and regional distribution. Students from elementary schools in Wake County were assigned to summer school sites primarily through geographic proximity. Thus all students in a school close to a myON summer school site used the myON program during summer school. Overall, students from 20 schools were exposed to the myON intervention, while students from 29 schools were not exposed to the myON treatment. Principals and schools themselves had no input into program participation.

Given that the intervention was assigned to entire elementary schools, we conduct a clustered observational study of the effectiveness of the myON program.

1.2. *Clustered Observational Studies.* When interventions are randomly assigned, differences between treated and control groups can be interpreted as causal effects, but when subjects select their own treatments, differing outcomes may reflect initial differences in treated and control groups rather than treatment effects (Cochran 1965; Rubin 1974). Pretreatment differences or selection biases amongst subjects come in two forms: those that have been accurately measured, which are overt biases, and those that are unmeasured but are suspected to exist, which are hidden biases. In an observational study of treatment effects, analysts typically use pretreatment covariates and a statistical adjustment strategy to remove overt biases, whereas hidden biases can be considered through a sensitivity analyses, as we show.

Matching estimators are one method of statistical adjustment for the removal of overt biases designed to mimic a randomized trial by constructing a highly comparable set of treated and control units. In many settings, treatments are applied to clusters of units instead of to single units. Clustered treatments are common in educational settings as treatments are applied to or withheld from entire schools rather than individual students or teachers. The myON reading intervention is a treatment of this type, as the reading program was offered to all students in schools selected for treatment, and withheld from all students in schools that did not receive the myON reading intervention. Moreover, students did not select whether their school participated in the myON program.

When treatment is randomly assignment to clusters, this is often referred to as a group randomized controlled trial (RCT). In a clustered observational study, one might attempt to mimic a group RCT by creating comparable pairs of treated and control clusters, since differences in outcomes might reflect overt bias. When treatments are clustered, data typically have a multilevel structure with observed and unobserved covariates at both the cluster and unit levels. For example, in the myON intervention, we observe student-level covariates such as pretreatment test scores as well as school-level covariates such as the number of students enrolled.

Thus to remove overt bias in a clustered observational study , researchers may need to remove treated and control differences in the distributions of covariates at the cluster level, the unit level, or perhaps both levels. That is, we require a statistical adjustment strategy that takes into account the multilevel structure of the data. Standard methods of adjustment for data with

multilevel structure include hierarchical regression or propensity score methods (Hong and Raudenbush 2006; Arpino and Mealli 2011; Li, Zaslavsky and Landrum 2013). Recent work developed a matching algorithm for multilevel data based on integer programming (Keele and Zubizarreta 2016).

Here, we extend the method in Keele and Zubizarreta (2016) by developing a new matching algorithm based on network flow optimization. Network flow optimization is frequently used in operations research and in statistics for optimal matching (Rosenbaum 1989). Our method is optimal in that it produces the smallest set of distances between matched clusters and units, here schools and students. While our method allows less flexibility in requiring specific levels of balance on individual covariates, it is faster and can be scaled to much larger data sets than methods based on integer programming. We also modify the basic algorithm to allow treated units to be excluded from the match in an optimal, dynamic manner, so that researchers can find the matches that are balanced but retain the largest possible sample size. We then apply our algorithm to the data from Wake County to evaluate the myON reading program.

This article is organized as follows. Section 2 describes the causal framework that we employ and the design of the study. Section 3 reviews matching algorithms based on integer programming and network flows. In this section, we develop an optimal multilevel matching algorithm based on network flows. and illustrate via simulation the superiority of our algorithms over a multilevel matching strategy that would match clusters and then units in a two-step process.

We then perform two matches, one in which we retain all treated units, and another where we optimally discard treated units to improve balance. Section 4 shows the resulting matches and analyzes the comparative effectiveness of the myON program in Wake County. Section 5 concludes with a summary and a discussion.

2. Notation, Definitions, and Causal Framework. We begin by defining notation and our causal framework. After matching, there are S matched pairs of schools, $s = 1, \ldots, S$, each with two schools, $j = 1, 2$, one treated and one control for $2S$ total units. The ordered pair sj thus identifies a unique school. Each school sj contains $n_{sj} > 1$ students, $i = 1, \ldots, n_{sj}$. Each pair is matched on a vector of observed, pretreatment covariates: $\mathbf{x}_{sji}$. We let $\mathbf{x}_{sj}$ represent the matrix whose rows consist of the $\mathbf{x}_{sji}$ vectors for each student i in the school indexed by sj with support $\mathbb{X} \subset \mathbb{R}$. A student i in school sj is described by both observed covariates and possibly an unobserved covariate u_{sji}. Included among the student covariates may be

some covariates common to all other students attending that school, and we call such covariates school-level covariates. For example, while gender is a student-level covariate, the proportion of male students in a school is a school-level covariate which takes the same value for all students in the same school. In our study, treatment assignment occurs at the school level as whole schools are assigned to treatment or control. If the j^{th} school in pair s receives the treatment of myON readers, write $Z_{sj} = 1$, whereas if this school receives the control and students are not given myON readers, write $Z_{sj} = 0$, so $Z_{s1} + Z_{s2} = 1$, for each s as each pair contains one treated school and one control school. If $n_{sj} = 1$ for all sj then the clusters are individuals, and we have unclustered treatment assignment.

We use the potential outcomes framework to define causal effects (Neyman 1923; Rubin 1974). In this framework, each student has two potential responses; one response that is observed under treatment $Z_{sj} = 1$ and the other observed under control $Z_{sj} = 0$. We denote these responses with (y_{Tsji}, y_{Csji}), where y_{Tsji} is observed from the ith subject in pair s under $Z_{sj} = 1$, and y_{Csji} is observed from this subject under $Z_{sj} = 0$. Here, y_{Tsji} is the reading test score that student sji would exhibit if he or she uses the myON software, and y_{Csji} is the test score this same student would exhibit if he or she does not use the myON software. Under this notation, we allow for any arbitrary pattern of interference among students in the same school but not across schools. In this context, y_{Tsji} denotes the response of student sji if all students in school sj receive the treatment, while y_{Csji} denotes the response of student sji if all students in school sj receive the control. Therefore, we do not assume that we would observe the same response from student sji if the treatment were assigned to some but not all of the students in school sj. We let $\mathbf{y}_{Tsj}$ and $\mathbf{y}_{Csj}$ represent the vectors of potential outcomes y_{Tsji} and y_{Csji} respectively for each student i in school sj. We do not observe both potential outcomes, but we do observe responses: $Y_{sji} = Z_{sj}y_{Tsji} + (1 - Z_{sj})y_{Csji}$. Under this framework, the observed response Y_{sji} varies with Z_{sj} but the potential outcomes do not vary with treatment assignment. Write $\mathbf{Y} = (Y_{111}, \ldots, Y_{S2,n_{s2}})^T$ for the $N = \sum_{s,j} n_{s,j}$ dimensional vector of observed responses, and write $\mathbf{y}_c$, for the vector of potential responses under control.

Next, we define the causal estimand. First, it is important to note that the estimand depends on the form of the multilevel match. One approach to the planning and design of observational studies is to use an analogous randomized experiment as template (Cochran 1965; Rubin 1974). The form of multilevel match will change depending on the type of experimental design used as a template. In general, we would argue there are two experimental

designs that we might use as templates for the matching. The first is the paired clustered randomized controlled trial. Under this design, clusters are first paired on covariates, and then within each pair, one cluster is selected for treatment, and all units within that cluster are treated. Under this experimental template, we would use the matching algorithm to pair schools and not students, since in the experimental design only clusters are paired. As such, under this design template, the causal estimand is the student level level contrast $y_{Tsji} - y_{Csji}$ caused by *group* level treatment assignment. Note that if we were to assume the existence of an appropriate superpopulation, it might be natural to focus on an average causal effect of the form $E[y_{Tsji} - y_{Csji}]$ where the expectation is taken over the superpopulation. We focus on finite sample inference, although the methods we use can easily be adapted for superpopulation inference (Lehmann and Romano 2005).

