Optimal Multilevel Matching Using Network Flows: An Application to a Summer Reading Intervention*

Samuel D. Pimentel [†] Lindsay Page[‡] Matthew Lenard[§] Luke Keele[¶]

December 4, 2015

Abstract

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. In this paper, 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. While we cannot target balance on specific covariates, our algorithm is quite fast and scales easily to larger data sets. We also consider complications that arise from the common support assumption. In one variant of the algorithm, where we match both schools and students, we change the causal estimand 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 apply both algorithms to an intervention in North Carolina. In the intervention, students in treated schools were exposed to a new reading program during summer school. We find that the intervention does not appear to increase reading test scores, however, in a sensitivity analysis, we find that an unobserved confounder could easily mask a larger treatment effect.

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

*For comments and suggestions, we thank ...

[†]University of Pennsylvania, Philadelphia, PA, Email: spi@wharton.upenn.edu

[‡]University of Pittsburgh, Pittsburgh, PA, Email: lpage@pitt.edu

[§]Wake County Public Schools, Raleigh, NC, Email: mlenard@wcpss.net

[¶]Penn State University, University Park, PA, Email: ljk20@psu.edu.

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 et al. 2005; Cooper et al. 2000; Entwisle and Alexander 1992). Students whose families can substitute school-based with home-based learning resources appear to be affected to a lesser degree than their low-SES counterparts and, as a result, retain learning from the previous school year throughout the summer (Borman and Dowling 2006; Entwisle et al. 2000). 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), 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 an overall loss of 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, by the late 1990s, approximately nine percent of public school students participated in summer school (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 didn't 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. The myON program was used by teachers at eight of the 19 summer school sites. These summer school sites were selected 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. 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 and Chambers 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.

Matching estimators are one method of statistical adjustment 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 individuals instead of to single individuals. Clustered treatments are common in educational settings as treatments are applied or withheld to entire schools rather to individual students or teachers. The myON reading intervention is a treatment of this type, as the reading program was offered to all students in school 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). When clustered treatments are assigned randomly, it may reduce efficiency compared to when treatments are assigned at the individual level, but one avoids biases from nonrandom treatment assignment. However, 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 has a multilevel data structure with observed and unobserved covariates both at the cluster and unit levels. For example, in the myON intervention, we observe student-level covariates such as pretreatment test scores, but also school-level covariates such as the size of the school.

Thus in a clustered observational study to remove overt bias, researchers need to remove treated and control differences in the distributions of covariates both at the cluster and unit levels. That is, we require a statistical strategy to account for the multilevel structure of the data. In this way, we seek to mimic not the standard group RCT design, but instead a blocked group RCT where both clusters and units are matched before treatment is assigned. Standard methods of adjustment for data with multilevel structure include hierarchical regression or methods based on propensity score stratification (Hong and Raudenbush 2006; Arpino and Mealli 2011; Li et al. 2013). Recent work developed a matching algorithm for multilevel data based on integer programming (Keele and Zubizarreta 2015).

Here, we develop an optimal matching method for multilevel data structures based on a network

flow framework. Network flows are a method of optimization 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 lacks the ability to add specific covariate constraints, 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 include optimal subsetting, 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. 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 definig notation and our causal framework. After matching, there are S matched pairs of schools, $s = 1, \ldots, S$, 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 for observed, pretreatment covariates: $\mathbf{x}_{sji}$. A student i in school sj is described by both observed covariates and possibly an unobserved covariate u_{sji}. The set $(\mathbf{x}_{sji}, u_{sji})$ may describe either the student sji or the school sj containing this student. 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 could use the myON software, and y_{Csji} is the test score this same student would exhibit if he or she could not use the myON software. We allow for 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.

For each student, the unobservable effect of treatment is $y_{Tsji} - y_{Csji}$, which is the change in reading test scores caused by use of the myON reading program. 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 with the same notation for $\mathbf{y}_c$, for the vector of potential responses under control. Below we test the sharp null hypothesis of no treatment effect on (y_{Tsji}, y_{Csji}) which stipulates that $H_0 : y_{Tsji} = y_{Csji}$ for all sji (Fisher 1935). This hypothesis asserts that changing the treatment assigned to school sj would leave the response of student sji unchanged.

To identify the causal estimand above, we assume that assignment to Z_{sj} depends on observable

covariates only. Formally, we must assume that

$$\pi_{sji} = Pr(Z_{sj} = 1|y_{Tsji}, y_{Csji}, \mathbf{x}_{sji}, u_{sji}) = Pr(Z_{sj} = 1|\mathbf{x}_{sji}).$$

We also assume that all units have some probability of being treated such that $0 < \pi_{sji} < 1$. The assumption of observable treatment assignment is often referred to as conditional ignorability or selection on observables (Rosenbaum and Rubin 1983; Barnow et al. 1980). If this assumption holds, potential outcomes will be 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 may not hold if for some students the probability of being exposed to the myON intervention is zero or if for some treated students the probability of being exposed to the myON intervention exceeds the probability that some control students were exposed to the treatment. 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.

