Building Representative Matched Samples with Multi-valued Treatments in Large Observational Studies*

Magdalena Bennett†   Juan Pablo Vielma‡   José R. Zubizarreta§

Abstract

In this paper, we present a new way of matching in observational studies that overcomes three limitations of existing matching approaches. First, it directly balances covariates with multi-valued treatments without requiring the generalized propensity score. Second, it builds self-weighted matched samples that are representative of a target population by design. Third, it can handle large data sets, with hundreds of thousands of observations, in a couple of minutes. The key insights of this new approach to matching are balancing the treatment groups relative to a target population and positing a linear-sized mixed integer formulation of the matching problem. We formally show that this formulation is more effective than alternative quadratic-sized formulations, as its reduction in size does not affect its strength from the standpoint of its linear programming relaxation. We also show that this formulation can be used for matching with distributional covariate balance in polynomial time under certain assumptions on the covariates and that it can handle large data sets in practice even when the assumptions are not satisfied. This algorithmic characterization is key to handle large data sets. We illustrate this new approach to matching in both a simulation study and an observational study of the impact of an earthquake on educational attainment. After matching, the results can be visualized with simple and transparent graphical displays: while increasing levels of exposure to the earthquake have a negative impact on school attendance, there is no effect on college admission test scores.

Keywords: Causal Inference; Multi-valued Treatment; Observational Studies; Optimal Matching; Propensity Score; Representative Study.

†Department of Education Policy and Social Analysis, Teachers College at Columbia University; email: mb3863@tc.columbia.edu.

‡Operations Research and Statistics Group, Sloan School of Management, Massachusetts Institute of Technology; email: jvielma@mit.edu.

§Department of Health Care Policy, and Department of Statistics, Harvard University; email: zubizarreta@hcp.med.harvard.edu.

1 Introduction

1.1 Practical appeal of matching

In observational studies, matching is a general method for covariate adjustment that approximates the ideal experiment that would be conducted if controlled experimentation was possible. Under the assumption of strong ignorability (Rosenbaum and Rubin 1983), or no unmeasured confounders (Imbens 2004), matching methods are often used to estimate treatment effects in observational studies with binary treatments (e.g., Dehejia and Wahba 1999; Haviland et al. 2007), longitudinal (Lu 2005; Zubizarreta et al. 2014) and multilevel data (Li et al. 2013; Zubizarreta and Keele 2017), but also under different assumptions, for example with instrumental variables (Baiocchi et al. 2010; Zubizarreta et al. 2013b) or in discontinuity designs (Keele et al. 2015; Mattei and Mealli 2016).

The practical appeal of matching methods lies in part in the conceptual simplicity and transparency of their adjustments (Cochran and Rubin 1973). These adjustments are an interpolation instead of an extrapolation based on a model that can be misspecified (Rosenbaum 1987a). Matching also enables the integration of quantitative and qualitative analyses (Rosenbaum and Silber 2001). Furthermore, since matching methods do not routinely use the outcomes for their adjustments, it is often argued that they promote the objectivity of the study by separating the design and analysis of an observational study into two distinct stages (Rubin 2008). Finally, matching methods can facilitate simpler forms of statistical inference and sensitivity analyses to hidden biases (see, for instance, Chapter 3 of Rosenbaum 2010). See Stuart (2010), Imbens (2015), and Rosenbaum (2017) for overviews of matching methods.

1.2 Three challenges

Most of the work in matching has been for binary treatments and, although there are methods for more general treatments (e.g., Lu et al. 2001, 2011; Yang et al. 2016; Lopez et al. 2017), some challenges remain. One of these challenges relates to the difficulty of balancing

covariates by means of the estimated propensity score (Rosenbaum and Rubin 1983) or generalizations thereof (e.g., Joffe and Rosenbaum 1999; Imbens 2000; Imai and Van Dyk 2004). The propensity score is the conditional probability of treatment assignment given observed covariates. It has the important property that matching on the propensity score tends to balance the covariates used to estimate the score; however, in any given data set it might be difficult to balance the covariates even if the propensity score model is correctly specified (Yang et al. 2012). This challenge has been discussed, e.g., by Hill (2011) and Zubizarreta et al. (2011), and alternative methods that directly balance the covariates have been proposed (e.g., Diamond and Sekhon 2013; Zubizarreta 2012), yet these are for binary treatments. The difficulties of balancing covariates by matching on generalizations of the propensity score can easily be exacerbated with multi-valued treatments.

The second challenge with matching methods, both for binary and multi-valued treatments, is targeting parameters of general scientific and policy interest, especially when there is limited overlap in covariate distributions across treatment groups. When there is sufficient overlap, it is possible to estimate the average effect of one treatment in place of another treatment (possibly the control treatment) on one of these two treatment groups, but it is difficult to estimate other parameters such as the average effect on all treatment groups (in resemblance to the standard average treatment effect; ATE), or on a particular group as defined by covariates (the conditional average treatment effect; CATE), without matching with replacement or reweighting which can make inference more complicated, as discussed by Abadie and Imbens (2008). When there is limited overlap, it is not possible to target the average effect of one treatment in place of another treatment for any of these treatment groups without imposing strong parametric assumptions, and it is common in practice to settle for more local average effects which may have more internal validity but be less generalizable. Often times, these analyses are criticized for their limited scientific and policy interest (see, e.g., Imbens 2010).

A third challenge relates to matching in large data sets. Most common optimization-based matching methods rely on quadratic-sized formulations that cannot handle data sets with hundreds of thousands of observations quickly. As we describe below, the problem is that

3

such formulations are too big to be practical or that they require additional structure on the covariates in order to run quickly.

Through a case study of the impact of an earthquake on educational achievement, our goal in this paper is to overcome these three challenges and present a new matching method that: (i) handles multi-valued treatments and directly balances covariates without estimating a generalization of the propensity score; (ii) finds self-weighting matched samples that are not only balanced but also have a similar structure to the one of a target population (thereby allowing the investigator to target parameters of general policy and scientific interest); and (iii) runs quickly in large data sets, for example, with hundreds of thousands of observations in a couple of minutes. For this, we use the following ideas.

1.3   Main ideas

Optimization-based methods for matching with binary treatments can be divided, roughly, into network flow methods (e.g., Rosenbaum 1989; Hansen 2004; Pimentel et al. 2015) and integer or mixed integer programming (MIP) methods (e.g., Nikolaev et al. 2013; Sauppe et al. 2014; Zubizarreta 2012; Zubizarreta et al. 2014). For the most, network flow methods minimize an aggregate measure of covariate distances, and not covariate balance directly, although there is the clever extension by Pimentel et al. (2015) for nested nominal covariates. The advantage of MIP methods is that they target covariate balance (and distances) more directly and flexibly. However, MIP based methods require the solution of a theoretically intractable or NP-hard problem, whereas network flow matching methods only require the solution of a polynomially solvable problem that is tractable both in theory and practice. Still, thanks to state-of-the-art MIP solvers which nearly double their speeds every year (Bixby 2012; Achterberg and Wunderling 2013), MIP methods are mostly tractable in practice, and hence the computational advantage of network flow methods has been steadily decreasing.

If we consider matching with three or more treatments or exposures, this computational advantage vanishes as network flow methods for minimum distance problems also become

4

theoretically intractable (Michael and David 1979). In this case, direct adaptations of both network flow and MIP-based methods are also intractable in practice. To circumvent this computational limitation, we match each treatment group to a representative sample of a target population and develop a new MIP-based matching formulation. A key to develop this method is a linear programming (LP) view of the difference in computational complexity between minimum distance matching and MIP-based methods. This view models the central problem of both methods as a MIP problem and then analyses the strength of their LP relaxation. For minimum distance matching the LP relaxation has the strongest possible *integral* property, which implies the associated MIP problem can be solved as an LP problem in polynomial time (Schrijver 2003). For MIP based techniques the LP relaxation fails to be integral, but having an LP relaxation that is "close" to integral is known to result on small solve times (Vielma 2015). Unfortunately, we often face a trade-off between strong, but large formulations and smaller, but weaker formulations. With this in mind, we show that a linear-sized MIP formulation for covariate balance is as strong as an alternative, and more common, quadratic-sized formulation. We also show that this formulation is in fact integral when only two covariates are considered or when the nested covariate structure of Pimentel et al. (2015) is present. While the formulation is not integral when more than three covariates are used, the associated MIP nonetheless remains strong and can be solved in minutes in large data sets. In contrast, attempting to solve the problem with existing formulations (both network and quadratic-sized MIP formulations) results in an out-of-memory error even in a workstation with 32GB of RAM. In this age of big data, this algorithmic characterization is key to handle large data sets in a practical manner.

In addition to allowing a computationally practical approach for matching with multi-valued treatments, the use of this auxiliary reference sample also increases the versatility of the method. By selecting this reference or template sample in different ways, we can target different estimands beyond the common average treatment effect on the treated (ATET), such as the (general) average treatment effect (ATE) or a conditional average treatment effect (CATE) for a particular group of the population (e.g., subjects with a particular age in a particular ethnic group). In this manner we can build representative matched samples with multi-valued treatments that may be numerous and unordered. In our procedure,

covariate distances between matching play a secondary role (relegated to the rematching of the balanced units, in the spirit of Zubizarreta et al. 2014). In fact, our procedure does not require explicitly modeling the propensity score and directly balances the covariates by design; this is, as specified before matching by the investigator. Furthermore, it facilitates checking covariate balance, as one can tabulate and plot the distributions of the observed covariates as opposed to evaluating balance in the context of a model.

1.4   Case study and outline

We illustrate all these ideas in the context of an observational study of the impact of a natural disaster on educational achievement; in particular, of the impact of the 2010 Chilean earthquake on test scores of university admission exams. This is a very important question because due to its location in the Pacific Ring of Fire, Chile is considered one of the most seismically active countries in the world (CFE-DM 2017), concentrating 78 earthquakes above 7.0 magnitude since 1900 (CSN Universidad de Chile 2018) and holding the record for the largest earthquake ever registered (USGS 2018). Furthermore, despite its sustained economic growth and being one of the most robust economies in Latin America, Chile has the highest Gini coefficient among the 35 OECD countries, followed by Mexico and the United States, which makes it the country with the most unequal income distribution of the OECD countries, according to this indicator (OECD 2016). In this context of extreme income inequality it is important to understand the impact of these recurring natural disasters on standardized university admissions examination scores as they almost fully determine students' entrance to university, one of the main avenues of social mobility and opportunity in Chile (Torche 2005).