However, we might implement a match that pairs both schools and students. This second design may be of interest if the intervention is assigned at the group level but is only targeted at a differentially-selected subset of units within each cluster. For example, Page and Scott-Clayton (2016) conducted a school-level randomized experiment in which they targeted college-intending high school seniors with information and personalized reminders about the process of applying for financial aid. Next, we describe two possible experimental analogues for this type of match. First, data of this type could result from a clustered randomized trial (CRT) with non-random unit level selection. That is, clusters are assigned to treatment and control, but within the cluster some units are targeted for treatment. Despite randomization at the cluster level, an investigator would need to correct for this selection bias via modeling. The same is true in an observational study. Matching on student level covariates as well as school level covariates would allow analysts to model selection at both the school and unit level.

Alternatively, one could conceive of this design as a double blocked design where first schools are paired on the basis of baseline covariates. Then once schools are paired, the students within those schools are paired. Such blocking at the student level would ensure comparability in student level covariate distributions beyond what would be achieved by aggregating student covariates to the school level. However, while such a design is hypothetically possible, we are unaware of any design actually being implemented. In all likelihood, blocking on student level covariates would result in some of the students being discarded especially if the number of students differs across the two schools. Once both students and schools are paired via the blocking, the treatment would then be assigned within school pairs at the school level. Thus, even though treatment assignment is recorded at the school level, the

treatment would only be applied to a subset of the students within the school. Such a design would ensure maximum comparability between both schools and students before randomization occurs.

Theoretically, the myOn intervention conforms to this second template. That is, the myOn intervention was assigned at the school level but was only used by the subset of students within each school that were required to attend summer school, and the method of selecting summer school students may differ from school to school. Therefore we might wish to match both schools and students in our application to correct for selection at the student level. Under this second design template, the causal estimand is a group level contrast for a set of students within the school who are at risk for the treatment. For this match, we define the student level covariate space, $\mathbb{A}_s \subset \mathbb{X}$. Let $\mathbb{1}_{\mathbf{x}_{sji} \in \mathbb{A}_s}$ be an indicator function for the event that $\mathbf{x}_{sji}$ is an element of the set $\mathbb{A}_s$. The causal estimand under this design is $\mathbb{1}_{\mathbf{x}_{sji} \in \mathbb{A}_s}(y_{Tsji} - y_{Csji})$ where $\mathbb{A}_s \subset \mathbb{X}$ is the school-pair-specific portion of the support $\mathbb{X}$ describing students who may be required to attend summer school in both school $s1$ and $s2$. The question of which estimand and associated match is more appropriate is a question for investigators and varies from application to application. As we outline later, our algorithm accommodates both types of matches. Henceforth, we denote a match that mimics a group RCT and does not pair students as Design 1. For a match that pairs both schools and students, we call this match Design 2.

To identify the causal estimand above, we assume that assignment to Z_{sj} depends on observable covariates only. Formally, we must assume that

$$Pr(Z_{sj} = 1 | \mathbf{y}_{Tsj}, \mathbf{y}_{Csj}, \mathbf{x}_{sj}, \mathbf{u}_{sj}) = Pr(Z_{sj} = 1 | \mathbf{x}_{sj}).$$

For brevity, represent the probability on the left-hand side of this statement by π_{sj}. We also assume that all schools have some probability of being treated such that $0 < \pi_{sj} < 1$. The assumption of observable treatment assignment is often referred to as conditional ignorability or selection on observables (Rosenbaum and Rubin 1983; Barnow, Cain and Goldberger 1980). If this assumption holds, potential outcomes will be conditionally independent of treatment assignment and the causal effect of the treatment will be identified. Later, we will probe the plausibility of this assumption using a sensitivity analysis.

The second part of the conditional ignorability assumption is known as the assumption of common support. Common support fails if for any student the true probability of being exposed to the myON intervention is zero or one. Very high or low estimates of the propensity score or high pre-treatment covariate imbalances for students often signal problems with common support,

and in these settings we must either relax this assumption or remove study units to maintain the assumption. Trimming to maintain common support changes the causal estimand such that it only applies to the population of units for which the effect of treatment is marginal: units that may or may not receive the treatment. As such, we could characterize the estimand as more local, since it applies to only a subset of the treated units. Changing the estimand through trimming of treated units may be unproblematic if the data do not represent a well-defined population (Rosenbaum 2012a). See Crump et al. (2009), Traskin and Small (2011), and Rosenbaum (2012a) for further discussion of the common support assumption and methods for dealing with a lack of overlap. The matching algorithm we develop also includes a form of optimal subset matching for applications where common support does not hold.

3. Multilevel Matching.

3.1. *Review: Multilevel Matching.* The goal with matching methods in an observational study is to mimic the structure of an experimental design: treated and control units that are similar in terms of observed covariates. With a multilevel data structure, such as students and schools, the covariates are observed at both levels and units are found within clusters. That is students are nested within school level clusters. As such, a matching algorithm for multilevel data must make units similar on observables at both levels. Of course, even the most successful match provides us with no confidence about the similarity of unobservables for the matched data. One method for matching with multilevel data is as follows:

1. Pair clusters using cluster level covariates.
2. Pair units within paired clusters (i.e. match units while requiring exact matches on cluster-level pair IDs).

Keele and Zubizarreta (2016) show that such an approach is not optimal with respect to minimizing covariate distance. They show that the optimal matching strategy for multilevel data is:

1. Consider each possible combination of one treated cluster and one control cluster. Using unit level covariates, calculate a unit level distance matrix. Summary statistics based on this distance matrix can be used to assess the quality of the units within each possible set of clustered pairs.
2. Pair clusters using cluster level covariates and the score information (based on unit-level covariates) from step 1.

3. Once clusters are paired, optionally form unit level pairs depending on the causal estimand of interest. Under Design 1, we would not pair students, under Design 2 we would pair students.

Such an approach is optimal since it utilizes unit level information in the cluster pairing unlike the aforementioned naive approach. Keele and Zubizarreta (2016) developed an optimal multilevel match using mixed integer programming (Zubizarreta 2012). The key advantage of integer based matching algorithms is that they allow the analyst to target explicit levels of balance for mean differences across treated and control units. For example, these methods allow for the explicit balancing of statistics such as the Kolmogorov-Smirnov (KS) statistic. The drawback to such methods is that the computational time necessary for the match may be lengthy.

Many matching algorithms use network flows instead of integer programming to balance covariates by minimizing the total sum of distances between treated and control units. While network flow algorithms for matching cannot always incorporate the specific balance goals allowed by integer programming, the algorithms are fast and can be applied to very large data sets with fewer difficulties. To that end, we develop an optimal multilevel matching algorithm based on network flows.

3.2. *Multilevel matching based on network flows.* Next, we outline our matching algorithm in its simplest form. First, we denote the number of matched cluster pairs as S. For the Design 1 match, we seek to create S matched pairs of schools such that school-level covariates are balanced across all schools in the matched sample. Under Design 1, matching school level covariates may balance student level covariates, but it may not. For the Design 2 match, we create S matched pairs of schools and, for each such pair, $m_s \leq \min(n_{s1}, n_{s2})$ matched pairs of students (one student from each school) such that school level covariates are balanced and student-level covariates are balanced within each school pair.