As noted in Keele and Zubizarreta (2015), a multilevel match may need to trim both clusters and units, here schools and students, to maintain the common support assumption. 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 myON intervention is marginal. 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. Since treatment status was defined through proximity to summer school sites with the technical capacity to support the myON intervention, we do not feel that our study population represents any natural larger population. As such, we allow in our match for the removal of both treated schools and students to maintain overlap in the treated and control distributions. However, for the purposes of comparison, we also include one additional match that does not trim students, but allows for trimming of treated schools. 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.

3 Multilevel Matching

3.1 Review: Matching based on integer programming and network flows

The goal with matching methods in an observational study is to created a set of comparable treated and control units. We seek to replicate the result from randomization: treated and control units that are similar in terms of observed covariates. A large number of algorithms exist to perform matching of this type.

One newer type of matching algorithm is based on mixed integer programming (Zubizarreta 2012) and has been adapted to multilevel data structures (Keele and Zubizarreta 2015). The key advantage of such matching algorithms is that they allow the analyst to target explicit levels of balance for mean differences across treated and control units. These methods also allow for explicit balancing of statistics such as the Kolmogorov-Smirnov (KS) statistic. These methods allow the analysts to target specific levels of covariate balance directly. The drawback to such

methods is that they are difficult to scale for larger data sets and the computational time can be extensive for even relatively small problems.

Many other matching algorithms use network flows to balance covariates by minimizing the total sum of distances between treated and control units. The difficulty with methods based on network flows is that covariate balance is an indirect product of this minimization, and in practice, an iterative process is needed to achieve an acceptable level of balance. The advantage to using network flows is that the algorithms are fast and can be applied to very large data sets with fewer difficulties. Next, we develop a multilevel matching algorithm based on network flows.

3.2 Multilevel matching based on network flows

In multilevel matching, our goal is to create S matched pairs of schools and, for each such pair, $m_s \leq \min(n_{s1}, n_{s2})$ matched pairs of students (one from each school) such that:

1. School-level covariates are balanced across all schools in the matched sample.
2. Student-level covariates are balanced within each school pair.
3. As many students are included in the final match as possible under these constraints.

One way to attempt construction of such a match would be to start by pairing schools based on school covariates, then matching students within schools. However, this approach is not guaranteed to produce an optimal match according to our criteria. Whenever there are multiple school-level matches that meet school-covariate balance constraints (which will generally be the case), it is not clear a priori which of these configurations will produce the student-level match with the largest sample size, and the school match selected may be suboptimal. This problem can be resolved by first conducting student-level matches for every possible school pair, then incorporating this information into the school-level match (Keele and Zubizarreta 2015).

Our algorithm follows this approach. 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 and imposing refined covariate balance constraints on school-level covariates. This ensures that we choose the lowest-score school pairs such that school-level covariates are balanced as well as possible.

One important issue that arises in both the student- and school-level matches is the possible need to trim units from the treated group to maintain the common support assumption. Unless the samples of students in control schools are uniformly substantially larger than the samples in 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, pair matching is either infeasible or unsuccessful in removing bias unless some treated units are trimmed or excluded from the match. Even when there are more control units than treated units, it may be desirable to exclude some treated units to improve covariate balance. 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. To allow for trimming of treated units in a principled manner, we use and extend optimal subset matching.

3.3 Optimal Subset Matching and Extension

Optimal matching based on network flows may result in levels of covariate balance that does not sufficiently remove overt biases. When this occurs a number of possible refinements are possible. One such refinement is optimal subset matching. Another refinement is refined covariate balance. We review each of these methods below and then develop a combination of the methods for the

purpose of multilevel matching.

Optimal subset matching is a network flow algorithm for pair matching which allows the match to exclude treated units but 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. 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.

Refined balance constraints can also be used to increase balance in observables (Pimentel et al. 2015). 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 not small relative to the number of observations. 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. Matching with a refined covariate balance constraint produces the lowest-cost match among matches where the covariate in question has been balanced as closely as possible. When refined covariate balance constraints are imposed on several covariates at once, they 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.

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. However, their framework does not permit treated units to be excluded from the match

in any case. We adapt matching with refined covariate balance to cases in which exclusion of treated units is desirable. Specifically, we introduce the familiar penalty parameter $\widetilde{\delta}$, which represents 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 formal proof of this result and a technical description of the alterations to the algorithm of Pimentel et al. (2015), see the Appendix.

3.4 A General Algorithm

We summarize our approach to multilevel matching with network flows in the following algorithm (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 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, 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 the above algorithm 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. As we noted above, such a match alters the estimand such that it only applies to a subset of marginal students at marginal schools.

We may wish to preserve the causal estimand as a school level effect. By a small modification of our matching algorithm, we can construct a matched sample that maintains the group level estimand. We do this by pairing schools, but not pairing students within schools. However, we can use student level information in the school level match. In particular, we define a new algorithm (Algorithm 2) by replacing step 1 of Algorithm 1 with the following procedure.