To address this question, the rest of this paper is organized as follows. In Section 2, we describe the Chilean educational system, the 2010 Chilean earthquake, and a particularly rich data set that is a census of the same students before and after the earthquake. In Section 3 we explain how to use the ideas in Silber et al. (2014) for building representative matched samples of target populations with possibly unordered and many treatment or exposure groups. In Section 4, we formally analyze and contrast the properties of linear- and

quadratic-sized MIP formulations for distributional balance, and evaluate their performance in data sets of increasing size. In Section 5 we present estimates of the impact of levels of exposures to the earthquake both on school attendance and standardized test scores for university admission. In Section 6 we conclude with a summary and remarks.

2 Impact of a natural disaster on educational opportunity

2.1 On the Chilean educational system

In Chile, most students complete the required 12 years of school education (MDS 2013). In their last year of high school, students take a college admission test called *Prueba de Selección Universitaria* or PSU in order to apply to college. The PSU is a high-stakes test as it comprises nearly 80% of the final college admission score and it can only be taken once a year, at the end of the Chilean academic year. Most students register for the test in their last year of high school, and the vast majority of them takes it only once in their lives. The importance of the PSU does not only relate to being admitted into a given college, but also to obtaining financial aid and having the possibility to actually attend to college (Dinkelman and Martinez 2014). In Chile, higher education is one of the main avenues for social mobility and the improvement later life outcomes (Torche 2005). Studies have shown that even though the returns to higher education in Chile are heterogeneous, there are still high, positive returns to attending college, specially highly-selective ones (Hastings et al. 2013; González-Velosa et al. 2015).

2.2 The 2010 Chilean earthquake

In February 27th, 2010, an earthquake of magnitude 8.8 struck near Concepción, Chile's second largest city, going down in history as the 6th most severe earthquake registered since 1900 so far (USGS 2014). The disaster severely damaged over 500,000 homes and affected

more than 2 million people. Losses amounted to 18% of the gross domestic product (MINE-DUC 2013). Nearly 20% of the schools in the affected regions suffered moderate damages or worse, and many schools had to be completely reconstructed. In addition, over 40% of the students in these regions could not start their academic year on time. 24 million US dollars were redirected to rebuild and restore the affected schools and the academic calendar was adjusted (MINEDUC 2013).

2.3   A longitudinal census of students

In our analyses, we use a new and rich administrative data set of the same high school students measured before and after the earthquake. This data set is a longitudinal census as it comprises all Chilean students in 10th grade in 2008 (before the earthquake) and it collects their information again in 12th grade in 2010 (after the earthquake), when they are finishing high school and applying for college. In 2008, the data provides detailed measures of the students, their schools, and respective households. For instance, the data provides the standardized test scores from the Education Quality Measurement System (SIMCE), school attendance, GPA ranking within the school, and school characteristics such as its socioeconomic status. In addition, it offers extensive information about socioeconomic characteristics of the students and their households, such as their parents education and the household income.

In 2010, the data contains two outcomes of interest: (i) the students' school attendance that academic year, and (ii) the language and mathematics PSU scores (both measured after the earthquake). The first outcome is assesed as percentage. The scale of the PSU is the same for the two tests and it ranges between 150 and 850 points, with a mean of 500 points and a standard deviation of 110 points (DEMRE 2018). The data comprises 121279 students measured before and after the earthquake.
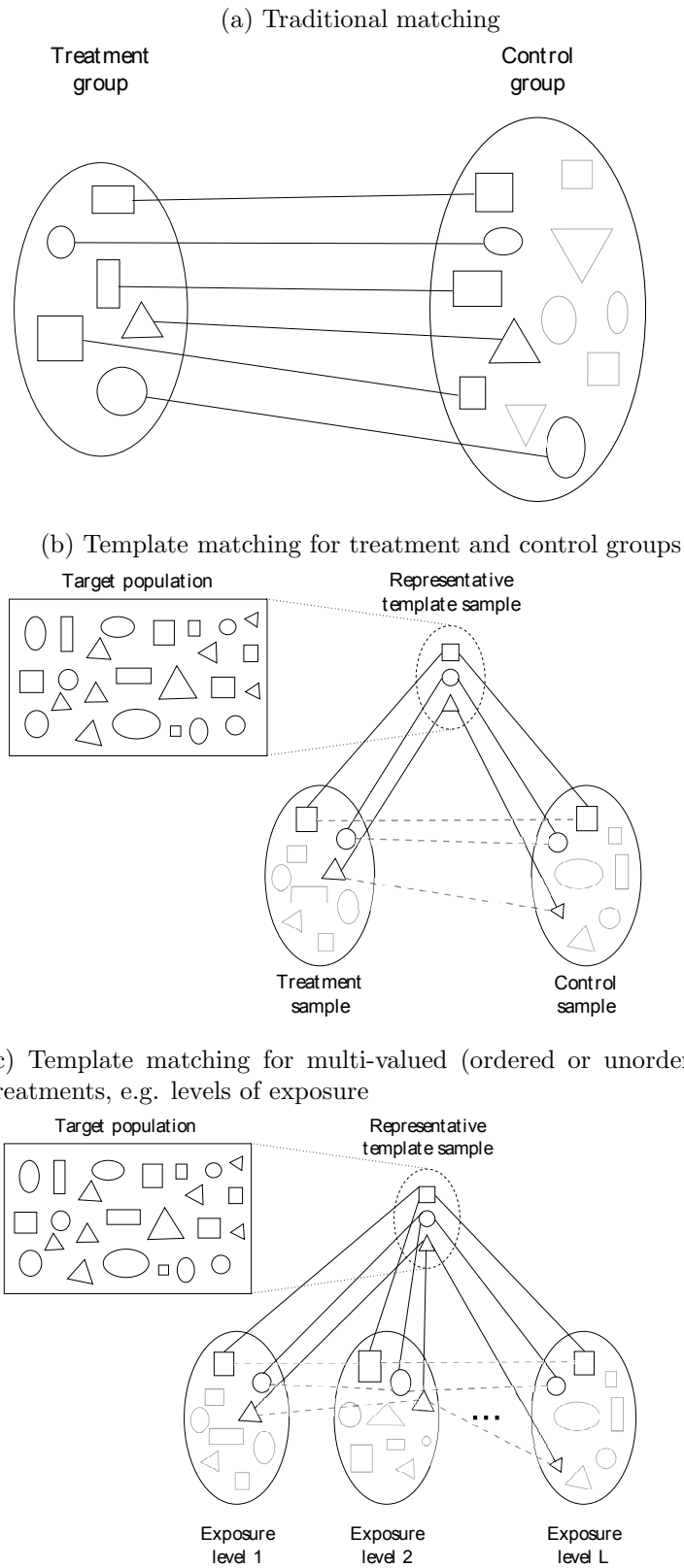
2.4 Earthquake intensity levels

We use peak ground acceleration (PGA) to measure the strength of the earthquake. In contrast to other seismic intensity scales, for example, the Mercalli or the Richter scales, PGA is a purely physical measure of the shaking of the earthquake at a given location. Following Zubizarreta et al. (2013a), we used the PGA values provided by the United States Geological Survey (USGS 2011) to estimate the PGA in each of the counties where we have data. Using these values, we created three measures of exposure to the earthquake: one with three levels of shaking, another with five levels, and a final one with ten, basically defined by quantiles of exposure (see Appendix A in the Supplementary Materials for details).

3 Representative matching with multi-valued treatments

As noted in Section 1.3, minimum distance matching (arguably, the mainstream form of optimal matching) with two treatments is computationally tractable (i.e. polynomially solvable), but it is computationally intractable (NP-hard) for three or more treatments. To circumvent this limitation, instead of explicitly matching samples across treatments, we will separately match each treatment sample to a representative or template sample of a target population of special interest, extending the ideas of Silber et al. (2014). This will allow us to address the two aforementioned difficulties of handling multi-valued treatments and building representative samples in matching. In the following section, we incorporate the idea into a MIP-based matching approach for large data sets. The basic idea is depicted in Figure 1.

Figure 1(a) shows traditional bipartite matching with a smaller treatment group than a control group. The goal is to match all the treatment and control units with similar observed covariates, represented in the figure by the shapes of the elements. Here, the target of inference is the average treatment effect on the treated (ATET). But how to target a different estimand, one for a particular population of policy or scientific interest? Call this estimand the target average treatment effect (TATE; Kern et al. 2016). Figure 1(b) illustrates how

Figure 1: Template matching for multi-valued treatments.

(a) Traditional matching

(b) Template matching for treatment and control groups

(c) Template matching for multi-valued (ordered or unordered) treatments, e.g. levels of exposure

to extend template matching for this purpose. Here, a representative template sample is selected from the target population of interest and then matched to the treatment and control groups separately. By construction, the matched groups will be balanced relative to the template sample and therefore to each other. We can repeat this process with multi-valued (ordered or un-ordered) treatments, possibly tens or hundreds treatments — as many as the data allows us to balance. This is illustrated in Figure 1(c).

We select the template sample by drawing random samples of a given size from the population of interest. In our case study, the population of interest was the population of all high school students in tenth grade in 2008 in Chile, and we drew 500 random samples of size 1000 from this population. We chose this sample size so that it was large enough to represent all levels of exposure to the earthquake. In this paper, we fixed this sample size for simplicity of exposition, but the matching formulations in Section 4 can be extended to relax this requirement and its implied fixed matching ratio by finding the largest sample for each treatment that is balanced relative to certain moments of the empirical distribution of the covariates that characterizes the target population. From the 500 random samples, we selected the sample that was closest to the population in terms of a robust version of the Mahalanobis distance (Rosenbaum 2010, Chapter 8). Table 5 in the Supplementary Materials describes the population and the selected template sample in terms of the means for each category of the observed covariates. In our case study, we used this template to match representative samples for each level of exposure to the earthquake as follows.

As described in Table 5, we have 14 categorical covariates with 78 categories in total, or 78 binary covariates for each of these categories. These covariates are: the student's gender, and ethnicity (3 categories); the father and mother's education (5 categories each), household income (7 categories), and number of books at home (6 categories); student's school attendance (10 categories), GPA ranking within the school (10 categories), and SIMCE score (10 categories); the school's type (public, voucher, or private), location (rural or urban), whether the school is catholic or not, its socioeconomic status (SES; 5 categories), and the average SIMCE score of the school (10 categories). We perfectly balanced the marginal distributions of all these covariates by matching with fine balance (Rosenbaum et al. 2007). Fine bal-

ance perfectly balances the marginal distributions of the observed covariates, but without constraining units to be matched within the same categories of the covariates (see Visconti and Zubizarreta 2018 for an illustration). By matching with fine balance to the template, we can guarantee that the matching across multiple values of the exposure will be perfectly balanced relative to each other and to template of the population in terms of the observed covariates (see Table 3 below).