For either design, the process starts in an identical fashion. First, student-level matches are conducted for all possible pairings of treated and controlled schools. If there are N_1 treated schools and N_2 control schools, the number of such possible pairings will be $N_1 \times N_2$. Each of these student-level matches is then scored based on the balance it achieves on student-level covariates (worse scores are given to matches with insufficient balance) and on the size of the sample it produces (worse scores are given to matches with small sample sizes). The scoring system is inverted, so that the best matches receive low scores and the poorest ones receive high scores. The scores are then stored in an N_1 by N_2 matrix. Next, schools are matched optimally using

the score matrix as a distance matrix. Below we outline a refinement to this step to better balance school level covariates.

At this point, the investigator can either choose to rematch students such that schools and students within schools are paired or stop with paired clusters such that students are not paired within schools. Importantly, even if the investigator chooses not to rematch students, student level information has been used in the school level match.

We now describe two important refinements to the basic algorithm. Both are designed to allow analysts to improve balance in contexts when the simplest form of the algorithm produces a match where imbalances on covariates are still deemed to be too large by the investigator.

3.3. *Role of refined balance.* Matching on the score matrix alone does not provide any guarantee of balance on school-level covariates (since scores are computed from student-level covariates only). To allow the investigator to improve balance over and above that produce by matching on the score matrix, we include the method of refined covariate balance (Pimentel et al. 2015) in our multilevel matching algorithm. Refined covariate balance is an extension of fine or near-fine balance. Under fine and near-fine balance constraints an optimization routine seeks the closest individual matched pairs such that the overall match has the closest possible balance on a prespecified nominal covariate (Rosenbaum 2010; Yang et al. 2012). Refined covariate balance involves matching under multiple nested near-fine balance constraints that act in order of priority, balancing the first covariate as closely as possible, the second as closely as possible under the constraint of the first, and so on. For example, one might match under 2 levels of near-fine balance, the first requiring that the match exhibit the best possible balance on Title I status, and the second requiring that (once Title I status is balanced) the match achieve the best remaining possible balance on the interaction of Title I status with an indicator for above-average proportion of new teachers.

Adding refined covariate balance to the matching algorithm has two advantages. First, it allows investigators to prioritize balance on some school-level covariates relative to other school-level covariates. If scientific knowledge dictates that some covariates are of higher priority, balance on those covariates can be targeted for improvement via refined covariate balance. Second, in multilevel matching applications, the number of covariates may be large relative to the number of observations at the cluster level. This is the case in our school-level match, where only 49 schools are present (20 of them treated) and 11 important school-level covariates have been identified. In situations of this type, the use of refined covariate balance provides

much stronger guarantees of balance than merely including the covariates in a propensity score formula or a pairwise Mahalanobis distance (Zubizarreta et al. 2012). It is true that in the limit, as sample size approaches infinity while holding the number of covariates fixed, pair matching on a covariate distance will adjust for all observed confounding and bring observed covariates into perfect balance in the matched sample. However, in finite sample situations pair matching often struggles to balance all observed variables, especially when the number of covariates is large relative to the number of observations. Therefore refined balance, which guarantees optimal finite-sample balance for a given set of constraints, is a better choice for multilevel data structures where the number of clusters to match is likely to be small relative to the number of covariates.

3.4. *Optimal Subset Matching in a Multilevel Match.* For a specific data set, we may find either that there is a lack of overlap in the covariate distributions or that balance is poor after the matching is complete. In both cases, too much overt bias remains when the match uses all treated observations. One solution is to trim units from the treated group to maintain the common support assumption or improve balance. Methods such as optimal subset matching (Rosenbaum 2012a) and cardinality matching (Zubizarreta, Paredes and Rosenbaum 2014) are designed to find the largest subset of the treated group such that covariate overlap or balance is deemed acceptable. Next, we add a form of optimal subset matching to our multilevel match to deal with such constraints.

Before we review technical details, we discuss conceptual issues that arise when using optimal subsetting in a multilevel matching context. A multilevel match complicates such trimming since one can choose to trim either treated clusters, units, or both. The type of trimming will depend on the form of the match. If one pairs both clusters and units as would be the case under Design 2, some unit level trimming is almost impossible to avoid. That is, unless the samples of students in control schools are uniformly substantially larger than the samples in all treated schools, some of the student-level matches in step 1 will involve treated groups that are larger or very similar in size to their control groups. In these settings, under student level pair matching, some treated units will invariably be trimmed or excluded from the match. Here the trimming is not done to enforce balance or common support, but is simply a byproduct of the structure of the pair match. Under Design 2, one might also choose to trim students to improve student-level covariate balance. Our algorithm allows for either form of trimming. Finally, under either Design 1 or 2 one may trim treated clusters to maintain common sup-

port and improve covariate balance at the cluster level. This would involve removing complete treated schools from the match.

Optimal subset matching is a network flow algorithm for pair matching which allows the match to exclude treated units by paying a penalty for each match excluded (Rosenbaum 2012a). For a given penalty $\widetilde{\delta}$, optimal subset matching considers only subsets of treated units such that adding any additional treated unit increases the best overall matched cost (distance) by at least $\widetilde{\delta}$. Among these treated subsets, it chooses the one for which the overall matched distance can be made smallest and the matched control group associated with that configuration. The choice of which individuals to exclude is made concurrently with the choice of matched pairs. To prevent the exclusion of overly large numbers of treated units, optimal subset matching may incorporate an additional parameter $\underline{n}$ which specifies a minimum number of treated units that must be included. When $\underline{n}$ is equal to the size of the treated population and there are as many controls as treated, optimal subset matching is equivalent to ordinary optimal pair matching.

In addition to offering a formal definition and an applied example of matching with refined covariate balance, Pimentel et al. (2015) provide a network flow algorithm to solve such matching problems, although it does not allow treated units to be excluded. We use an adaptation of this algorithm that also allows exclusion of treated units. Specifically, the familiar penalty parameter $\widetilde{\delta}$ is used to represent the cost of excluding a treated individual from the match. For sufficiently large values of $\widetilde{\delta}$, the match does not exclude anyone and the algorithm behaves exactly as in the original paper; as $\widetilde{\delta}$ is decreased, more and more treated units will be excluded. For any given value of $\widetilde{\delta}$ and given set of balance constraints, the algorithm guarantees that the match produced has optimal refined balance among matches with the same number of treated units excluded. For a characterization of the optimality of matches produced by this modified algorithm and a technical description of the alterations it requires in the original algorithm of Pimentel et al. (2015), see Pimentel and Kelz (2016).

If it is necessary to trim either treated schools or students then the causal estimand has been changed due to the match. For example, if we trim at the school level only, we would exclude schools with covariates outside set $\mathbb{A}$, a subset of school level covariates. Here, we denote $\mathbb{1}_{\mathbf{x}_{sj} \in \mathbb{A}}$ is an indicator function for the event that $\mathbf{x}_{sj}$ is an element of the set $\mathbb{A}$. The estimand is now $\mathbb{1}_{\mathbf{x}_{sj} \in \mathbb{A}}(y_{Tsji} - y_{Csji})$, a subsample treatment effect defined by the school level covariates in $\mathbb{A}$, as the estimand now depends on school level covariates. If we trim students, our causal estimand is now the subsample treatment effect defined by the student level covariates in $\mathbb{A}_s$. This estimand is identical in

form to the one defined for Design 2: $\mathbb{1}_{\mathbf{x}_{sji} \in \mathbb{A}_s}(y_{Tsji} - y_{Csji})$. In our example, this would be the set of students for whom there is some probability that they attend summer school. If it is necessary to trim both students and schools, then our causal estimand focuses on the population of schools and students for whom the intervention is marginal. Alternatively, it is the subsample treatment effect defined by the school and student level covariates in $\mathbb{A}$ and $\mathbb{A}_s$, respectively. As outlined above, this may be a reasonable decision in an observational study when interest is in a marginal population who might or might not receive the treatment of interest rather than a known, a priori well-defined population. This applies to the myON treatment in that the treated schools are those that happen to be located near a summer school site with the technical capacity for the intervention.