1. For each possible combination of one treated school i and one control school j:
    - Let m^* be the harmonic mean of the treated and control group sizes
    - Assign a score ℓ_{ij} to this match using a pre-specified scoring rule which depends on m^* and the balance of the two groups on student covariates, and store it in a matrix.

As such, we still use student level balance information in the pairing of schools, but we do not pair the students directly after schools are matched.

3.5 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. We created one matched sample using Algorithm 1, pairing both schools and students within schools. In addition, we created a second matched sample using Algorithm 2, which paired schools using student level balance information but retained all students within the matched schools.

For the match of both schools and students, we first created a robust Mahalanobis distance

matrix 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. Six layers of refined covariate balance constraints were added, each layer an interaction of categorical school covariates (e.g. Title 1 status) and appropriate quantiles of continuous school covariates (e.g. proportion of new teachers, proportion of English-proficient students, etc.). We set the penalty $\widetilde{\delta}_2$ to 3×10^7, which resulted in 3 treated schools being excluded for a final matched sample of 17 school pairs. Combining the first-stage school-to-school matches corresponding to the matched school pairs, we obtained an overall matched sample of 2,086 students, 1,043 from schools with the myON

intervention and 1,043 without (meaning a total of 328 treated students were trimmed from the treated sample).

For the second match using Algorithm 2, a similar scoring function was 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 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 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 1: the same set of refined covariate balance constraints was used. The $\widetilde{\delta}_2$ parameter was lowered to 2.5×10^7 to ensure that the same number of treated schools (3) were excluded for better comparability of the two matches. As a result, the resulting matched sample had 17 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 of total size 2268, 1181 from schools with the myON intervention and 1087 from schools without (meaning a total of 190 treated students were excluded). However, in this match, treated students were excluded only because we trimmed 3 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

Next, we report balance results. To assess the quality of the match, we used the standardized distance, which for a given variable is computed by taking the mean difference between matched patients 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 0.10 or less than one-tenth of a standard deviation, which is often considered an acceptable discrepancy, since we might expectation discrepancies of this size from a randomized experiment (Silber et al. 2001; Rosenbaum and Rubin 1985; Cochran and Rubin 1973; Rosenbaum 2010). We see that treated schools tend to have higher test scores but lower staff turnout over and a lower percentage of nonwhite teachers.

We plot the distribution of standardized differences for each match in Figure 1. The school-only match and the match where we trim students and schools have similar distributions of standardized differences. In neither match are we able to reduce all imbalances to less than one-tenth of standard deviation, but we are able to reduce the imbalance in school-level covariates significantly compared to the unmatched data. Given the small number of possible control schools, we found it impossible to produce a match with lower levels of overt bias. Table 1 contains more specific balance statistics. In particular, we find there remain imbalances larger than 0.10 for a measure of composite test scores, percentage of staff turnover, average daily attendance, and percent proficient in mathematics.

Next, we report the balance for the student-level covariates in Table 2. For the student-level covariates, even in the unmatched data, the imbalances are quite modest as none exceed 0.10. Again, the two matches are very similar, with neither producing uniformly better balance on all variables than the other.

Table 1: 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	0.17	0.17
Percent Proficient Reading	0.11	0.12	0.09
Percent Proficient Math	0.20	0.15	0.14
Percentage Students Disadvantaged	-0.10	-0.09	-0.10
Percentage Limited English	-0.29	-0.09	-0.14
Average Daily Attendance	0.03	0.13	0.18
Percentage of Teachers Beginners	0.28	0.10	0.11
Percentage of Staff Turnover	-0.28	-0.20	-0.18
Percentage of Nonwhite Teachers	-0.26	-0.07	-0.06
Title 1 School	-0.11	0.00	0.00
Title 1 Focus School	0.02	0.00	0.13
Composite Test Score Upper Quantile 0/1	0.21	0.13	0.13
Proficient Reading Upper Quantile 0/1	0.21	0.27	0.27
Proficient Math Upper Quantile 0/1	0.18	0.24	0.24
% Teachers Beginners Mid Quantile 0/1	0.03	-0.12	-0.12
% Teachers Beginners Upper Quantile 0/1	0.26	0.00	0.00
% Teachers Beginners Lower Mid Quantile 0/1	-0.19	0.00	0.00
% Teachers Beginners Lowest Quantile 0/1	-0.17	-0.12	-0.12
% Turnover Upper Quantile 0/1	-0.07	0.12	0.12
% Turnover Lower Quantile 0/1	0.09	0.12	0.12
% Teachers Nonwhite Upper Quantile 0/1	-0.10	0.00	0.00
% LEP 0/1	-0.20	0.00	0.00
Average Daily Attendance Upper Quantile 0/1	-0.08	0.00	0.00