In our case study, identification of the TATE relies on strong ignorability (Rosenbaum and Rubin 1983; Imbens and Rubin 2015) of both the treatment assignment and the sample selection mechanisms (see, e.g., Tipton 2013 and Kern et al. 2016). The fact that the entire territory of the country is prone to earthquakes that people cannot anticipate, the broad scope of the covariates in our data set for which we adjust by matching, and the act of randomly selecting the template sample from the (entire) population, make these assumptions plausible. We also assume the stable unit treatment value assumption (SUTVA; Rubin 1980, 1986) holds.

4 Effective formulations for matching using integer programming

In this section, we analyze and contrast the effectiveness linear- and quadratic-sized mixed integer programming matching formulations. In particular, we show that a particular linear-sized formulation is more effective as it permits us to handle considerably larger data sets in a shorter time. For example, in our case study, one of the data sets involved more than 70,000 units, yet the linear-sized formulation could find the solution in less than one minute in a standard laptop computer. With quadratic-sized formulations, the problem did not even fit in memory. Coupled with the matching strategy in the above section, this allows us to build balanced and representative matched samples with multi-valued treatments in considerably larger data sets.

Parts of this section are technical and so we begin by describing its structure and main results.

In Section 4.1 we describe the notation and a quadratic-sized formulation for distributional balance that works well in some instances, but that cannot handle the data in our case study due to its size. In Section 4.2 we describe how the quadratic-sized formulation can be reduced to a linear-sized formulation through a simple procedure that could potentially result in a loss of formulation strength. In Section 4.3 we show that no strength is actually lost with this reduction and prove three results: the linear programming (LP) relaxation (i.e., the "strength") of the linear-sized formulation is equal to the one of the quadratic-sized formulation; with two covariates or nested covariates, the LP relaxation of both formulations is integral; and with three or more covariates, the LP relaxation of both formulations can fail to be integral. In practice, however, the smaller formulation is quite practical. As we show in Section 4.4, this formulation can handle data sets with more than 700,000 units in only a couple of minutes.

4.1 A large formulation for distributional balance

Let $\mathcal{T} = \{t_1, \ldots, t_T\}$ represent the T units in the template sample, $\mathcal{L} = \{\ell_1, \ldots, \ell_L\}$ be the L units in the sample under a given treatment or exposure level, and $\mathbb{L}$ be the family containing such sets $\mathcal{L}$ of units for each treatment or exposure level (which we do not explicitly index to emphasize that exposure levels do not need to be ordered). In the following, we fix $\mathcal{L} \in \mathbb{L}$ because, as explained in the previous section (Figure 1(c)), the matching process is identical for every exposure level $\mathcal{L} \in \mathbb{L}$. In addition, let $\mathcal{P} = \{p_1, \ldots, p_P\}$ denote the P observed covariates that we aim to balance and $\mathbf{x}_i = (x_{i,p})_{p \in \mathcal{P}}$ define the covariate values for the units in the template and exposure level samples for $i \in \mathcal{T}$ and $i \in \mathcal{L}$, respectively. Finally, let $\mathcal{K}(p) = \{k_1, \ldots, k_{K_p}\}$ stand for the categories of covariate $p \in \mathcal{P}$, specify the template (level l) units in category k in covariate p as $\mathcal{T}_{p,k} = \{t \in \mathcal{T} : x_{t,p} = k\}$ ($\mathcal{L}_{p,k} = \{\ell \in \mathcal{L} : x_{\ell,p} = k\}$) and let $N_{p,k} = |\mathcal{T}_{p,k}|$.

We can minimize the imbalances in the marginal distributions of the P covariates, through

the following mixed integer programming problem

$$
\begin{aligned}
\operatorname*{minimize}_{\bm{v},\,\bm{m}} \quad & \sum_{p\in\mathcal{P}}\sum_{k\in\mathcal{K}(p)} v_{p,k} && \text{(1a)}\\
\text{subject to} \quad & \left|\sum_{\ell\in\mathcal{L}_{p,k}}\sum_{t\in\mathcal{T}} m_{t,\ell} - N_{p,k}\right| \le v_{p,k}, && \forall p\in\mathcal{P}, \quad k\in\mathcal{K}(p), && \text{(1b)}\\
& \sum_{t\in\mathcal{T}} m_{t,\ell} \le 1, && \forall \ell\in\mathcal{L}, && \text{(1c)}\\
& \sum_{\ell\in\mathcal{L}} m_{t,\ell} = 1, && \forall t\in\mathcal{T}, && \text{(1d)}\\
& m_{t,\ell}\in\{0,1\}, && \forall t\in\mathcal{T},\quad \ell\in\mathcal{L}, && \text{(1e)}
\end{aligned}
$$

where $v_{p,k}$ defines the imbalance or violations from fine balance (Rosenbaum et al. 2007) for category k of covariate p, and $m_{t,\ell}$ is the binary decision variable which takes the value of 1 if template unit t is matched to level $\mathcal{L}$ unit ℓ, and 0 otherwise.

Formulation (1) has $T\times L+\sum_{p\in\mathcal{P}} K_p$ variables and $T+L+\sum_{p\in\mathcal{P}} K_p$ constraints excluding the variable bounds. While this problem size is polynomial on the number of units, the quadratic term $T\times L$ can be prohibitive in practice. For instance, in our case study one of the data sets results in 1.8×10^6 decision variables. This makes the problem very hard in practice and attempting to solve it with a state-of-the-art solver, Gurobi v7.5.1 on a Core i7-3770 3.40GHz workstation with 32GB of RAM, results in an out-of-memory error. We get the same error even if we attempt to solve the much easier LP relaxation obtained by relaxing constraint (1e) into $m_{t,\ell}\in[0,1]$.

4.2 Reducing the size of the formulation

One approach to reduce the size of formulation (1) is to define variables $z_\ell = \sum_{t\in\mathcal{T}} m_{t,\ell}$ for each $\ell\in\mathcal{L}$ and use these new variables to eliminate the $m_{t,\ell}$ variables from formulation (1).

For this, we first sum constraints (1d) to obtain

$$\underset{\boldsymbol{v},\,\boldsymbol{m}}{\text{minimize}} \quad \sum_{p\in\mathcal{P}}\sum_{k\in\mathcal{K}(p)} v_{p,k} \tag{2a}$$

$$\text{subject to} \quad \left|\sum_{\ell\in\mathcal{L}_{p,k}}\sum_{t\in\mathcal{T}} m_{t,\ell} - N_{p,k}\right| \leq v_{p,k}, \qquad \forall p\in\mathcal{P}, \quad k\in\mathcal{K}(p) \tag{2b}$$

$$\sum_{t\in\mathcal{T}} m_{t,\ell} \leq 1, \qquad \forall \ell\in\mathcal{L} \tag{2c}$$

$$\sum_{t\in\mathcal{T}}\sum_{\ell\in\mathcal{L}} m_{t,\ell} = T, \tag{2d}$$

$$m_{t,\ell} \in \{0,1\}, \quad \forall t\in\mathcal{T}, \quad \ell\in\mathcal{L}. \tag{2e}$$

We claim that formulation (1) is equivalent to (2). For that, first note that any solution to (1) is a solution to (2). Second, suppose that a solution $(\boldsymbol{v}^*,\ \boldsymbol{m}^*)$ to (2) has some $t'\in\mathcal{T}$ such that $\sum_{\ell\in\mathcal{L}} m_{t',\ell}^* \geq 2$. Then, because of (2d), there is some $t''\in\mathcal{T}$ with $t''\neq t'$ such that $\sum_{\ell\in\mathcal{L}} m_{t'',\ell}^* = 0$. If $\ell'\in\mathcal{L}$ is such that $m_{t',\ell'}^* = 1$ we can change the solution by letting $m_{t',\ell'}^* = 0$ and $m_{t'',\ell'}^* = 1$. This does not change $\sum_{t\in\mathcal{T}} m_{t,\ell}^*$ for any $t\in\mathcal{T}$ so the solution remains feasible for (2). If we repeat this step, we eventually get a solution to (1).

Noting that the order of the sums in (2d) can be exchanged, we can replace every occurrence of $\sum_{t\in\mathcal{T}} m_{t,\ell}$ by z_ℓ in (2) to obtain the following equivalent formulation

$$\underset{\boldsymbol{v},\,\boldsymbol{z}}{\text{minimize}} \quad \sum_{p\in\mathcal{P}}\sum_{k\in\mathcal{K}(p)} v_{p,k} \tag{3a}$$

$$\text{subject to} \quad \left|\sum_{\ell\in\mathcal{L}_{p,k}} z_\ell - N_{p,k}\right| \leq v_{p,k} \qquad \forall p\in\mathcal{P}, \quad k\in\mathcal{K}(p) \tag{3b}$$

$$\sum_{\ell\in\mathcal{L}} z_\ell = T \qquad \forall p\in\mathcal{P}, \quad k\in\mathcal{K}(p) \tag{3c}$$

$$z_\ell \in \{0,1\} \quad \ell\in\mathcal{L}, \tag{3d}$$

where z_ℓ is a binary decision variable that takes the value 1 if unit ℓ of the exposure level group $\mathcal{L}$ is matched to the template $\mathcal{T}$, and 0 otherwise. Formulation (3) has $L + \sum_{p\in\mathcal{P}} K_p$ variables and $1 + \sum_{p\in\mathcal{P}} K_p$ constraints excluding the variable bounds, which is significantly smaller than formulation (1). Indeed, for the aforementioned data set in our case study, going from formulation (1) to (3) yields a reduction in the number of variables from over

1.8×10^6 to 19208 and a reduction in the number of constraints from 19370 to 162. Thanks to this drastic decrease in size we can solve the LP relaxation of (3) for this data set in less than one second.