3.5. *A General Algorithm.* Next, we precisely define our approach to multilevel matching with network flows. We outline two different algorithms depending on the design chosen by the investigator. First, we might wish to match under Design 1, where we intend to preserve the causal estimand as a school level effect. We do this by pairing schools, but not pairing students within schools. We define this form of match under Algorithm 1:

1. Create a distance matrix M for all students in the dataset, using student covariates only. For each possible combination of one treated school i and one control school j:
   - Match the students in school i to the students in school j on the appropriate submatrix of M.
   - Assign a score ℓ_{ij} to this match using a pre-specified scoring rule which depends on the size of the matched samples and the balance achieved on student covariates, and store it in a matrix.
2. The score matrix produced by the pairwise school matches now gives distances between all pairwise treated-control school combinations. If desired, this score matrix may be combined with a distance matrix computed from school-level covariates. Use this matrix to match schools, with (optional) refined covariate balance constraints on school covariates via fine balance. Optionally, the analyst may use an optimal subset match with subset penalty $\widetilde{\delta}_2$ to further improve balance. The student-level matched samples consists of all students in any school that was selected in the school match.

Under Algorithm 1, we still use student level information in the pairing of schools, but we do not pair the students directly after schools are matched.

Next, we define Algorithm 2, which matches both schools and students.

1. Create a distance matrix M for all students in the dataset, using student covariates only. For each possible combination of one treated school i and one control school j:
   - Match the students in school i to the students in school j on the appropriate submatrix of M using optimal subset matching with penalty $\widetilde{\delta}_1$ and minimum sample size $\underline{n}_{ij}$.
   - Assign a score ℓ_{ij} to this match using a pre-specified scoring rule which depends on the size of the matched samples and the balance achieved on student covariates, and store it in a matrix.
2. Using the score matrix produced by the pairwise school matches (or if desired, the score matrix and a school-level distance matrix computed from school covariates may be combined), match schools using an optimal subset match with refined covariate balance constraints on school covariates, with subset penalty $\widetilde{\delta}_2$.
3. Combine the student matches computed in step 2 for the school pairs computed in step 3 to produce an overall matched sample of students.

Application of Algorithm 2 results in a set of matched schools with students within those schools that are also paired. The algorithm may trim either treated schools, treated students or both in order to balance covariates and maintain the common support assumption.

3.6. *Choosing appropriate penalty parameters.* Notice that besides requiring the researcher to specify relevant student and school covariates and a set of balance constraints (for school covariates), Algorithm 2 relies on tuning parameters $\widetilde{\delta}_1$, $\widetilde{\delta}_2$, and (for each choice of i and j) $\underline{n}_{ij}$. Algorithm 1 relies only on $\widetilde{\delta}_2$. How should these parameters be chosen effectively?

When excluding treated students, we recommend setting the values $\underline{n}_{ij}$ as a fixed, relatively large proportion (perhaps 80%) of the smaller of the two sample sizes in schools i and j. This ensures that even when students are excluded, most students will be retained and limits the degree to which a matched sample selected in a school pair can differ from the full samples in both schools. Following the recommendation in Rosenbaum (2012b), $\widetilde{\delta}_1$ can be set as a fixed, small percentile of the pairwise Mahalanobis distances among all students in the dataset, perhaps the fifth or the tenth percentile.

Setting $\widetilde{\delta}_2$ in order to exclude entire treated schools can be done by taking quantiles of the values in the school distance matrix (analogous to the strategy for setting $\widetilde{\delta}_1$), especially when no balance constraints on school covariates are used. When balance constraints are present, appropriate values of $\widetilde{\delta}_2$ are more difficult to derive a priori and to interpret, since balance

constraints are implemented internally by imposing additional penalties not reflected in the pairwise distances. Because of the difficulty of setting $\widetilde{\delta_2}$ in this context, the R package `matchMulti` which implements both algorithms offers users the option to supply a desired number of schools to retain, and conducts a binary search in the penalty space (essentially, repeating step 2 of the algorithm over many values of $\widetilde{\delta_2}$ until it finds a penalty inducing the desired sample size.

As a general rule, we recommend excluding as few units as possible; i.e. the initial match may be run with $\underline{n}_{ij}$ values set at 100% of the smaller sample size in the pair i, j (excluding as few students as possible) and instructing the algorithm to set $\widetilde{\delta_2}$ so all treated schools are retained. If balance on student covariates is poor, values for $\underline{n}_{ij}$ and $\widetilde{\delta_1}$ can be decreased gradually, to encourage exclusion of more students, until balance is achieved; similarly, if school balance is poor, the desired number of schools retained can be decreased in increments of 1. Note, however, that the analysts may wish to first impose refined balance constraints on school level covariates via fine balance before using penalties to remove treated schools. The fine balance constraints can be a very effective way to improve balance without discarding schools. Importantly, since outcome data is not examined until after a final match has been chosen, the validity of statistical tests is not affected by this iterative processes checking balance and rematching.

3.7. *Simulation Study.* Next, we evaluate our proposed matching method through a simulation. In the simulation, we compare our multilevel matching algorithm to a match that first matches schools without reference to student covariates. Hereafter, we denote the match that only uses school level covariate information in the first stage as the "naive" match. Before proceeding to the simulation, we review some relevant analytic results. Zubizarreta and Keele (2016) considered the optimality of a multilevel matching algorithm based on integer programming. While our algorithm is based on network flows, the proof in Zubizarreta and Keele (2016) extends to our algorithm. Therefore, we should never expect the naive match to perform better than our optimal approach. However, in a specific application, the naive match may perform as well as the optimal method we propose. Thus a simulation will help us understand whether our optimal approach is clearly superior or whether a naive approach will tend to produce similar results in data like that in the myON evaluation.

To bolster the realism of the simulations, we generated the simulated data from the myON data. However, in the simulated data, we increased the imbalance in the covariates that measure student and school level test

scores in both reading and mathematics. We increased the imbalance by increasing the treated means and by adding a random draw from a Normal distribution with a mean of four and a standard deviation of one. Thus we systematically increased the imbalance in these four covariates, but also introduced stochastic variability in the imbalance across the simulations. Finally, for the two treated schools with the largest multivariate distances from the controls, we increased the imbalance on test scores by additional two-tenths of a standard deviation at both the student and school level for one school and for the other school by a quarter of a standard deviation. To generate outcomes in the simulated data, we first generated an outcome model by regressing the test score outcome in the myON data on the baseline covariates. We then generated simulated outcomes via a linear model based on the simulated data, the coefficients from the outcome model, a treatment effect of two-tenths of a standard deviation, and Normally distributed errors with a mean and standard deviation of one.

We repeated the simulation 1,000 times. For each simulation, we applied our multilevel match algorithm twice. First, we did not allow for optimal subsetting, and for the second run of the multilevel match, we allowed optimal subsetting. We also applied the naive matching method that matched directly on school level covariates. For both matches, we matched under Design 1 and did not match students. We did this for a one specific reason. In the simulated data, we increased the imbalance on student level test scores. The multilevel match incorporates student level information into the school match, while the naive method only does so through aggregated test scores. Thus in the simulation, we should be able to observe whether our matching algorithm can reduce imbalance in student level covariates while preserving the structure of a clustered randomized trial.

For the naive match, we used the optimal matching algorithm in the `R` package `optmatch`. While there are many different matching algorithms that we could have used, we selected this matching algorithm for a specific reason. Our multilevel match also relies on the basic algorithm in the `optmatch` library. The key difference is that our algorithm alters the order of the match by first matching students and then using that information to match schools. A naive implementation of optmatch matches schools first without fully incorporating student level information into the match. Therefore, by implementing the naive method using `optmatch`, we reduce the comparison between the matching algorithms to the order of the match. We did not apply a caliper to any of the matches. For the multilevel matches, we prioritized balanced on test score covariates in the simulation. Prioritizing balance on a subset of covariates will tend to make balance worse on other covariates.