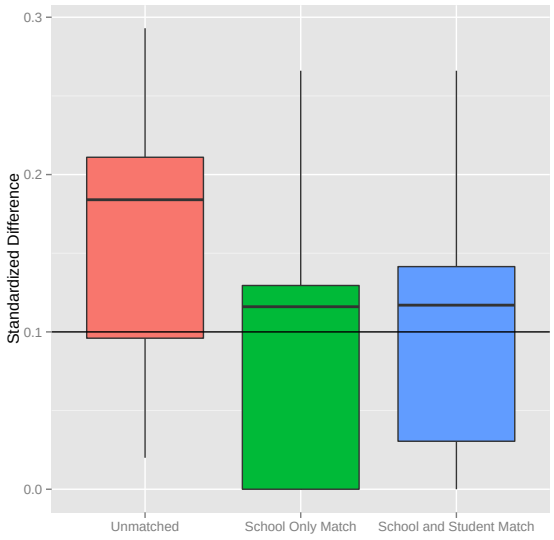

Figure 1: Distribution of absolute standardized differences for unmatched data and two matches: one that only pairs students and one that pairs both students and schools.

Table 2: Balance on student level covariates. St-diff is the absolute standardized difference.

	Unmatched –St-diff–	School Only Match –St-diff–	School & Student Match –St-diff–
Reading Pretest Score	0.02	0.03	0.05
Math Pretest Score	0.02	0.08	0.08
Male 0/1	0.09	0.05	0.03
Special Education 0/1	0.09	0.00	0.00
Hispanic 0/1	0.02	0.03	0.01
African-American 0/1	0.00	0.02	0.04

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 it may not be true $\Pr(Z_{sj} = 1) = 1/2$ for each unit j in pair s due to an unobserved covariate u_{sji}. We explore this possibility through a sensitivity analysis described below.

To test Fisher's sharp null hypothesis of no treatment effect, we define T a test statistic which is a function of $\mathbf{Z}$ and $\mathbf{R}$ where $T = t(\mathbf{Z}, \mathbf{R})$. If the the sharp null hypothesis holds, then $\mathbf{R} = \mathbf{y}_c$ and $T = t(\mathbf{Z}, \mathbf{y}_c)$. If the model for treatment assignment above holds, the randomization distribution for T is

$$\text{Pr}\{t(\mathbf{Z}, \mathbf{R}) \geq w | y_{Tsji}, y_{Csji}, \mathbf{x}_{sji}, u_{sji}, \mathbf{\Omega}\} = \text{Pr}\{t(\mathbf{Z}, \mathbf{y}_c) \geq w | y_{Tsji}, y_{Csji}, \mathbf{x}_{sji}, u_{sji}, \mathbf{\Omega}\}$$

since $\mathbf{y}_c$ is fixed by conditioning on $y_{Tsji}, y_{Csji}, \mathbf{x}_{sji}, u_{sji}$ and $\text{Pr}(\mathbf{Z} = Z_{sj} | y_{Tsji}, y_{Csji}, \mathbf{x}_{sji}, u_{sji}, \mathbf{\Omega}) = 1/|\mathbf{\Omega}|$.

We define T as a test statistic from Hansen et al. (2014). For this test statistic, q_{sji} is 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}. To make q_{sji} resistant to outliers, we use the ranks of the residuals when Y_{sji} is regressed on the student-level covariates using Huber's method of m-estimation Small et al. (2008). We regressed the outcome, reading test scores recorded after summer school on student level test scores recorded prior to summer school. 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

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

where w_s are weights which are a function of n_{sj}. Hansen et al. (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 $\mathrm{var}(T) = \sum_{s=1}^{S} Q_s^2$. The central limit theorem implies that as $S \to \infty$, then $T/\sqrt{\mathrm{var}(T)}$ converges in distribution to the standard Normal distribution.

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 weight we use 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 $Y_{sji} - Z_{sj}\tau_0$. Using constant effects is convenient, but this assumption can be relaxed; see Rosenbaum (2003). If the treatment has an 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.

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 first match that paired both students and schools, 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 treated population which attended a marginal school.

We first test the sharp null hypothesis that the myON intervention is without effect. For the

school and student match, if we use constant weights $w_s \propto 1$, the approximate one-sided p-value is 0.136. Using weights proportional to cluster size, the approximate one-sided p-value is 0.137. Thus we are unable to reject the null that the myON intervention is completely without effect. For the school only match, with constant weights $w_s \propto 1$, the approximate one-sided p-value is 0.289. Using weights proportional to cluster size, the approximate one-sided p-value is 0.245. Again, we are unable to reject the null that the myON intervention is completely without effect. In the absence of bias from hidden confounders, the point estimate is $\hat{\tau} = 2.61$ with a 95% confidence interval of -1.22 and 8.52 for the school and student match, and the point estimate is $\hat{\tau} = 1.60$ with a 95% confidence interval of -4.04 and 9.61, for the school-only match. 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).