However, this reduction in problem size and solve time for the LP relaxation may not necessarily translate into a reduction in solve time for the MIP formulation (3). To analyze this solve time, it is convenient to consider the tractability of minimum distance matching (e.g., Rosenbaum 1989) from an LP perspective instead of the traditional graph theoretical one (i.e., the assignment problem or bipartite matching one). From this perspective, the tractability of minimum distance matching can be attributed to the fact that the LP relaxation of its standard assignment problem formulation has extreme points or basic feasible solutions that always satisfy the integrality constraints of the formulation. A formulation with this property is usually denoted *integral* and this property implies that solving the formulation is equivalent to solving its LP relaxation and hence minimum distance matching is polynomially solvable (Schrijver 2003). In contrast, the extreme points of the LP relaxations of (1) and (3) do not necessarily satisfy the integrality constraints. Hence, these problems are not necessarily polynomially solvable and may need to be solved with a MIP solver instead of just an LP solver. In such cases, a relatively good predictor of the difficulty of solving the optimization problem with a MIP solver is the distance between the optimal value of the original MIP problem and its LP relaxation, which is often denoted the *integrality* or *LP GAP* (Vielma 2015). If the GAP is small, the formulation is *strong* and it is expected to lead to faster solve times than similar sized *weaker* formulations with a larger GAPs. Unfortunately, the elimination of variables through $z_\ell = \sum_{t \in \mathcal{T}} m_{t,\ell}$ to obtain (3) from (1) does not automatically guaranteed to preserve formulation strength, and hence the improvements from size reduction could be negated by a loss of strength.

4.3   Complexity and strength of the formulation

Using results from Balas and Pulleyblank (1983) we can fortunately show that no formulation strength is lost in the transformation from (1) to (3). We formalize this fact in the following simple proposition for which we give a self-contained proof.

Proposition 4.1. *The LP relaxations of formulations* (1) *and* (3) *are equivalent.*

Proof. We follow a similar logic to our arguments for the equivalence between (1) and (2), but using continuous instead of binary variables.

First, let $(\boldsymbol{v},\, \boldsymbol{m})$ be a feasible solution to the LP relaxation of (1). Then $(\boldsymbol{v},\, \boldsymbol{z})$ for $\boldsymbol{z}$ given by $z_\ell = \sum_{t \in \mathcal{T}} m_{t,\ell}$ for all $\ell \in \mathcal{L}$ is a feasible solution to the LP relaxation of (3) obtained by relaxing (3d) into $z_\ell \in [0, 1]$.

For the converse, let $(\boldsymbol{v}, \boldsymbol{z})$ be a feasible solution to the LP relaxation of (3). We obtain a feasible solution $(\boldsymbol{v}, \boldsymbol{m})$ to the LP relaxation of (1) by constructing $\boldsymbol{m}$ as follows. Let $\underline{s}_1 = 0$, $\overline{s}_1 = \min\left\{ s \geq 1 : \sum_{i=\underline{s}_1+1}^{s} z_{\ell_i} \geq 1 \right\}$, $m_{t_1, \ell_{\overline{s}_1}} = 1 - \sum_{i=\underline{s}_1+1}^{\overline{s}_1-1} z_{\ell_i}$ and for each $i \in \{\underline{s}_1+1, \ldots, \overline{s}_1-1\}$ let $m_{t_1, \ell_i} = z_{\ell_i}$. Then, for each $j \in \{2, \ldots, T\}$ let $\underline{s}_j = \overline{s}_{j-1}$, $\overline{s}_j = \min\left\{ s \geq s_{j-1} : \left(z_{\ell_{\underline{s}_j}} - m_{t_{j-1}, \ell_{\underline{s}_j}} \right) + \sum_{i=\underline{s}_j+1}^{s} z_{\ell_i} \geq 1 \right\}$, $m_{t_j, \ell_{\overline{s}_j}} = 1 - \sum_{i=\underline{s}_j+1}^{\overline{s}_j-1} z_{\ell_i} - \left(z_{\ell_{\underline{s}_j}} - m_{t_{j-1}, \ell_{\underline{s}_j}} \right)$, $m_{t_j, \ell_{\overline{s}_j}} = \left(z_{\ell_{\underline{s}_j}} - m_{t_{j-1}, \ell_{\underline{s}_j}} \right)$, and for each $i \in \{\underline{s}_j+1, \ldots, \overline{s}_j-1\}$ let $m_{t_j, \ell_i} = z_{\ell_i}$. Because of (3c) we have that $m_{t_T, \ell_{\overline{s}_T}} = z_{\ell_{\overline{s}_T}}$ and $z_{\ell_i} = 0$ for all $i > \overline{s}_T$. Then, by construction m satisfies (1c)–(1d), $m_{t,\ell} \in [0, 1]$ for all $t \in \mathcal{T}$ and $\ell \in \mathcal{L}$, and $z_\ell = \sum_{t \in \mathcal{T}} m_{t,\ell}$ for all $\ell \in \mathcal{L}$. Finally, because of this last equation and the fact that (z, v) satisfies (3b) we have that $(\boldsymbol{v}, \boldsymbol{z})$ satisfies (1b). $\square$

Proposition 4.1 shows that we do not loose formulation strength with the size reduction, but it does not tell us how strong the equivalent formulations are. The following proposition, that we prove in Appendix C in the Supplementary Materials, shows that for problems with at most two covariates or with nested covariates the LP relaxations of both formulations are integral and that, similar to the assignment formulations for propensity score matching, their solution is equivalent to the solution of their LP relaxations.

Proposition 4.2. *The LP relaxations of* (1) *and* (3) *are integral (i.e. they have integral extreme points) if*

1. $P \leq 2$, *or*

2. *for all $i < j$ and $k \in \mathcal{K}(p_j)$ there exist $k' \in \mathcal{K}(p_i)$ such that $\mathcal{L}_{p_j,k} \subseteq \mathcal{L}_{p_i,k'}$.*

In particular, under these conditions (1) *and* (3) *can be solved in polynomial time by solving their LP relaxations.*

Unfortunately, the following lemma, that we prove in Appendix B, shows that the formulations can fail to be integral for non-nested covariates even if $P = 3$.

Lemma 4.3. *Even for $P = 3$ and $K_1 = K_2 = K_3 = 3$, there exist covariates for which the LP relaxations of* (1) *and* (3) *fail to be integral.*

This loss of integrality for more general covariate structures is not surprising as the problem is NP-hard. Fortunately, enough of the formulation strength is preserved so that formulation (3) remains extremely effective. Indeed the size and strength of formulation (3) allows to solve all the matching problems in our case study in a few seconds in a regular laptop (see Table 2) even though they do not satisfy the covariate assumptions of Proposition 4.2.

4.4 Performance of the linear-sized formulation

To evaluate the performance of the linear-sized formulation, we implemented it in the new function `cardmatch` in `designmatch` for `R` (Zubizarreta et al. 2018) and tested it in data sets of different sizes. We increased the largest exposure sample in our data set ($L = 70118$) up to ten times ($L = 701180$) by creating random copies of it. For each copy, we randomly modified all the nominal covariates by adding 1, 0, or -1, and truncating the resulting values to preserve the original ranges. We also created templates of different sizes of up to 10000 observations.

Table 1 shows the computing times. The largest data set we considered had a template sample size of $T = 10000$ and an exposure sample size of $L = 701180$, and took approximately three minutes. Most of the matchings took less than two minutes. Naturally, the computing time tends to increase both with the template and exposure sample sizes.

Table 2 shows the matching time in our case study with 3, 5, and 10 levels of exposure, and a template sample size of $T = 1000$. The table shows that all matching times are well under a minute, and most of them are under 10 seconds. The computing times do not increase monotonically with the sample size in part due to random variation in the generation of

18

the covariate values. We also compared the performance of the linear-sized formulation in `cardmatch` to other matching packages based on quadratic-formulations, such as `optmatch` (Hansen 2007) and `rcbalance` (Pimentel et al. 2015). Trying to run `optmatch` results in an out-of-memory error, even for our original case study sample ($T = 1000$ and $L = 70118$). In the case of `rcbalance`, we were only able to match the original sample ($L = 70118$) to template sizes up to size $T = 6000$. For larger template sizes, we were not able to get a solution in the allowed time of 8 hours.

Table 1: Computing time (in minutes) of the proposed matching formulation as implemented in the new function `cardmatch` in `designmatch` for R (Zubizarreta et al. 2018). We considered different template samples sizes (ranging from 1000 to 10000) and varying exposure sample sizes, increasing from the largest exposure sample size in our case study, 70118, to $10 \times 70118 = 701180$.

Template size T	Exposure size L									
	70118	140236	210354	280472	350590	420708	490826	560944	631062	701180
1000	0.28	0.50	0.65	0.79	1.11	1.20	1.49	2.13	2.58	2.63
2000	0.20	0.72	0.91	1.14	1.49	1.56	1.87	2.20	2.53	2.67
3000	0.19	0.73	1.08	1.37	1.62	1.51	2.02	2.26	2.53	3.15
4000	0.22	0.44	1.09	1.57	1.74	1.98	2.00	2.29	2.48	2.62
5000	0.18	0.33	0.87	1.26	1.52	1.94	3.05	1.73	2.93	3.51
6000	0.26	0.47	0.64	1.66	2.07	2.40	2.78	2.94	3.18	3.04
7000	0.18	0.36	0.56	0.76	1.62	2.09	2.28	2.36	2.71	8.54
8000	0.25	0.40	0.57	0.82	1.87	2.25	2.42	2.95	3.08	3.66
9000	0.25	0.46	0.74	0.82	0.99	2.18	2.94	3.13	4.13	3.85
10000	0.19	0.39	0.63	0.83	1.08	2.55	2.58	2.93	3.13	3.42

Table 2: Computing time (in minutes) using `cardmatch` in the case study with 3, 5, and 10 levels of exposure to the earthquake.

Exposure level	Sample size	Time (min)
1	18208	0.05
2	70118	0.34
3	32953	0.08
1	22075	0.06
2	25977	0.06
3	24896	0.06
4	24279	0.06
5	24052	0.06
1	12084	0.03
2	9991	0.03
3	12513	0.04
4	13464	0.03
5	13119	0.03
6	11777	0.04
7	12813	0.03
8	11466	0.03
9	12071	0.03
10	11981	0.03

5 Results from the case study

5.1 Assessing balance

Table 3 shows covariate balance for the matchings with 3, 5, and 10 levels of exposure to the earthquake. In the template (target) sample there are 490 female and 510 male students, and the same is true across all exposure levels. Due to space constraints, the table shows the counts for 3 covariates only (gender, school SES, and mother's education), but same pattern holds for all other 11 covariates (see Figure 3 in the Supplementary Materials). In other words, the marginal distributions of the 14 covariates are perfectly balanced relative to each other and to the representative template sample. As a result, Figure 3 shows that means of the indicators for the 78 covariate categories are perfectly balanced after matching.

Table 3: Distributional balance or fine balance across matched samples for 3, 5, and 10 levels of exposure to the earthquake. Due to space constraints, the counts are shown for three covariates only but the same pattern holds for all other 11 covariates.