Thus, we seek to observe whether we can improve balance on these covariates while maintaining acceptable balance on the other covariates.

For each simulation, we record measures of both balance and bias in the treatment effect estimates. First, for each test score covariate, we calculated the percentage change in bias reduction. We calculated this by taking the percentage change from before and after matching in the absolute standardized difference for each test score covariate. Second, we recorded the average absolute standardized difference after matching for all the covariates included in the match. Third, we record bias as the average difference between the estimated effect and the true treatment effect. Finally, we calculated a relative measure of bias using the percentage change between the unadjusted estimate of the treatment effect and the adjusted estimate of the treatment effect after matching.

Table 1 contains the results from the simulation. Column 1 of the table contains the results from the naive match. The naive match produces an average standardized difference of 0.30 while reducing test score imbalance by 2.4% to 10%. The multilevel match without optimal subsetting, produces superior imbalance reduction on the test score covariates. The multilevel match reduces bias by at least 11% and in one case by as much as 44%. In particular, it is worth noting the imbalance reduction in the student level covariates. The multilevel match produces greater imbalance reduction for the student level data, since the imbalance in student level covariates is accounted for in the school level match. Moreover, the overall level of balance is improved under the multilevel match as the average standardized difference is 0.21. Column 3 of Table 1 contains the results from the multilevel match where the algorithm performed optimal subsetting by removing two schools from the match. The multilevel match with optimal subsetting produces the best results as the amount of bias reduction is now between 15% to 60%, with an average standardized difference of 0.18. Reducing imbalance also translates into reduced bias in the estimated treatment effect. The naive match reduced bias by 4.5% relative to the unadjusted estimate, while the multilevel match reduced bias by 6.6% and 11.5% when optimal subsetting is applied.

In the simulations, we find that a multilevel match algorithm which considers balance at both student and school level performs better than a naive match that breaks the match into sequential steps. Moreover, the simulation also clearly demonstrates that optimal subsetting can be a useful strategy for reducing bias. Of course, this comes at the cost of altering the causal estimand, since the estimand no longer applies to the entire treated population of schools.

TABLE 1. *Simulation Results*

	Naive Match	ML Match	ML Match w/ Optimal Subset
% Imbalance Reduction[a] - Math Scores - School Level	2.5	39	50
% Imbalance Reduction - Reading Scores - School Level	4	44	60
% Imbalance Reduction - Math Scores - Student Level	5.2	11	15
% Imbalance Reduction - Reading Scores - Student Level	10	22	39
Average Std. Diff.	0.30	0.21	0.18
Bias	-4.92	-4.56	-3.70
% Bias Reduction[b]	4.54	6.60	11.5

Note: [a]Imbalance Reduction records the percentage change in the absolute standardized difference from the unmatched sample to the matched sample. [b]Bias reduction records the difference between the estimated treatment effect after matching and the unadjusted treatment effect.

3.8. *Two Matched Comparisons.* The data from the myON study contain 3434 summer school students from 49 schools, of which students from 20 schools (containing a total of 1371 summer school students) received the myON intervention. Since treatment status was defined through proximity to summer school sites with the technical capacity to support the myON intervention, we judge that our study population does not represent any natural larger population. Moreover as we noted, within each school the intervention only applies to the subset of students who must attend summer school. Thus Design 2 may be the most appropriate for the myOn application. Therefore, we created one matched sample using Algorithm 2, pairing both schools and students within schools. Moreover, we allow in our match for the removal of both treated schools and students to maintain overlap in the treated and control distributions. Therefore, our study population will not be representative of the larger population of students for whom the myON intervention is not marginal. It also implies that our causal estimand which is defined at the school level only applies to a set of marginal students, not the entire set of students that receive the school-level myON treatment. However, for purposes of comparison, we also include one additional match under Design 1. This match will at least maintain the status of our group level causal estimand. However, this may come at the cost of higher levels of overt bias. We created this matched sample using Algorithm 1, which paired schools using student level balance information but retained all students within the matched schools.

For the match based on Design 2, we first created a robust Mahalanobis distance matrix (Rosenbaum 2010, section 8.3) among all students in the data based on individual pre-treatment reading and math test scores, Hispanic and African American indicator variables, sex, and indicators for participation in the special education program. The $\widetilde{\delta}_1$ parameter was set as the 75th percentile of the costs in the overall robust Mahalanobis matrix, meaning the match will prefer to exclude treated units rather than form pairs with distances from the largest quantile of possible pair distances, and the $\underline{n}_{ij}$ parameter was set to $\min\{0.8T_i, C_j\}$ where T_i is the number of students from treated school i and C_j is the number of students from control school j. This ensured that wherever possible at least 80% of the treated students in each school were retained.

Once student-level matches were computed, they were scored as follows:

1. Initialize score to a large value L.
2. Subtract the number of matched samples formed.
3. Check post-match balance on the following student-level covariates: individual pre-treatment reading and math test scores, Hispanic and

African American indicator variables, sex, and indicators for participation in special education. For each absolute standardized difference (see Section 4.1) above 0.2, add a penalty of $10L$.

This scoring strategy prioritizes large matches over small matches (since large matches have lower scores after step 2), but the large penalties in step 3 ensure that adequate balance is the primary criterion in assigning scores.

Next, schools were paired following step 2 of the algorithm. Two layers of refined covariate balance constraints were added, each layer an interaction of categorical school covariates (e.g., Title 1 status) and appropriate coarsenings of continuous school covariates (i.e. indicators for whether individual values of the proportion of new teachers, proportion of English-proficient students, etc. exceed certain quantiles of their distributions). After examining larger matches and finding them insufficiently balanced, we ultimately excluded four treated schools for a final matched sample of 16 school pairs (this result was obtained by setting $\widetilde{\delta}_2$ to 10^7). Combining the first-stage school-to-school matches corresponding to the matched school pairs, we obtained an overall matched sample of 1,532 students, 766 from schools with the myON intervention and 766 without (meaning a total of 605 treated students were trimmed from the treated sample).

For the second match under Design 1, we used a scoring function similar to the one used in Step 1:

1. Initialize score to a large value L.
2. Subtract the harmonic mean of the number of treated students and the number of control students.
3. Check balance on the following student-level covariates: individual pretreatment reading and math test scores, Hispanic and African American indicator variables, sex, and indicators for participation in special education. For each absolute standardized difference above 0.2, add a penalty of $10L$.

The school match in Step 2 was very similar to the one conducted as part of Algorithm 2: the same set of refined covariate balance constraints was used. The new match also excluded the same number of treated schools (4) for better comparability of the two matches (this corresponded to reducing the $\widetilde{\delta}_2$ parameter to 10^6). As a result, the resulting matched sample had 16 sets of paired schools (although these were not the same schools as those selected by Algorithm 1). Combining the full student samples from each school in the paired samples, we obtained a student sample size of 2284, 1106 from schools with the myON intervention and 1178 from schools without (meaning a total of 265 treated students were excluded). However, in

this match, treated students were excluded only because we trimmed four treated schools. We did not trim any students from treated schools that were retained and paired to control students.