Under a test of equivalence, the null hypothesis asserts $H_{\neq}^{(\delta)} : |\tau| > \delta$ for some specified $\delta > 0$. We set δ to .20 of a standard deviation, which is considered to be an educationally significant effect size in the education literature. Rejecting $H_{\neq}^{(\delta)}$ provides a basis for asserting with confidence that $|\tau| < \delta$. $H_{\neq}^{(\delta)}$ is the union of two exclusive hypotheses: $\overleftarrow{H}_0^{(\delta)} : \tau \leq -\delta$ and $\overrightarrow{H}_0^{(\delta)} : \tau \geq \delta$, and $H_{\neq}^{(\delta)}$ is rejected if both $\overleftarrow{H}_0^{(\delta)}$ and $\overrightarrow{H}_0^{(\delta)}$ are rejected (Rosenbaum and Silber 2009). We can apply the two tests without correction for multiple testing since we test two mutually exclusive hypotheses. Thus we can test whether the estimate from our study is different from other possible treatment effects which are represented by δ. With a test of equivalence, it is not possible to demonstrate a total absence of effect, but if this were a randomized trial we could safely test that

our estimated effect is not as large as δ. That is we may be able to reject $H_{\neq}^{(\delta)} : |\tau| > \delta$. In an observational study, however, it may be the case that we reject the null hypothesis of equivalence due to hidden confounding. However, using a sensitivity analysis we may find evidence that the test of equivalence is insensitive to biases from nonrandom treatment assignment.

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 students 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 $\mathbf{x}_{sji} = \mathbf{x}_{sji'} \ \forall \ s, j, i, i'$, but possibly $u_{sji} \neq u_{sji'}$. If true, the probability of being exposed to treatment may not be constant within matched pairs. To explore this possibility, we use a sensitivity analysis that imagines that before matching, student i in pair s had a probability, π_s, of being exposed to the myON intervention. For two matched students in pair s, say i and i', because they have the same observed covariates $\mathbf{x}_{sji} = \mathbf{x}_{sji'}$ it may be true that $\pi_s = \pi_{s'}$. However, if these two students differ in an unobserved covariate, $u_{sji} \neq u_{sji'}$, then these two students may differ in their odds of being exposed to the myON intervention by at most a factor of $\Gamma \geq 1$ such that

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

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 $H_{\neq}^{(\delta)} : |\tau| > \delta$ if the p-value from the test is less than some threshold, typically 0.05. Rejecting this null allows us to infer that the estimate treatment effect is not as large as δ. 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 the sensitivity analysis as applied to our test statistic T can be found in Hansen et al. (2014).

Sensitivity to hidden bias may vary with the choice of weights w_s (Hansen et al. 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. 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?

We present results for the test of equivalence for both matches. We begin with the results from the school and student match. We first assume that there is no hidden bias such that $\Gamma = 1$, and we test $\overleftarrow{H}_0^{(\delta)}$ and find that the one-sided p-value from this test is 0.011. We then test $\overrightarrow{H}_0^{(\delta)}$ and we find that the one-sided p-value is 0.027. Therefore, we are able to reject the null that the treatment effect we observe in this study is equivalent to an educationally significant effect. Is this inference sensitive to bias from a confounder? We find that when Γ is as small as 1.30 the p-value for the test of equivalence is 0.049. Thus if students differed by as much as 30 percent in the odds of being treated that could explain our inference.

Next, we present the results from the school only match. If we assume that there is no hidden bias such that $\Gamma = 1$, and test $\overleftarrow{H}_0^{(\delta)}$. we find that the one-sided p-value from this test is 0.014. We then test $\overrightarrow{H}_0^{(\delta)}$, and we find that the one-sided p-value is 0.035. Therefore, we are able to reject the null that the treatment effect we observe in this study is equivalent to an educationally significant effect. Again, we use a sensitivity test to query whether this this inference sensitive

to bias from a confounder? We find that when Γ is as small as 1.16 the p-value for the test of equivalence is 0.049. Thus if students differed by as much as 16 percent in the odds of being treated that could explain our inference. For both matches, the treatment effect estimate is statistically smaller than an educationally significant effect. The conclusions from the test of equivalence are more resistant to hidden bias in the school only match. However, in both cases, the results are fairly sensitive to the possibility that our results could be explained by 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. As such, we cannot target specific balance constaints. However, 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. One that seeks to maintain the common support assumption, and a second that seeks to maintain a group-level estimand. The choice between algorithm will depend on the context and the extent that the common support assumption is violated.

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

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. and Chambers, S. P. (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., 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.

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.

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.

Entwisle, D. R., Alexander, K. L., and Olson, L. S. (2000), "Summer learning and home environment," *A notion at risk: Preserving public education as an engine for social mobility*, 9–30.

Fisher, R. A. (1935), *The Design of Experiments*, London: Oliver and Boyd.

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. (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. (2015), "Optimal Multilevel Matching in Clustered Observational Studies: A Case Study of the School Voucher System in Chile," Unpublished Manuscript.

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).

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.