(a) 3 exposure levels

| | | | Exposure level | | |
Covariate	Template	1	2	3
Gender				
Male	490	490	490	490
Female	510	510	510	510
School SES				
Low	90	90	90	90
Mid-low	318	318	318	318
Medium	303	303	303	303
Mid-high	174	174	174	174
High	115	115	115	115
Mother's education				
Primary	321	321	321	321
Secondary	434	434	434	434
Technical	120	120	120	120
College	115	115	115	115
Missing	10	10	10	10

⋮

(b) 5 exposure levels

| | | | | Exposure level | | |
Covariate	Template	1	2	3	4	5
Gender						
Male	490	490	490	490	490	490
Female	510	510	510	510	510	510
School SES						
Low	90	90	90	90	90	90
Mid-low	318	318	318	318	318	318
Medium	303	303	303	303	303	303
Mid-high	174	174	174	174	174	174
High	115	115	115	115	115	115
Mother's education						
Primary	321	321	321	321	321	321
Secondary	434	434	434	434	434	434
Technical	120	120	120	120	120	120
College	115	115	115	115	115	115
Missing	10	10	10	10	10	10

⋮

(c) 10 exposure levels

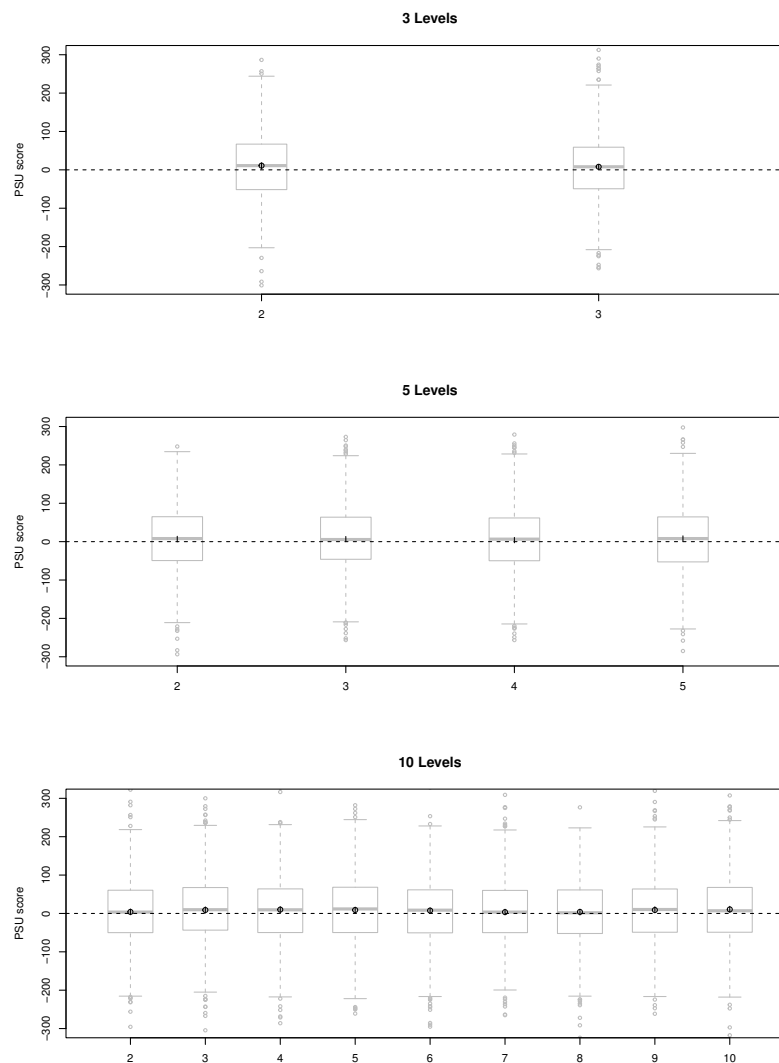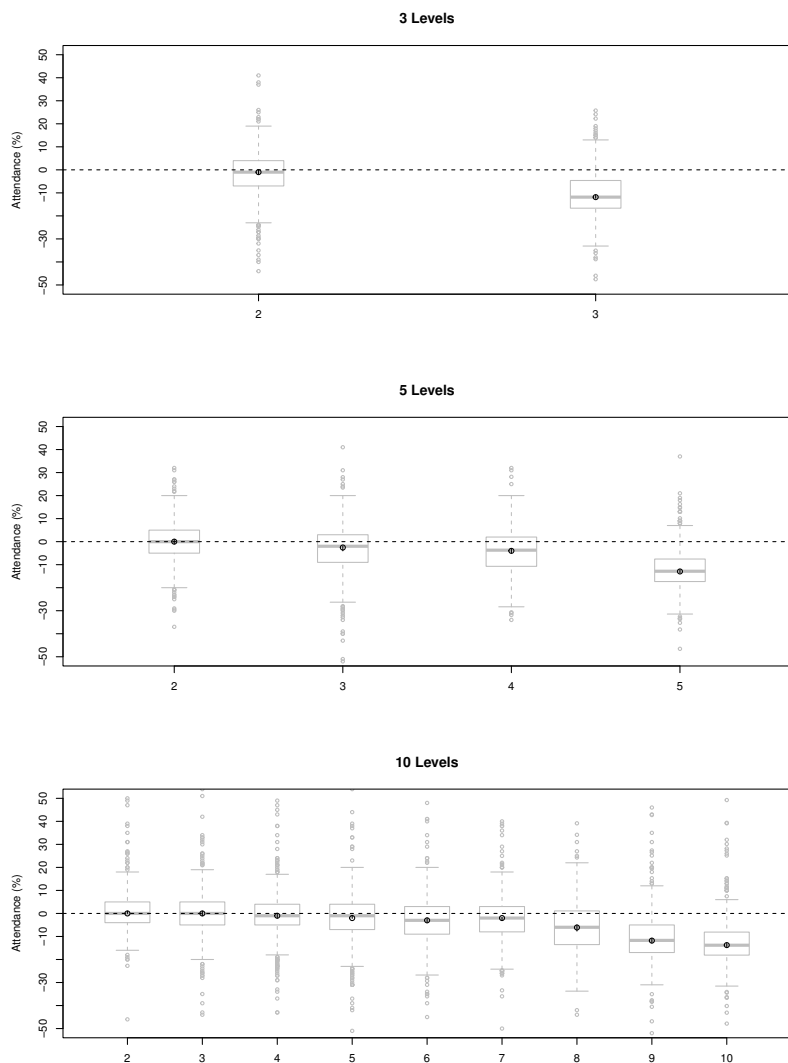| | | | | | | Exposure level | | | | | |
Covariate	Template	1	2	3	4	5	6	7	8	9	10
Gender											
Male	490	490	490	490	490	490	490	490	490	490	490
Female	510	510	510	510	510	510	510	510	510	510	510
School SES											
Low	90	90	90	90	90	90	90	90	90	90	90
Mid-low	318	318	318	318	318	318	318	318	318	318	318
Medium	303	303	303	303	303	303	303	303	303	303	303
Mid-high	174	174	174	174	174	174	174	174	174	174	174
High	115	115	115	115	115	115	115	115	115	115	115
Mother's education											
Primary	321	321	321	321	321	321	321	321	321	321	321
Secondary	434	434	434	434	434	434	434	434	434	434	434
Technical	120	120	120	120	120	120	120	120	120	120	120
College	115	115	115	115	115	115	115	115	115	115	115
Missing	10	10	10	10	10	10	10	10	10	10	10

⋮

5.2 Visualizing effects

One of the advantages of matching methods is that their adjustments are transparent, as illustrated in Table 3. Also, the adjustments are made without looking at the outcomes, which aids the objectivity of the observational study (Rubin 2008). Furthermore, the structure of the data after the adjustments is simple enough that we can analyze the effects by simply taking differences in means and even by plotting the outcomes. This is illustrated in the Figure 2.

Relative to level 1, for each exposure level the figure shows the distribution of matched pair differences in outcomes for both school attendance (Figure 2(a)) and PSU scores (Figure 2(b)). As expected, Figure 2(a) shows that as the exposure level increases, the impact on school attendance becomes more severe. However, this pattern stands in stark contrast to the one in Figure 2(b), where we see no effect of the earthquake on university admission test scores. This is striking, given the magnitude of the earthquake and its impact on school attendance that year. Previous studies have documented a positive effect of school attendance on student achievement as measured by test scores (e.g., Lamdin 1996; Gottfried 2010; Paredes and Ugarte 2011), but this does not appear to be the case in the last year of school in Chile in terms of PSU scores. In the following subsection we estimate the actual effects.

Figure 2: Matched-pair differences in outcomes with respect to level 1

(a) Pair differences for attendance

(b) Pair differences for the PSU scores

5.3 Estimating contrasts

To estimate the effect of the earthquake, we contrast the outcomes in the matched samples under the different levels of exposure to the earthquake. For this, we compute Hodges-Lehmann point estimates as well as 95% confidence intervals following the procedure in Hollander et al. (2015). We compare the outcome of interest for each level $\mathcal{L} \in \mathbb{L}$ with respect to the control level $\mathcal{L}_1$. The null hypothesis H_0 is defined as

$$H_0: \ \tau_u^i = \tau_1^i \text{ if } |R_u^i - R_1^i| < r_{\alpha/2}^{i*}$$

where τ_j^i is the treatment effect for level j in $\mathcal{L}$, R_j^i is the sum of the within-matched group ranks, and $r_{\alpha/2}^{i*}$ is a scalar such that, under the null hypothesis,

$$P_0(|R_u^i - R_1^i| < r_{\alpha/2}^{i*}, u = 2, ..., L) = 1 - \alpha.$$

In this way, by setting an experiment-wise error rate of α, we address the issue of multiple comparisons across different level levels. We obtain the value of $r_{\alpha/2}^{i*}$ under the assumption that under the null hypothesis H_0 all $(L!)^n$ rank combinations are equally likely.

Table 6 in Appendix E in the Supplementary Materials shows the effect estimates for the different exposure levels. The results are consistent with the boxplots in Figure 2. For example, in Table 6(c) with 10 exposure levels, the impact of the earthquake on school attendance increases with the exposure level, having an effect of 3 percentage points for level 6 and of almost 14 points for level 10, both relative to exposure level 1. For exposure levels lower than 6, the effects on school attendance are not statistically significant. In the second column of the table, it is again surprising to see that despite this negative and significant impact on school attendance, the earthquake did not have a significant impact on the PSU test scores, a test that was taken at the end of that school year.