4. Analysis of the myON Intervention.

4.1. *Balance Across the Two Matches.* First, we report balance results for the two matches compared to the unmatched data. To assess the quality of the match, we used the standardized difference, which for a given variable is computed by taking the mean difference between matched schools or students and dividing by the pooled standard deviation before matching (Silber et al. 2001; Rosenbaum and Rubin 1985; Cochran and Rubin 1973). We attempted to make all standardized differences less than one-tenth of a standard deviation, which 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 2 contains the results for balance on school level covariates. First, while there are clear differences between treated and control schools in the unmatched data, those discrepancies are not extreme as none of the standardized differences exceed 0.30, however, most of the standardized differences exceed 0.20. In general, treated schools tend to have higher test scores, lower staff turnover, and a lower percentage of nonwhite teachers. Next, balance on school level covariates after matching is identical for both matches, since our matching algorithm under Design 1 and 2 does not differ at the school level and thus produces identical balance results. While the unmatched standard differences are not large, reducing them further proved difficult. We were unable to lower all the standardized differences below the 0.10 benchmark, even after we discarded 4 treated schools in the match. However, overall balance is markedly improved compared to the unmatched data, Figure 1 provides a visual summary of the overall improvement in balance.

Next, we report the balance for the student-level covariates in Table 3. Again, the differences in the unmatched data are all quite small. We now observe differences in the balance statistics, since in one match we did not pair students, and in the other, each student within a matched school is paired with a student. In this data, matching schools only (Design 1) improves student balance modestly. Here, matching students as well as schools (Design 2) produces very similar levels of balance to Design 1.

TABLE 2
Balance at the school level for unmatched data and two matched comparisons. Both means and standardized differences are weighted by the number of students in each school. St-diff is the standardized difference.

	Unmatched –St-diff–	School Only Match –St-diff–	School & Student Match –St-diff–
Composite Test Score	0.21	0.04	0.04
Percent Proficient Reading	0.11	0.07	0.07
Percent Proficient Math	0.20	0.06	0.06
Percentage Student With Free Lunch	-0.10	0.10	0.10
Percentage LEP	-0.29	0.06	0.06
Average Daily Attendance	0.03	0.13	0.13
Percentage of Teachers Beginners	0.28	0.18	0.18
Percentage of Staff Turnover	-0.28	0.17	0.17
Percentage of Nonwhite Teachers	-0.26	0.07	0.07
Title 1 School	-0.11	0.00	0.00
Title 1 Focus School	0.02	0.14	0.14

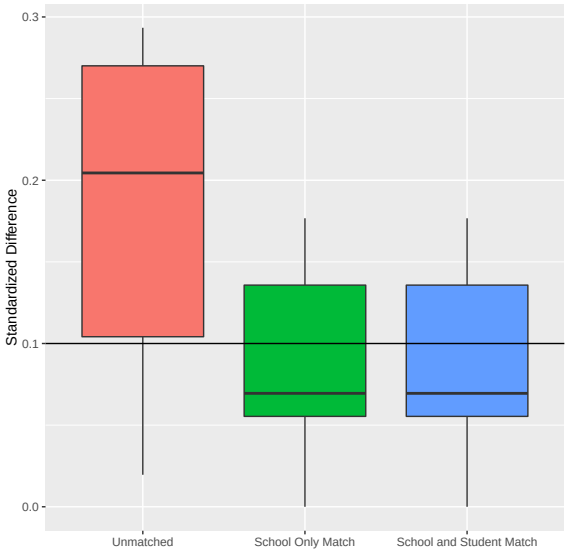

FIG 1. *Boxplots of the distribution of absolute standardized differences for school level covariates. The first boxplot is for the unmatched data, the second for a match that pairs schools only, and the third for a match that pairs both students and schools.*

TABLE 3
Balance on student level covariates. St-diff is the standardized difference.

	Unmatched –St-diff–	School Only Match –St-diff–	School & Student Match –St-diff–
Reading Pretest Score	-0.02	-0.06	-0.02
Math Pretest Score	-0.02	-0.07	-0.07
Male 0/1	-0.09	-0.08	-0.09
Special Education 0/1	0.09	0.13	0.09
Hispanic 0/1	0.02	-0.03	-0.01
African-American 0/1	-0.00	-0.08	-0.06

4.2. *Randomization Inference in Clustered Designs.* In our analysis, we assume that, after matching, treatment assignment is as-if randomly assigned to schools. That is, after matching, it is as if the toss of a fair coin was used to allocate the myON reading program within matched school pairs. The set $\mathbf{\Omega}$ contains the 2^S treatment assignments for all $2S$ clusters: $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{S2})^T$. Under our identification strategy, we assume that the probability of receiving treatment is equal for both schools in each matched pair. If true, the conditional distribution of $\mathbf{Z}$ given that there is exactly one treated unit in each pair equals the randomization distribution, and $\Pr(Z_{sj} = 1) = 1/2$ for each unit j in pair s (see Rosenbaum 2002 for details). However, in an observational study, $\Pr(Z_{sj} = 1) = 1/2$ may not hold for each unit j in pair s due to the presence of an unobserved covariate u_{sji}. We explore this possibility throughthe sensitivity analysis in Section 4.5.

To test Fisher's sharp null hypothesis of no treatment effect, we define a test statistic $T = t(\mathbf{Z}, \mathbf{R})$. If the sharp null hypothesis holds, then $\mathbf{R} = \mathbf{y}_c$ and $T = t(\mathbf{Z}, \mathbf{y}_c)$. We choose T to be a test statistic from Hansen, Rosenbaum and Small (2014). Specifically let q_{sji} be a score or rank given to Y_{sji}, so that under the null hypothesis, the q_{sji} are functions of the y_{Csji} and $\mathbf{x}_{sji}$, and they do not vary with Z_{sk}. The test statistic T is a weighted sum of the mean ranks in the treated school minus the mean ranks in the control school. Formally the test statistic is

$$T = \sum_{s=1}^{S} B_s Q_s$$

where

$$B_s = 2Z_{s1} - 1 = \pm 1, \quad Q_s = \frac{w_s}{n_{s1}} \sum_{i=1}^{n_{s1}} q_{s1i} - \frac{w_s}{n_{s2}} \sum_{i=1}^{n_{s2}} q_{s2i}.$$

and where w_s are weights which are a function of n_{sj}. Hansen, Rosenbaum and Small (2014) show that T is the sum of S independent random variables each taking the value $\pm Q_s$ with probability 1/2, so $E(T) = 0$ and $\text{var}(T) = \sum_{s=1}^{S} Q_s^2$. The central limit theorem implies that as $S \to \infty$, then $T/\sqrt{\text{var}(T)}$ converges in distribution to the standard Normal distribution. Note the framework of Hansen, Rosenbaum and Small (2014) applies directly to either Design 1 or Design 2, since under both designs we assume treatment is applied at the cluster level.

We use two different sets of weights. The first set of weights, $w_s \propto 1$, weight each set of matched pairs equally. The second set of weights are proportional to the total number of students in a matched cluster pair: $w_s \propto n_{s1} + n_{s2}$ or $w_s = (n_{s1} + n_{s2})/\sum_{l=1}^{S}(n_{11} + n_{12})$. These weights allow the treatment effect to vary with cluster size. This would be true if, for example, the effect of the myON reading intervention was perhaps larger in smaller schools. Below we discuss how we incorporate the different weights into the sensitivity analysis.

If we test the hypothesis of a shift effect instead of the hypothesis of no effect, we can apply the method of Hodges and Lehmann (1963) to estimate the effect of being offered the myON reading program. The Hodges and Lehmann (HL) estimate of τ is the value of τ_0 that when subtracted from Y_{sji} makes T as close as possible to its null expectation. Intuitively, the point estimate $\hat{\tau}$ is the value of τ_0 such that T equals 0 when T_{τ_0} is computed from $Y_{sji} - Z_{sj}\tau_0$. If the treatment has a constant additive effect, $Y_{sji} = y_{Csji} + \tau$ then a 95% confidence interval for the additive treatment effect is formed by testing a series of hypotheses $H_0 : \tau = \tau_0$ and retaining the set of values of τ_0 not rejected at the 5% level. Using constant effects is convenient, but this assumption can be relaxed; see Rosenbaum (2003).