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

— (2003), "Exact Confidence Intervals for Nonconstant Effects by Inverting the Signed Rank Test," *The American Statistician*, 57, 132–138.

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

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

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

— (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.

— (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.

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.

Small, D. S., Have, T. R. T., and Rosenbaum, P. R. (2008), "Randomization Inference in a Group–Randomized Trial of Treatments for Depression: Covariate Adjustment, Noncompliance, and Quantile Effects," *Journal of the American Statistical Association*, 103, 271–279.

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.*, Washington D.C.: ERIC.

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.

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.

6 Appendix

Here we characterize the optimality of matches with refined balance that may also exclude some treated units. To do so we consider the matching problem of Pimentel et al. (2015) and the associated notation: treated units $\mathcal{T} = \{\tau_1, \ldots, \tau_T\}$, controls $\mathcal{C} = \{\kappa_1, \ldots, \kappa_C\}$, a set $\mathcal{A}$ of allowed pairs (τ_t, κ_c) with costs δ_{tc}, a series of nested refined balance covariates $\nu_1, \ldots, \nu_K$ each mapping each treated or control unit to a category in the set $\{\lambda_{k1}, \ldots, \lambda_{kL_k}\}$. Consider the following additions and modifications to their framework:

Definition 6.1. *Acceptable 1-to-1 match with cardinality T': a subset $\mathcal{M} \subset \mathcal{A}$ such that $|\mathcal{M}| = T'$ and each element in $\mathcal{T} \cup \mathcal{C}$ appears in at most one pair.*

For an acceptable 1-to-1 match with cardinality T' define:

$$\beta_{k\ell}(\mathcal{M}) = |\{(\tau_t, \kappa_c) \in \mathcal{M} : \nu_k(\tau_t) = \lambda_{k\ell}\}| - |\{(\tau_t, \kappa_c) \in \mathcal{M} : \nu_k(\kappa_c) = \lambda_{k\ell}\}|$$

Definition 6.2. *An acceptable 1-to-1 match $\mathcal{M}$ with cardinality T' has refined balance if*

1. *It minimizes $\sum_{\ell=1}^{L_1} |\beta_{1\ell}(\mathcal{M}')|$ among 1-to-1 acceptable matches $\mathcal{M}$ with cardinality T'.*
2. *It minimizes $\sum_{\ell=1}^{L_2} |\beta_{2\ell}(\mathcal{M}')|$ among 1-to-1 acceptable matches $\mathcal{M}$ with cardinality T' that also satisfy 1.*

   $\vdots$

K. *It minimizes $\sum_{\ell=1}^{L_2} |\beta_{2\ell}(\mathcal{M}')|$ among 1-to-1 acceptable matches $\mathcal{M}$ with cardinality T' that also satisfy 1, $\ldots, K-1$.*

Definition 6.3. *An acceptable 1-to-1 match $\mathcal{M}$ with cardinality T' has optimal refined balance if it has refined balance and minimizes*

$$\sum_{(\tau_t, \kappa_c) \in \mathcal{M}'} \delta_{tc}$$

among 1-to-1 acceptable matches $\mathcal{M}$ with cardinality T' and refined balance.

Notice that when $T = T'$ these definitions are equivalent to those in Pimentel et al. (2015). In addition to changing definitions to allow for matches that exclude treated units, we must alter the network optimization algorithm. To review, the network (for 1-to-1 matching) includes treated nodes $\tau_1, \ldots, \tau_T$ with supply one each; control nodes $\kappa_1, \ldots, \kappa_C$, and three nodes $(\lambda_{k\ell}, \lambda'_{k\ell}, \lambda''_{k\ell})$ for each category $\ell = 1, \ldots, L_k$ of each balance level $k = 1, \ldots, K$ and , all with supply 0; and one sink node ω with demand T. The network also includes an edge for each pair $(\tau_t, \kappa_c) \in \mathcal{A}$ (with cost δ_{tc} and capacity 1), and edges connecting the "triangle" of nodes for each fine balance level: $(\lambda_{k\ell}, \lambda'_{k\ell})$ with infinite capacity and cost Υ^{K-k+1} (where Υ is a network-wide penalty parameter), $(\lambda'_{k\ell}, \lambda''_{k\ell})$ with infinite capacity and cost zero, and $(\lambda_{k\ell}, \lambda''_{k\ell})$ with capacity $d_{k\ell} = |\{\tau_t \in \mathcal{T} : \nu_k(\tau_t) = \lambda_{k\ell}\}|$ and cost zero. The fine balance "triangles" with $k > 1$ and $\ell \in 1, \ldots, L_k$ are connected to each other via edges $(\lambda''_{k\ell}, \lambda_{(k-1)\ell'})$ (with infinite capacity and zero cost) where $\lambda_{(k-1)\ell'}$ is the potentially coarser category in level $k-1$ in which the category $\lambda_{k\ell}$ nests. Finally, each control node κ_c is connected to the level-K balance node $\lambda_{K\ell}$ such that $\nu_K(\kappa_c) = \lambda_{K\ell}$ and each level-1 balance node $\lambda''_{1\ell}$ is connected to the sink ω, all by edges of zero cost and infinite capacity.