To assess the sensitivity of these findings to hidden biases, we compute Rosenbaum (1987b; 2002; 2015) bounds. These bounds quantify the magnitude that an unobserved covariate would need to have in order to explain away a significant effect and qualitatively change its

interpretation. This magnitude is summarized by the parameter Γ_c in Table 4. We see that for high exposure levels the estimates are quite insensitive to hidden biases. For example, with 10 exposure levels (Table 4(c)) in order to explain away the estimated effect of exposure level 10 relative to level 1 on school attendance, an unobserved covariate that is perfectly associated with this outcome would need to increase the odds of being exposed to level 10 as opposed to level 1 by a factor of 28. For exposure levels 9 and 10, these values are 13.6 and 3.8, respectively. By way of contrast, Hammond's (1964) classical study on the effect of smoking on lung cancer becomes sensitive at $\Gamma_c \approx 6$ (Rosenbaum 2002; Section 4.3.2).

Table 4: Critical Γ (Γ_c) for Sensitivity Analysis

(a) 3 levels

Level of exposure (compared to level 1)	Γ_c
2	1.34
3	10.68

(b) 5 levels

Level of exposure (compared to level 1)	Γ_c
3	1.82
4	2.58
5	17.64

(c) 10 levels

Level of exposure (compared to level 1)	Γ_c
5	1.36
6	1.95
7	1.72
8	3.48
9	9.49
10	18.47

6 Summary and remarks

In this paper, we have proposed a new approach to address three challenges in matching in observational studies. The first challenge relates to handling multi-valued treatments (possibly tens or hundreds of them, either ordered or un-ordered) without estimating the generalized propensity score, directly and flexibly balancing the observed covariates, and facilitating transparent analysis for the outcomes, such as graphical displays. The second challenge relates to building matched samples that are not only balanced but that are repre-

sentative of a population of interest, in such as way that we obtain "representative estimates" of target causal effects. Arguably, this second challenge goes beyond matching and also applies to regression as a method for covariate adjustment for causal inference (see Aronow and Samii 2016). The third challenge relates to matching in larger data sets than usually considered, with hundred of thousands of observations.

To overcome these challenges, instead of simultaneously matching across treatments groups, we separately match each treatment group to a representative random sample of the population of interest. For this, we impose exact distributional or fine balance constraints. This guarantees that the marginal distributions of the matched samples across treatment groups will be identical to each other and to the template in terms of the observed covariates. The effectiveness of the approach relies on the use of a linear-sized MIP formulation, which we show is (i) as strong as much larger quadratic-sized formulations, and (ii) is integral (and hence polynomially solvable) when only two covariates or nested covariates are considered. This integrality property is not preserved for more general covariate structures, but the formulation still retains its practical effectiveness in such settings.

We have used this new matching approach to estimate the impact of an earthquake on the educational achievement of high school students. In particular, we estimated the effect of levels of exposure to the 2010 Chilean earthquake on both school attendance and university admission test scores. We documented negatively increasing effects of the strength of the earthquake on school attendance, but no effect on university admission test scores.

References

Abadie, A. and Imbens, G. W. (2008), "On the failure of the bootstrap for matching estimators," *Econometrica*, 76, 1537–1557.

Achterberg, T. and Wunderling, R. (2013), "Mixed Integer Programming: Analyzing 12 Years of Progress," in *Facets of Combinatorial Optimization: Festschrift for Martin Grötschel*, eds. Jünger, M. and Reinelt, G., Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 449–481.

Aronow, P. M. and Samii, C. (2016), "Does regression produce representative estimates of causal effects?" *American Journal of Political Science*, 60, 250–267.

Baiocchi, M., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2010), "Building a stronger instrument in an observational study of perinatal care for premature infants," *Journal of the American Statistical Association*, 105, 1285–1296.

Balas, E. and Pulleyblank, W. (1983), "The perfectly matchable subgraph polytope of a bipartite graph," *Networks*, 13, 495–516.

Bixby, R. E. (2012), "A brief history of linear and mixed-integer programming computation," *Documenta Mathematica*, 107–121.

CFE-DM (2017), "Chile: Disaster Management Reference Handbook," Tech. rep., Center for Excellence in Disaster Management and Humanitarian Assistance.

Cochran, W. and Rubin, D. (1973), "Controlling bias in observational studies: A review," *Sankhyā: The Indian Journal of Statistics, Series A*, 417–446.

CSN Universidad de Chile (2018), "Registro de Eventos Significativos," http://evtdb.csn.uchile.cl/, accessed: 2018-03-10.

Dehejia, R. and Wahba, S. (1999), "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs," *Journal of the American Statistical Association*, 94, 1053–1062.

DEMRE (2018), "Informe Técnico PSU," Tech. rep., Universidad de Chile.

Diamond, A. and Sekhon, J. S. (2013), "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies," *Review of Economics and Statistics*, 95, 932–945.

Dinkelman, T. and Martinez, C. (2014), "Investing in Schooling in Chile: The Role of Information About Financial Aid for Higher Education," *The Review of Economics and Statistics*, 92, 244–257.

Fukuda, K. (2001), "cddlib reference manual, cddlib Version 0.94j," *ETHZ, Zürich, Switzerland*.

González-Velosa, C., Rucci, G., Sarzosa, M., and Urzúa, S. (2015), "Returns to Higher Education in Chile and Colombia," *IDB Working Paper Series*.

Gottfried, M. (2010), "Evaluating the relationship between student attendance and achieve-

ment in urban elementary and middle schools: An instrumental variables approach," *American Educational Research Journal*, 47, 434–465.

Hammond, E. C. (1964), "Smoking in relation to mortality and morbidity," *Journal of the National Cancer Institute*, 32, 1161–1188.

Hansen, B. (2007), "Flexible, optimal matching for observational studies," *R News*, 7, 18–24.

Hansen, B. B. (2004), "Full matching in an observational study of coaching for the SAT," *Journal of the American Statistical Association*, 99, 609–618.

Hastings, J., Neilson, C., and Zimmerman, S. (2013), "Are Some Degrees Worth More than Others? Evidence from College Admission Cutoffs in Chile," *NBER Working Paper*.

Haviland, A., Nagin, D., and Rosenbaum, P. (2007), "Combining propensity score matching and group-based trajectory analysis in an observational study," *Psychological Methods*, 12, 247.

Hill, J. L. (2011), "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, 217–240.

Hollander, M., A. Wolfe, D., and Chicken, E. (2015), *Nonparametric statistical methods*, A Wiley publication in applied statistics, New York [u.a.]: Wiley.

Imai, K. and Van Dyk, D. A. (2004), "Causal inference with general treatment regimes: Generalizing the propensity score," *Journal of the American Statistical Association*, 99, 854–866.

Imbens, G. W. (2000), "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87, 706–710.

— (2004), "Nonparametric estimation of average treatment effects under exogeneity: A review," *The Review of Economics and Statistics*, 86, 4–29.

— (2010), "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic literature*, 48, 399–423.

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

Imbens, G. W. and Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Joffe, M. and Rosenbaum, P. (1999), "Propensity scores," *American Journal of Epidemiology*, 150, 327–333.

Keele, L., Titiunik, R., and Zubizarreta, J. R. (2015), "Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout," *Journal of the Royal Statistical Society: Series A*, 178, 223–239.

Kern, H. L., Stuart, E. A., Hill, J., and Green, D. P. (2016), "Assessing methods for generalizing experimental impact estimates to target populations," *Journal of research on educational effectiveness*, 9, 103–127.

Lamdin, D. (1996), "Evidence of student attendance as an independent variable in education production functions," *The Journal of Educational Research*, 89, 155–162.

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

Lopez, M. J., Gutman, R., et al. (2017), "Estimation of causal effects with multiple treatments: a review and new ideas," *Statistical Science*, 32, 432–454.

Lu, B. (2005), "Propensity score matching with time-dependent covariates," *Biometrics*, 61, 721–728.

Lu, B., Greevy, R., Xu, X., and Beck, C. (2011), "Optimal nonbipartite matching and its statistical applications," *The American Statistician*, 65, 21–30.

Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. R. (2001), "Matching with doses in an observational study of a media campaign against drug abuse," *Journal of the American Statistical Association*, 96, 21–30.

Mattei, A. and Mealli, F. (2016), "Regression discontinuity designs as local randomized experiments," *Observational Studies*, 66, 156–173.

MDS (2013), "Encuesta CASEN 2011: Módulo Educación," Tech. rep., Ministerio de Desarrollo Social.

Michael, R. G. and David, S. J. (1979), *Computers and intractability: A guide to the theory of NP-completeness.*

MINEDUC (2013), "La Reconstrucción en Educación," Tech. rep., Gobierno de Chile.

Nikolaev, A. G., Jacobson, S. H., Cho, W. K. T., Sauppe, J. J., and Sewell, E. C. (2013), "Balance optimization subset selection (BOSS): an alternative approach for causal inference with observational data," *Operations Research*, 61, 398–412.

OECD (2016), "Income Inequality Update: Income Inequality Remains High in the Face of Recovery," *COPE (Center for Opportunity and Equality).*

Paredes, R. and Ugarte, G. (2011), "Should students be allowed to miss?" *The Journal of Educational Research*, 194–201.

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.

Rosenbaum, P. R. (1987a), "Model-based direct adjustment," *Journal of the American Statistical Association*, 82, 387–394.

— (1987b), "Sensitivity analysis for certain permutation inferences in matched observational studies," *Biometrika*, 74, 13–26.

— (1989), "Optimal matching for observational studies," *Journal of the American Statistical Association*, 84, 1024–1032.

— (2002), *Observational studies*, Springer.

— (2010), *Design of observational studies*, Springer.

— (2015), "Two R packages for sensitivity analysis in observational studies," *Observational Studies*, 1, 1–17.

— (2017), *Observation and experiment: an introduction to causal inference*, Harvard University Press.

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007), "Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer," *Journal of the American Statistical Association*, 102, 75–83.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rosenbaum, P. R. and Silber, J. H. (2001), "Matching and thick description in an observational study of mortality after surgery," *Biostatistics*, 2, 217–232.

Rubin, D. B. (1980), "Randomization analysis of experimental data: the Fisher randomization test comment," *Journal of the American Statistical Association*, 75, 591–593.

— (1986), "Comment: which ifs have causal answers," *Journal of the American Statistical Association*, 81, 961–962.

— (2008), "For objective causal inference, design trumps analysis," *Annals of Applied Statistics*, 2, 808–840.

Sauppe, J. J., Jacobson, S. H., and Sewell, E. C. (2014), "Complexity and approximation results for the balance optimization subset selection model for causal inference in observational studies," *INFORMS Journal on Computing*, 26, 547–566.