4.3. *The Effectiveness of the myON Intervention.* Next, we report the results on the effectiveness of the myON intervention for both matches. The causal estimand for each match is slightly different. For the match that paired both students and schools (Design 2), the estimand pertains to the set of schools and students for whom treatment is marginal. As such, the causal estimand does not apply to all treated students. The school-only match represents a true group-level estimand, as such it represents the effect of the myON intervention on the population that attended a marginal treated school.

Hansen, Rosenbaum and Small (2014) suggest adjusting for baseline student covariates by applying a regression model to the matched data and using the ranks of the residuals when Y_{sji} is regressed on the student-level

covariates. We regressed the outcome, reading test scores recorded after summer school, on student level test scores recorded prior to summer school. We performed the regression analysis via Huber's method of m-estimation. To allow our analysis to be fully transparent, we report results for both matches with and without regression adjustment for baseline student level test scores.

We first test the sharp null hypothesis that the myON intervention is without effect. For the Design 1 match, without regression adjustment with constant weights $w_s \propto 1$, the approximate one-sided p-value is 0.415. Using weights proportional to cluster size, the approximate one-sided p-value is 0.456. The p-values after adjustment are 0.315 and 0.343, respectively. Thus we are unable to reject the null that the myON intervention is completely without effect. For the Design 2 match, If we do not apply regression adjustment and use constant weights $w_s \propto 1$, the approximate one-sided p-value is 0.205. If we use weights proportional to cluster size, the approximate one-sided p-value is 0.338. The p-values for the test of the sharp null for the regression adjusted data are very similar at 0.288 and 0.269 respectively. Again, we are unable to reject the null that the myON intervention is completely without effect.

Next, we report confidence intervals and point estimates. Table 4 contains both point estimates and 95% confidence intervals for both the Design 1 and Design 2 matches, with and without regression adjustment. In the absence of bias from hidden confounders, under Design 1, without adjustment, the point estimate is $\hat{\tau} = 4.7$ with a 95% confidence interval of $[-5.8, 20.5]$, and 1.81 with a 95% confidence interval of $[-4.2, 9.4]$, with adjustment. For Design 2, the point estimate is $\hat{\tau} = 0.745$ with a 95% confidence interval of $[-5, 8]$ without regression adjustment, and $\hat{\tau} = 1.5$ with a 95% confidence interval of $[-4.2, 9.4]$ with regression adjustment. Under Design 1, the role of adjustment via regression is clear as the confidence interval is much narrower. However, under Design 2, matching on students appears to serve a similar role. We next explore the likelihood that bias from a hidden confounder masks a treatment effect.

4.4. *Test of Equivalence and Sensitivity Analysis.* Next, we apply a test of equivalence to test the hypothesis that $\hat{\tau}$ is less than an educationally meaningful effect size. This will allow us to probe the possibility that bias from a hidden confounder is masking an actual treatment effect leaving the analyst to conclude there is no effect when in fact such an effect exists. We can explore this possibility by combining a test of equivalence with a sensitivity analysis (Rosenbaum 2008; Rosenbaum and Silber 2009; Rosenbaum 2010).

TABLE 4
Outcome estimates and confidence intervals for both matches. Estimates reported both with regression adjustment for baseline test scores and without adjustment.

Design 1: School Only Match		
	Unadjusted	Regression Adjusted
Point Estimate	4.74	1.81
95% Confidence Interval	[-5.8, 20.5]	[-4.22, 9.42]
Design 2: School & Student Match		
	Unadjusted	Regression Adjusted
Point Estimate	0.745	1.5
95% Confidence Interval	[-5, 8]	[-4.24, 9.42]

Under a test of equivalence, the null hypothesis asserts $H_{\neq}^{(\iota)} : |\tau| > \iota$ for some specified $\iota > 0$. Rejecting $H_{\neq}^{(\iota)}$ provides a basis for asserting with confidence that $|\tau| < \iota$. $H_{\neq}^{(\iota)}$ is the union of two exclusive hypotheses: $\overleftarrow{H}_0^{(\iota)} : \tau \leq -\iota$ and $\overrightarrow{H}_0^{(\iota)} : \tau \geq \iota$, and $H_{\neq}^{(\iota)}$ is rejected if both $\overleftarrow{H}_0^{(\iota)}$ and $\overrightarrow{H}_0^{(\iota)}$ are rejected (Rosenbaum and Silber 2009). We can apply the two tests without correction for multiple testing since we test two mutually exclusive hypotheses. Thus we can test whether the estimate from our study is different from other possible treatment effects which are represented by ι. With a test of equivalence, it is not possible to demonstrate a total absence of effect, but in the absence of unobserved confounders we may hope to demonstrate that our estimated effect does not exceed ι, by rejecting $H_{\neq}^{(\iota)} : |\tau| > \iota$.

Next, we use a sensitivity analysis to quantify the degree to which a key assumption must be violated in order for our inference to be reversed. We use a model of sensitivity analysis discussed in Rosenbaum (2002, ch. 4), which we describe below. In our study, matching on observed covariates $\mathbf{x}_{sji}$ made schools more similar in their chances of being exposed to the treatment. However, we may have failed to match on an important unobserved covariate u_{sji} such that $\mathbf{x}_{sj} = \mathbf{x}_{sj'} \; \forall \; s, j, j''$, but possibly $\mathbf{u}_{sj} \neq \mathbf{u}_{sj'}$. If true, the probability of being exposed to treatment may not be constant within matched school pairs (and hence within matched student pairs). To explore this possibility, we use a sensitivity analysis that imagines that before matching, school j in pair s had a probability, π_{sj}, of being exposed to the myON intervention. For two matched schools in pair s, say j and j', because they have the same observed covariates $\mathbf{x}_{sj} = \mathbf{x}_{sj'}$ it may be true that $\pi_{sj} = \pi_{sj'}$. However, if these two schools differ in their unobserved covariates, $\mathbf{u}_{sj} \neq \mathbf{u}_{sj'}$, then these two schools may differ in their odds of being exposed to the myON intervention by at most a factor of $\Gamma \geq 1$ such that

$$\frac{1}{\Gamma} \leq \frac{\pi_{sj}/(1-\pi_{sj'})}{\pi_{sj'}/(1-\pi_{sj})} \leq \Gamma, \quad \forall\, s, j, j', \text{ with } \mathbf{x}_{sj} = \mathbf{x}_{sj'}. \tag{1}$$

If $\Gamma = 1$, then $\pi_s = \pi_{s'}$, and the randomization distribution for T is valid. If $\Gamma > 1$, then quantities such as p-values and point estimates are unknown but are bounded by a known interval. Under a test of equivalence, we may be able to reject $H_{\neq}^{(\iota)} : |\tau| > \iota$ if the p-value from the test is less than some threshold, typically 0.05. Rejecting this null allows us to infer that the estimated treatment effect is not as large as ι. We then apply the sensitivity analysis to understand whether this inference is sensitive to biases from nonrandom treatment assignment. In the analysis, we observe at what value of Γ the upper-bound on the p-value exceeds the conventional 0.05 threshold for each test. If this Γ value is relatively large, we can be confident that the test of equivalence is not sensitive to hidden bias from nonrandom treatment assignment. The derivation for a sensitivity analysis appropriate for test statistic T can be found in Hansen, Rosenbaum and Small (2014).

Sensitivity to hidden bias may vary with the choice of weights w_s (Hansen, Rosenbaum and Small 2014). To understand whether different weights lead to different sensitivities to a hidden confounder, we can conduct a different sensitivity analysis for each set of weights and correct these tests using a Bonferroni correction. However, Rosenbaum (2012b) shows that the Bonferroni correction is overly conservative and develops an alternative multiple testing correction based on correlations among the test statistics. We use this correction for multiple testing correction which produces a single corrected p-value for each value of Γ.