We modify this network by adding T additional edges, each connecting a treated unit τ_t to the associated category of covariate ν_K, $\lambda_{K\ell} = \nu_K(\tau_t)$. These edges have capacity 1 and cost $\widetilde{\delta}$. Let $\mathcal{N}$ be the full node set of the resulting network, and let $\mathcal{E}$ be the full edge set. We can describe flows through this network via functions $f : \mathcal{E} \longrightarrow \mathbb{Z}^T \cup \{0\}$ that assign some flow to each edge such that edge capacities and node supplies/demands are all satisfied. Any flow f also has a cost $\mathfrak{C}(f)$ computed by multiplying the flow at each edge by its cost and adding them all up. This notation allows us to state the main theorem.

Theorem 6.1. *Consider any optimal network flow in the network just defined such that exactly T' of the edges in $\mathcal{A}$ have nonzero flow. Then there is a unique acceptable 1-to-1 match of cardinality T' associated with this flow, and for sufficiently large Υ this match has optimal refined balance*

among matches with cardinality T'.

Before proving the theorem we present and prove a lemma.

Lemma 6.2. *Every flow f in the network is uniquely associated with an acceptable 1-to-1 match $\mathcal{M}(f)$ of some cardinality (possibly 0). Furthermore, for any acceptable 1-to-1 match $\mathcal{M}$ of cardinality T', there exists an associated flow $f(\mathcal{M})$ in the network with $\mathcal{M}(f(\mathcal{M})) = \mathcal{M}$ and total edge cost*

$$\mathfrak{C}(f(\mathcal{M})) = \sum_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc} + \sum_{k=1}^{K} \Upsilon^{K-k+1} \sum_{\ell=1}^{L_k} |\beta_{k\ell}|/2 + (T - T')\widetilde{\delta}.$$

In addition, no other flow that is also associated with $\mathcal{M}$ has a strictly lower cost.

Proof. Consider any flow in the network. Define $\mathcal{M}(f) = \{(\tau_t, \kappa_c) \in \mathcal{A} : f(\tau_t, \kappa_c) = 1\}$ and let $T' = |\mathcal{M}(f)|$. This defines a unique 1-to-1 acceptable match of cardinality T'.

Now consider any acceptable 1-to-1 match with cardinality T'. Set the flow across each edge (τ_t, κ_c) in $\mathcal{M}$ to 1 (as well as the flow in the subsequent edges $(\kappa_c, \nu_K(\kappa_c))$). For all treated units not in any pair in $\mathcal{M}$, set the flow across edge $(\tau_t, \nu_K(\tau_t))$ to 1. Since there are exactly $T - T'$ treated units not included in any pair in $\mathcal{M}$, the cost of the flow across these edges is

$$\sum_{(\tau_t,\kappa_c)\in\mathcal{M}} \delta_{tc} + (T - T')\widetilde{\delta}$$

Now consider the refined balance nodes. Define

$$d'_{k\ell} = |\{(\tau_t, \kappa_c) \in \mathcal{M} : \nu_k(\tau_t) = \lambda_{k\ell}\}|$$

The size of the flow entering any node $\lambda_{K\ell}$ is the sum of the flow from the control units $|\{(\tau_t, \kappa_c) \in \mathcal{M} : \nu_K(\kappa_c) = \lambda_{K\ell}\}|$ and the flow directly from the excluded treated units $d_{K\ell} - d'_{K\ell}$. More

generally, at any balance level $k < K$ the flow entering node $\lambda_{k\ell}$ is

$$|\{(\tau_t, \kappa_c) \in \mathcal{M} : \nu_K(\kappa_c) = \lambda_{k\ell}\}| + d_{k\ell} - d'_{k\ell}$$

since $d_{k\ell} = \sum_{\ell \in \Lambda} d_{(k+1)\ell}$ and $d'_{k\ell} = \sum_{\ell \in \Lambda} d'_{(k+1)\ell}$ where $\Lambda \subset \{1, \ldots, L_{k+1}\}$ is the set of categories of ν_{k+1} nested within $\lambda_{k\ell}$. The $d_{k\ell} - d'_{k\ell}$ portion of flow can always be routed across edge $(\lambda_{K\ell}, \lambda''_{K\ell})$ which has capacity $d_{K\ell}$. If we then send as much of the remaining flow over this edge (which has zero cost) as well, the remaining portion of flow that must pass through the bypass edge $(\lambda_{K\ell}, \lambda'_{K\ell})$ is

$$\max\{0, \quad |\{(\tau_t, \kappa_c) \in \mathcal{M} : \nu_k(\kappa_c) = \lambda_{k\ell}\}| - d'_{k\ell}\} = \max\{0, \quad \beta_{k\ell}\}$$