Schrijver, A. (2003), *Combinatorial optimization - polyhedra and efficiency*, Springer.

Silber, J. H., Rosenbaum, P. R., Ross, R. N., Ludwig, J. M., Wang, W., Niknam, B. A., Mukherjee, N., Saynisch, P. A., Even-Shoshan, O., Kelz, R. R., et al. (2014), "Template matching for auditing hospital cost and quality," *Health Services Research*, 49, 1446–1474.

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

Tipton, E. (2013), "Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts," *Journal of Educational and Behavioral Statistics*, 38, 239–266.

Torche, F. (2005), "Unequal But Fluid: Social Mobility in Chile in Comparative Perspective," *American Sociological Review*, 70, 422–450.

USGS (2011), "Shakemap 2011," http://earthquake.usgs.gov/earthquakes/shakemap/global/shake/2010tfa

— (2014), "Largest earthquakes in the world since 1900," http://earthquake.usgs.gov/earthquakes/world/10_largest_world.php.

USGS (2018), "Registro de Eventos Significativos," https://earthquake.usgs.gov/earthquakes/browse/large accessed: 2018-03-10.

Vielma, J. P. (2015), "Mixed integer linear programming formulation techniques," *SIAM Review*, 57, 3–57.

Visconti, G. and Zubizarreta, J. R. (2018), "Handling limited overlap in observational studies with cardinality matching," *Observational Studies*, 4, 217–249.

Yang, D., Small, D., 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.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. (2016), "Propensity score matching and subclassification in observational studies with multi-level treatments," *Biometrics*, 72, 1055–1065.

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., Cerdá, M., and Rosenbaum, P. R. (2013a), "Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design," *Epidemiology*, 24, 79–87.

Zubizarreta, J. R. and Keele, L. (2017), "Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system," *Journal of the American Statistical Association*, 112, 547–560.

Zubizarreta, J. R., Kilcioglu, C., and Vielma, J. P. (2018), "designmatch: Matched samples that are balanced and representative by design," *R package version 0.3*, https://cran.r-project.org/package=designmatch.

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," *Annals of Applied Statistics*, 8, 204–231.

Zubizarreta, J. R., Reinke, C. E., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2011), "Matching for several sparse nominal variables in a case-control study of readmission following surgery," *The American Statistician*, 65, 229–238.

Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S. A., and Rosenbaum, P. R. (2013b), "Stronger instruments via integer programming in an observational study of late preterm birth outcomes," *Annals of Applied Statistics*, 7, 25–50.

Zubizarreta, J. R., Small, D. S., and Rosenbaum, P. R. (2014), "Isolation in the construction of natural experiments," *Annals of Applied Statistics*, 8, 2096–2121.

Online Supplementary Materials

Appendix A: Earthquake intensity levels

Using the estimated PGA, we created three measures of exposure to the earthquake: one with three levels, another with five levels, and a final one with ten levels.

We defined the exposure with three intensity levels as follows.

- Low PGA (PGA< 0.08): felt by many people indoors; no buildings received damage.

- Medium PGA ($0.08 \leq$ PGA ≤ 0.25): felt by most or all people indoors; some people were frightened; damages in some (non-resistant) buildings.

- High PGA (PGA> 0.25): many people were frightened; severe shaking; damages in resistant buildings.

For the versions of the exposure with five and ten levels, we divided the students into PGA quintiles and deciles, respectively.

Appendix B: Covariate profiles

Table 5: Covariate profile of the population and template sample

Covariate	Population	Template
Female	0.54	0.51
Indigenous		
Indigenous	0.08	0.07
Missing	0.15	0.14
Father's education		
Secondary	0.39	0.40
Technical	0.09	0.09
College	0.15	0.14
Missing	0.05	0.04
Mother's education		
Secondary	0.41	0.43
Technical	0.13	0.12
College	0.12	0.12
Missing	0.01	0.01
Household income (2008 CL$1000)		
100-200	0.26	0.26
200-400	0.30	0.31
400-600	0.13	0.12
600-1400	0.13	0.14
1400 or more	0.09	0.09
Missing	0.01	0.02
Number of books at home		
1-10	0.19	0.19
11-50	0.46	0.46
51-10	0.16	0.16
More than 100	0.16	0.16
Missing	0.01	0.01

Table 5 (continued): Covariate profile of the population and template sample

Covariate	Population	Template
Student's attendance (deciles)		
2	0.12	0.11
3	0.08	0.09
4	0.10	0.10
5	0.13	0.13
6	0.08	0.08
7	0.09	0.09
8	0.10	0.09
9	0.11	0.10
10	0.15	0.16
Student's GPA 2008 (deciles)		
2	0.11	0.11
3	0.11	0.09
4	0.10	0.11
5	0.10	0.10
6	0.10	0.11
7	0.09	0.09
8	0.10	0.10
9	0.09	0.10
10	0.08	0.08
Student's test scores (deciles)		
2	0.08	0.07
3	0.09	0.10
4	0.09	0.10
5	0.10	0.10
6	0.10	0.11
7	0.11	0.10
8	0.12	0.12
9	0.12	0.11
10	0.12	0.13
Missing	0.01	0.01
School administration		
Public	0.34	0.34
Private subsidized (voucher)	0.55	0.55
Rural school	0.03	0.02
Catholic school	0.24	0.23
School SES		
Mid-low	0.32	0.32
Medium	0.29	0.30
Mid-high	0.18	0.17
High	0.11	0.12
School's test scores (deciles)		
2	0.07	0.07
3	0.09	0.07
4	0.09	0.09
5	0.10	0.11
6	0.11	0.12
7	0.10	0.11
8	0.12	0.12
9	0.12	0.11
10	0.13	0.13

Appendix C: Proofs

To prove Proposition 4.2 we use the following lemma.

Lemma 6.1. *If $P \leq 2$ and $(\boldsymbol{z},\ \boldsymbol{v})$ is an non-integral feasible solution for the LP relaxation of* (1)*, then one of the following conditions holds:*

1. *There exist $\underline{z}$ and $\overline{z}$ such that*

   - $\sum_{\ell \in \mathcal{L}_{p,k}} \underline{z}_\ell = \sum_{\ell \in \mathcal{L}_{p,k}} \overline{z}_\ell = \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell$ *for all $p \in \mathcal{P}$ and $k \in \mathcal{K}(p)$,*

   - $\underline{z}_\ell, \overline{z}_\ell \in [0,1]$ *for all $\ell \in \mathcal{L}$,*

   - $\underline{z} \neq \overline{z}$, *and*

   - $z = (1/2)\underline{z} + (1/2)\overline{z}$.

2. *There exist $q_1, q_2 \in \mathcal{P}$, $u_1 \in \mathcal{K}(q_1)$, $u_2 \in \mathcal{K}(q_2)$, $(\underline{z}, \underline{v})$ and $(\overline{z}, \overline{v})$ such that*

   - $\sum_{\ell \in \mathcal{L}_{p,k}} \underline{z}_\ell = \sum_{\ell \in \mathcal{L}_{p,k}} \overline{z}_\ell = \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell$ *for all $p \in \mathcal{P}$ and $k \in \mathcal{K}(p)$ such that $(p,k) \notin \{(q_1, u_1), (q_2, u_2)\}$,*

   - $\left| \sum_{\ell \in \mathcal{L}_{p,k}} \underline{z}_\ell - N_{p,k} \right| = \underline{v}_{p,k}$ *and* $\left| \sum_{\ell \in \mathcal{L}_{p,k}} \overline{z}_\ell - N_{p,k} \right| = \overline{v}_{p,k}$ *for all $(p,k) \in \{(q_1, u_1), (q_2, u_2)\}$,*

   - $(\underline{z}, \underline{v}) \neq (\overline{z}, \overline{v})$, *and*

   - $(z, v) = \frac{1}{2}(\underline{z}, \underline{v}) + \frac{1}{2}(\overline{z}, \overline{v})$.

Proof. Note that $N_{p,k} \in \mathbb{Z}$ and (z, v) satisfies (3b). If there exist $p \in \mathcal{P}$ and $k \in \mathcal{K}(p)$ for which (z, v) satisfies (3b) as a strict inequality, the let $\varepsilon = v_{p,k} - \left| \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell - N_{p,k} \right|$, $(\underline{z}, \underline{v}) = (\overline{z}, \overline{v}) = (z, v)$ and change $\underline{v}_{p,k}$ to $v_{p,k} - \varepsilon$ and $\overline{v}_{p,k}$ to $v_{p,k} + \varepsilon$ to obtain case 2.

If all inequalities (3b) are satisfied at equality by (z, v), then for each $p \in \mathcal{P}$ and $k \in \mathcal{K}(p)$ we have that either $v_{p,k} \notin \mathbb{Z}_+$ or $|\{\ell \in \mathcal{L}_{p,k} : z_\ell \in (0,1)\}| \geq 2$. Hence, we may construct a sequences $\{s_1, \ldots, s_S\} \subseteq \{1, \ldots, L\}$ and $\{r_1, \ldots, r_{S+1}\} \subseteq \bigcup_{p \in \mathcal{P}} \mathcal{K}(p)$ such that (without loss of generality by possibly interchanging p_1 and p_2):

- $s_j \neq s_l$ and $r_j \neq r_S$ for all $j, l \in \{1, \ldots, S\}$ with $j \neq l$,

- $z_{\ell_{s_j}} \in (0,1)$ for all $j \in \{1, \ldots, S\}$,

- $r_j \in \mathcal{K}\left(p_{h(j)}\right)$ for all $j \in \{1, \ldots, S+1\}$ where

$$h(j) = 2 - j + 2\lfloor j/2 \rfloor = \begin{cases} 1 & j \text{ is odd} \\ 2 & j \text{ is even,} \end{cases}$$

- $\ell_{s_j} \in \mathcal{L}_{p_{h(j)}, r_j}$ and $\ell_{s_j} \in \mathcal{L}_{p_{h(j+1)}, r_{j+1}}$ for all $j \in \{1, \ldots, S\}$,

and either

$$S \text{ is even, } h(1) = h(S+1) \text{ and } r_1 = r_{S+1} \tag{4}$$

or

$$v_{p_{h(1)},r_1}, v_{p_{h(S+1)},r_{S+1}} \in (0,1), \tag{5a}$$

$$\left\{ \ell \in \mathcal{L}_{p_{h(1)},r_1} : z_\ell \in (0,1) \right\} = \left\{ z_{\ell_{s_1}} \right\}, \tag{5b}$$

$$\left\{ \ell \in \mathcal{L}_{p_{h(S+1)},r_{S+1}} : z_\ell \in (0,1) \right\} = \left\{ z_{\ell_{s_S}} \right\}. \tag{5c}$$

If (4) holds let $\varepsilon = \min_{i=1}^{S} \left\{ z_{\ell_{s_j}}, 1 - z_{\ell_{s_j}} \right\}$, $\underline{z}$ and $\overline{z}$ be such that

$$\underline{z}_\ell = \begin{cases} z_\ell - \varepsilon & \text{if } \ell = s_j \text{ for an odd } j \in S, \\ z_\ell + \varepsilon & \text{if } \ell = s_j \text{ for an even } j \in S, \\ z_\ell & \text{otherwise} \end{cases}$$

$$\overline{z}_\ell = \begin{cases} z_\ell + \varepsilon & \text{if } \ell = s_j \text{ for an odd } j \in S, \\ z_\ell - \varepsilon & \text{if } \ell = s_j \text{ for an even } j \in S, \\ z_\ell & \text{otherwise,} \end{cases}$$

to obtain case 1.