4.5. *How Much Bias Would Need to be Present to Mask an Educationally Significant Effect?.* First, we set ι to .20 of a standard deviation, which is considered to be an educationally significant effect size in the relevant literature. We do not present results for the test of equivalence for all four point estimates. We only apply the test of equivalence to the unadjusted point estimates. These are the largest and smallest estimates across both designs, thus the results we report will bracket the tests of equivalence for the adjusted point estimates.

First, we present the results for the unadjusted point estimate in Design 1, which is the largest of the four point estimates. If we assume that there is no hidden bias such that $\Gamma = 1$, and test $\overleftarrow{H}_0^{(\iota)}$, we find that the one-sided p-value from this test is 0.033. We then test $\overrightarrow{H}_0^{(\iota)}$, and we find that the one-sided p-value is 0.11. Therefore, we are unable to reject the null that the

treatment effect we observe in this study is educationally significant. Next, we apply the test of equivalence to the unadjusted point estimate in Design 2, which is the smallest of the four point estimates. We first assume that there is no hidden bias such that $\Gamma = 1$, and we test $\overleftarrow{H}_0^{(\iota)}$ and find that the one-sided p-value from this test is 0.025. We then test $\overrightarrow{H}_0^{(\iota)}$, and we find that the one-sided p-value is 0.034. Therefore, we are able to reject the null that the treatment effect we observe in this study is educationally significant. Is this inference sensitive to bias from a confounder? We find that when Γ is as small as 1.2 the p-value for the test of equivalence is 0.049. Thus if students differed by as much as 20 percent in the odds of being treated that could explain our inference. As such, our study's findings are fairly sensitive to possible bias from a hidden confounder.

5. Summary and Discussion. Here, we developed a new matching algorithm for hierarchical or multilevel data structures. Building on previous work, we follow the strategy of first matching individuals and then, considering these optimal individual level matches, match clusters. However, we use a more standard matching framework based on network flows as opposed to integer programming. Although we cannot target all balance constraints as directly as previous methods did, our algorithm is much faster and can be more easily scaled up to large matching problems without the use of specialized computing techniques such as parallel processing of the matches. We also develop two versions of the algorithm. The first is designed to closely follow the template of a group RCT and does not pair students within schools. The second algorithm pairs both students and schools. We think it is most applicable in contexts like the myON intervention where the treatment only applied to a subset of students within treated schools. Under both algorithms, analysts can choose to trim treated units to improve balance or maintain the common support assumption.

Our application highlights some clear limitations that can arise in clustered observational studies. Here, the pool of controls is fairly small, and as a result, we are unable to produce a match where satisfactory balance is achieved on all covariates. When this occurs, optimal subsetting of the treated group is often the only way to reduce imbalances. When trimming the sample, investigators should take care to communicate to readers how the sample has changed and the population that defines the causal estimand. Finally, we highlight how clustered observational studies often require design choices that are absent when treatment assignment is not clustered. Critically, the choice between either Design 1 or Design 2 alters the estimand, since pairing students will invariably trim the treated sample. Moreover,

these design choices should be made blind to outcomes. Ideally, outcome measures would be merged with the data after the matching is complete (Rubin 2008).

References.

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

Borman, G. D., Benson, J. and Overman, L. T. (2005). Families, schools, and summer learning. *The Elementary School Journal* **106** 131–150.

Borman, G. D. and Dowling, N. M. (2006). Longitudinal achievement effects of multiyear summer school: Evidence from the Teach Baltimore randomized field trial. *Educational Evaluation and Policy Analysis* **28** 25–48.

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

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

Cooper, H., Nye, B., Charlton, K., Lindsay, J. and Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research* **66** 227–268.

Cooper, H., Charlton, K., Valentine, J. C., Muhlenbruck, L. and Borman, G. D. (2000). Making the most of summer school: A meta-analytic and narrative review. *Monographs of the society for research in child development* 1–127.

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

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

Entwisle, D. R. and Alexander, K. L. (1992). Summer setback: Race, poverty, school composition, and mathematics achievement in the first two years of school. *American Sociological Review* 72–84.

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.

Hodges, J. L. and Lehmann, E. L. (1963). Estimates of Location Based on Ranks. *The Annals of Mathematical Statistics* **34** 598-611.

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.

Keele, L. J. and Zubizarreta, J. (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* **In press.**

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York, NY.

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

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).

Page, L. C. and Scott-Clayton, J. (2016). Improving college access in the United States: Barriers and policy responses. *Economics of Education Review* **51** 4–22.

Pimentel, S. D. and Kelz, R. (2016). Optimal Tradeoffs in Matching Designs for Ob-

servational Studies. Unpublished Manuscript.

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.

Quinn, D. M. (2015). Black–White Summer Learning Gaps Interpreting the Variability of Estimates Across Representations. *Educational Evaluation and Policy Analysis* **37** 50–69.

Rambo-Hernandez, K. E. and McCoach, D. B. (2015). High-Achieving and Average Students' Reading Growth: Contrasting School and Summer Trajectories. *The Journal of Educational Research* **108** 112–129.

Rosenbaum, P. R. (1989). Optimal Matching for Observational Studies. *Journal of the American Statistical Association* **84** 1024-1032.

Rosenbaum, P. R. (2002). *Observational Studies*, 2nd ed. Springer, New York, NY.

Rosenbaum, P. R. (2003). Exact Confidence Intervals for Nonconstant Effects by Inverting the Signed Rank Test. *The American Statistician* **57** 132-138.

Rosenbaum, P. R. (2008). Testing hypotheses in order. *Biometrika* **95** 248–252.

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

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

Rosenbaum, P. R. (2012b). Testing One Hypothesis Twice in Observational Studies. *Biometrika* **99** 763–774.

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

Rosenbaum, P. R. and Rubin, D. B. (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.

Rubin, D. B. (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.

Skibbe, L. E., Grimm, K. J., Bowles, R. P. and Morrison, F. J. (2012). Literacy growth in the academic year versus summer from preschool through second grade: Differential effects of schooling across four skills. *Scientific Studies of Reading* **16** 141–165.

Traskin, M. and Small, D. S. (2011). Defining the study population for an observational study to ensure sufficient overlap: a tree approach. *Statistics in Biosciences* **3** 94–118.

Wirt, J., Choy, S., Gruner, A., Sable, J., Tobin, R., Bae, Y., Sexton, J., Stennett, J., Watanabe, S., Zill, N. et al. (2000). *The Condition of Education, 2000.* ERIC, Washington D.C.

Yang, D., Small, D. S., Silber, J. H. and Rosenbaum, P. R. (2012). Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics* **68** 628–636.

Zubizarreta, J. R. (2012). Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure after Surgery. *Journal of the American Statistical*

Association **107** 1360–1371.

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* **in press**.

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

Zubizarreta, J. R., Reinke, C. E., Kelz, R. R., Silber, J. H. and Rosenbaum, P. R. (2012). Matching for several sparse nominal variables in a case-control study of readmission following surgery. *The American Statistician*.

Zvoch, K. and Stevens, J. J. (2015). Identification of Summer School Effects by Comparing the In-and Out-of-School Growth Rates of Struggling Early Readers. *The Elementary School Journal* **115** 433–456.

Department of Statistics
367 Evans Hall
University of California, Berkeley
Berkeley, CA 94720
E-mail: spi@berkeley.edu

School of Education
5918 Wesley W. Posvar Hall
230 South Bouquet Street
Pittsburgh, PA 15260
E-mail: lpage@pitt.edu

Dept of Data and Accountability
Crossroads 1
Cary, NC 27518
E-mail: mlenard@wcpss.net

McCourt School of Public Policy
304 Old North
37th & O St, NW
Washington, D.C. 20057
E-mail: lk681@georgetown.edu