The cost of the flow is the number just given multiplied by the penalty Υ^{K-k+1}. Adding up over all categories $\lambda_{k\ell}$ and all levels k, the total cost of flow through the balance nodes is

$$\sum_{k=1}^{K} \Upsilon^{K-k+1} \sum_{\ell=1}^{L_k} |\beta_{k\ell}|/2$$

Therefore the flow defined (call it f) has the desired cost. Now consider any other flow $f' \neq f$ associated with $\mathcal{M}$. For any edge $e = (\tau_t, \kappa_c)$ or $e = (\tau_t, \nu_K(\tau_t))$, $f(e) = f'(e)$ or the two flows could not be associated with the same match $\mathcal{M}$. Therefore the only way the two flows can differ is in the allocation of flow within edge pairs $(\lambda_{k\ell}, \lambda''_{k\ell})$ and $(\lambda_{k\ell}, \lambda'_{k\ell})$. But since f already sends as much flow as possible across the zero-cost edges $(\lambda_{k\ell}, \lambda''_{k\ell})$, f' can never send more flow than f across a zero-cost edge and must send strictly more flow than f across some positive-cost edge $(\lambda_{k\ell}, \lambda'_{k\ell})$. Therefore $\mathfrak{C}(f) < \mathfrak{C}(f')$ and $f = f(\mathcal{M})$ is the optimal flow associated with match $\mathcal{M}$. □

Proof of Theorem 6.1. The existence of a unique acceptable 1-to-1 match of cardinality T' for any flow follows from Lemma 6.2. Let $\mathcal{M} = \mathcal{M}(f)$. Suppose $\Upsilon > T(K + \max_{i,j}\{\delta_{ij}, \widetilde{\delta}\})$. Consider refined balance covariate ν_k. The cost of all the flow in the network before it reaches

level k is

$$\begin{aligned}\sum_{(\tau_t,\kappa_c)\in\mathcal{M}'}\delta_{tc}+\sum_{i=k+1}^{K}\Upsilon^{K-i+1}\sum_{\ell=1}^{L_i}|\beta_{i\ell}(\mathcal{M}')|/2+(T-T')\widetilde{\delta} &< \sum_{(\tau_t,\kappa_c)\in\mathcal{M}'}\delta_{tc}+(T-T')\widetilde{\delta}+\sum_{i=k+1}^{K}\Upsilon^{K-k}T \\ &\leq T\max_{i,j}\{\delta_{ij},\widetilde{\delta}\}+TK\Upsilon^{K-k} \\ &< \Upsilon^{K-k}T(K+\max_{i,j}\{\delta_{ij},\widetilde{\delta}\}) \\ &< \Upsilon^{K-k+1} \qquad (2)\end{aligned}$$

This means that the following facts hold:

1. $\mathcal{M}$ minimizes $\sum_{\ell=1}^{L_1}|\beta_{1\ell}(\mathcal{M}')|$ among 1-to-1 acceptable matches $\mathcal{M}'$ with cardinality T'.
2. $\mathcal{M}$ minimizes $\sum_{\ell=1}^{L_2}|\beta_{2\ell}(\mathcal{M}')|$ among 1-to-1 acceptable matches $\mathcal{M}'$ with cardinality T' that also satisfy item 1.

   $\vdots$

K. $\mathcal{M}$ minimizes $\sum_{\ell=1}^{L_K}|\beta_{K\ell}(\mathcal{M}')|$ among 1-to-1 acceptable matches $\mathcal{M}'$ with cardinality T' that also satisfy items 1,2, $\ldots, K-1$.

$K+1$. $\mathcal{M}$ minimizes $\sum_{(\tau_t,\kappa_c)\in\mathcal{M}'}\delta_{tc}+(T-T')\widetilde{\delta}$ among 1-to-1 acceptable matches $\mathcal{M}'$ with cardinality T' that also satisfy items 1,2, $\ldots, K$.

To see this, suppose one of the first K facts fails, i.e. some balance variable is not optimally balanced conditional on previous levels. Then expression (2) tells us that we can strictly improve the cost of $f=f(\mathcal{M})$ by balancing this variable, since the gain by improving balance at this level will outweigh any change in the total cost coming from subsequent level. Since f is the optimal flow in the network, this cannot be the case. This proves that $\mathcal{M}$ satisfies refined balance.

Now suppose that fact $K+1$ fails. Then there must be some other match $\mathcal{M}'$ of cardinality T' that satisfies refined balance but has lower pairwise distance costs $\sum_{(\tau_t,\kappa_c)\in\mathcal{M}'}\delta_{tc}$. But since $\mathfrak{C}(f(\mathcal{M}))-\mathfrak{C}(f(\mathcal{M}'))$ is equal to the difference in pairwise distance costs between the two matches, flow $f(\mathcal{M}')$ must have a lower cost than flow $f=f(\mathcal{M})$, a contradiction since f is

the optimal flow. Therefore all the facts must hold and $\mathcal{M}$ must satisfy optimal refined balance among acceptable 1-to-1 matches of cardinality T'. □