If instead (5) holds let $q_1 = p_{h(1)}$, $q_2 = p_{h(S+1)}$, $u_1 = r_1$, $u_2 = r_{S+1}$,

$$\varepsilon = \min \left\{ \min_{i=1}^{S} \left\{ z_{\ell_{s_j}}, 1 - z_{\ell_{s_j}} \right\}, v_{p_{h(1)},r_1}, v_{p_{h(S+1)},r_{S+1}} \right\}$$

$\underline{z}$ and $\overline{z}$ be as defined previously with this new ε. Then, let $\underline{v}$ and $\overline{v}$ be such that

$$
\underline{v}_{p,k} = \begin{cases}
v_{p,k} + \varepsilon & \text{if } p = p_{h(1)},\ k = r_1 \text{ and } N_{p,k} - \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell > 0 \\
v_{p,k} - \varepsilon & \text{if } p = p_{h(1)},\ k = r_1 \text{ and } N_{p,k} - \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell < 0 \\
v_{p,k} - \varepsilon & \text{if } p = p_{h(S+1)},\ k = r_{S+1},\ S \text{ is even and } N_{p,k} - \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell > 0 \\
v_{p,k} + \varepsilon & \text{if } p = p_{h(S+1)},\ k = r_{S+1},\ S \text{ is even and } N_{p,k} - \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell < 0 \\
v_{p,k} + \varepsilon & \text{if } p = p_{h(S+1)},\ k = r_{S+1},\ S \text{ is odd and } N_{p,k} - \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell > 0 \\
v_{p,k} - \varepsilon & \text{if } p = p_{h(S+1)},\ k = r_{S+1},\ S \text{ is odd and } N_{p,k} - \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell < 0 \\
v_{p,k} & \text{otherwise}
\end{cases}
$$

$$
\overline{v}_{p,k} = \begin{cases}
v_{p,k} - \varepsilon & \text{if } p = p_{h(1)},\ k = r_1 \text{ and } N_{p,k} - \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell > 0 \\
v_{p,k} + \varepsilon & \text{if } p = p_{h(1)},\ k = r_1 \text{ and } N_{p,k} - \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell < 0 \\
v_{p,k} + \varepsilon & \text{if } p = p_{h(S+1)},\ k = r_{S+1},\ S \text{ is even and } N_{p,k} - \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell > 0 \\
v_{p,k} - \varepsilon & \text{if } p = p_{h(S+1)},\ k = r_{S+1},\ S \text{ is even and } N_{p,k} - \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell < 0 \\
v_{p,k} - \varepsilon & \text{if } p = p_{h(S+1)},\ k = r_{S+1},\ S \text{ is odd and } N_{p,k} - \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell > 0 \\
v_{p,k} + \varepsilon & \text{if } p = p_{h(S+1)},\ k = r_{S+1},\ S \text{ is odd and } N_{p,k} - \sum_{\ell \in \mathcal{L}_{p,k}} z_\ell < 0 \\
v_{p,k} & \text{otherwise}
\end{cases}
$$

to obtain case 2. $\qquad\square$

Proof of Proposition 4.2. For case 1 we prove the result for $P = 2$. The result follows for $P = 1$ by copying the covariate as the result for $P = 2$ does not require that the covariates are different. Furthermore, because of Proposition 4.1 it suffices to prove the result for (3).

Assume for a contradiction that the LP relaxation of (3) has a non-integral extreme point (z, v). Then by Lemma 6.1 there exist $(\underline{z}, \underline{v})$ and $(\overline{z}, \overline{v})$ that are feasible for the LP relaxation of (3), $(\underline{z}, \underline{v}) \neq (\overline{z}, \overline{v})$ and $(z, v) = \frac{1}{2}(\underline{z}, \underline{v}) + \frac{1}{2}(\overline{z}, \overline{v})$, which contradicts (z, v) being an extreme point.

For case 2 we only need to prove the result for $P \geq 3$. Again, assume for a contradiction that the LP relaxation of (3) has a non-integral extreme point (z, v). If we restrict this v to the last two covariates to we can apply Lemma 6.1 to obtain $(\underline{z}, \underline{v})$ and $(\overline{z}, \overline{v})$ (with $\underline{v}/\overline{v}$ restricted to the last two covariates) such that $(\underline{z}, \underline{v}) \neq (\overline{z}, \overline{v})$ and $(z, v) = \frac{1}{2}(\underline{z}, \underline{v}) + \frac{1}{2}(\overline{z}, \overline{v})$ (with $v/\underline{v}/\overline{v}$ restricted to the last two covariates). Furthermore, $(\underline{z}, \underline{v})$ and $(\overline{z}, \overline{v})$ satisfy all the constraints of LP relaxation of (3) except for constraints (3b) associated to the first $P - 2$ covariates.

If $(\underline{z}, \underline{v})$ and $(\overline{z}, \overline{v})$ came from case 1 Lemma 6.1 we extend $\underline{v}/\overline{v}$ to all covariates by letting $\underline{v}_{p,k} = \overline{v}_{p,k} = v_{p,k}$ for all $p \in \{p_1, \ldots, p_{P-2}\}$ and $k \in \mathcal{K}(p)$. This extension is feasible for the complete LP relaxation of (3) because $\sum_{\ell \in \mathcal{L}_{p_{P-1},k}} \underline{z}_\ell = \sum_{\ell \in \mathcal{L}_{p_{P-1},k}} \overline{z}_\ell = \sum_{\ell \in \mathcal{L}_{p_{P-1},k}} z_\ell$ for all $k \in \mathcal{K}(p_{P-1})$, and because by assumption for all $p \in \{p_{P-1}, p_P\}$, $k \in \mathcal{K}(p)$ and $i \in \{1, \ldots, P-2\}$ there exist $k' \in \mathcal{K}(p_i)$ such that $\mathcal{L}_{p,k} \subseteq \mathcal{L}_{p_i,k'}$.

If $(\underline{z}, \underline{v})$ and $(\overline{z}, \overline{v})$ came from case 2 Lemma 6.1, let $q_1, q_2 \in \{p_{P-1}, p_P\}$, $u_1 \in \mathcal{K}(q_1)$, $u_2 \in \mathcal{K}(q_2)$ from the Lemma be such that $\left| \sum_{\ell \in \mathcal{L}_{p,k}} \underline{z}_\ell - N_{p,k} \right| = \underline{v}_{p,k}$ and $\left| \sum_{\ell \in \mathcal{L}_{p,k}} \overline{z}_\ell - N_{p,k} \right| = \overline{v}_{p,k}$ for all $(p,k) \in \{(q_1, u_1), (q_2, u_2)\}$. Then if we extend $\underline{v}/\overline{v}$ be letting

$$
\underline{v}_{p,k} = \overline{v}_{p,k} = \begin{cases} v_{q_1,u_1} & \text{if } k \in \mathcal{K}(p) \text{ is such that } \mathcal{L}_{q_1,u_1} \subseteq \mathcal{L}_{p,k} \\ v_{q_2,u_2} & \text{if } k \in \mathcal{K}(p) \text{ is such that } \mathcal{L}_{q_2,u_1} \subseteq \mathcal{L}_{p,k} \\ v_{p,k} & \text{otherwise} \end{cases}
$$

we again obtain an extension that is feasible for the complete LP relaxation of (3).

In both cases we end up with a contradiction of (z, v) being an extreme point. $\qquad \square$

Proof of Lemma 4.3. Let

$$
\begin{aligned}
\mathbf{x}_1 &= (k_1, k_1, k_1), & \mathbf{x}_2 &= (k_3, k_3, k_3), \\
\mathbf{x}_1 &= (k_1, k_2, k_3), & \mathbf{x}_2 &= (k_3, k_2, k_1), \\
\mathbf{x}_1 &= (k_2, k_1, k_2), & \mathbf{x}_2 &= (k_2, k_3, k_2),
\end{aligned}
$$

$T = 3$, $L = 6$ and $N_{p,k} = 1$ for all p and k. The feasible region of the LP relaxation of (3) for this case is given by

$$
\begin{aligned}
|z_{\ell_1} + z_{\ell_3} - 1| \le v_{p_1,k_1}, \quad & |z_{\ell_5} + z_{\ell_6} - 1| \le v_{p_1,k_2}, \quad & |z_{\ell_2} + z_{\ell_4} - 1| \le v_{p_1,k_3}, \\
|z_{\ell_1} + z_{\ell_5} - 1| \le v_{p_2,k_1}, \quad & |z_{\ell_3} + z_{\ell_4} - 1| \le v_{p_2,k_2}, \quad & |z_{\ell_2} + z_{\ell_6} - 1| \le v_{p_2,k_3} \\
|z_{\ell_1} + z_{\ell_4} - 1| \le v_{p_3,k_1}, \quad & |z_{\ell_5} + z_{\ell_6} - 1| \le v_{p_3,k_2}, \quad & |z_{\ell_2} + z_{\ell_3} - 1| \le v_{p_3,k_3} \\
\sum_{i=1}^{6} z_{\ell_i} = 1, \quad & 0 \le z_{\ell_i} \le 1 \quad & \forall i \in \{1, \ldots, 6\}.
\end{aligned}
$$

Using CDDLib (Fukuda 2001) we can check that this LP has 11 fractional extreme points out of a total of 31. In particular, $z_{\ell_i} = 1/2$ for all $i \in \{1, \ldots, 6\}$ and $v_{p,k} = 0$ for all p and k is one such fractional extreme point. $\qquad \square$

6

Appendix D: Covariate balance

Figure 3: Standardized differences in means in covariates before and after matching for 10 levels of exposure.
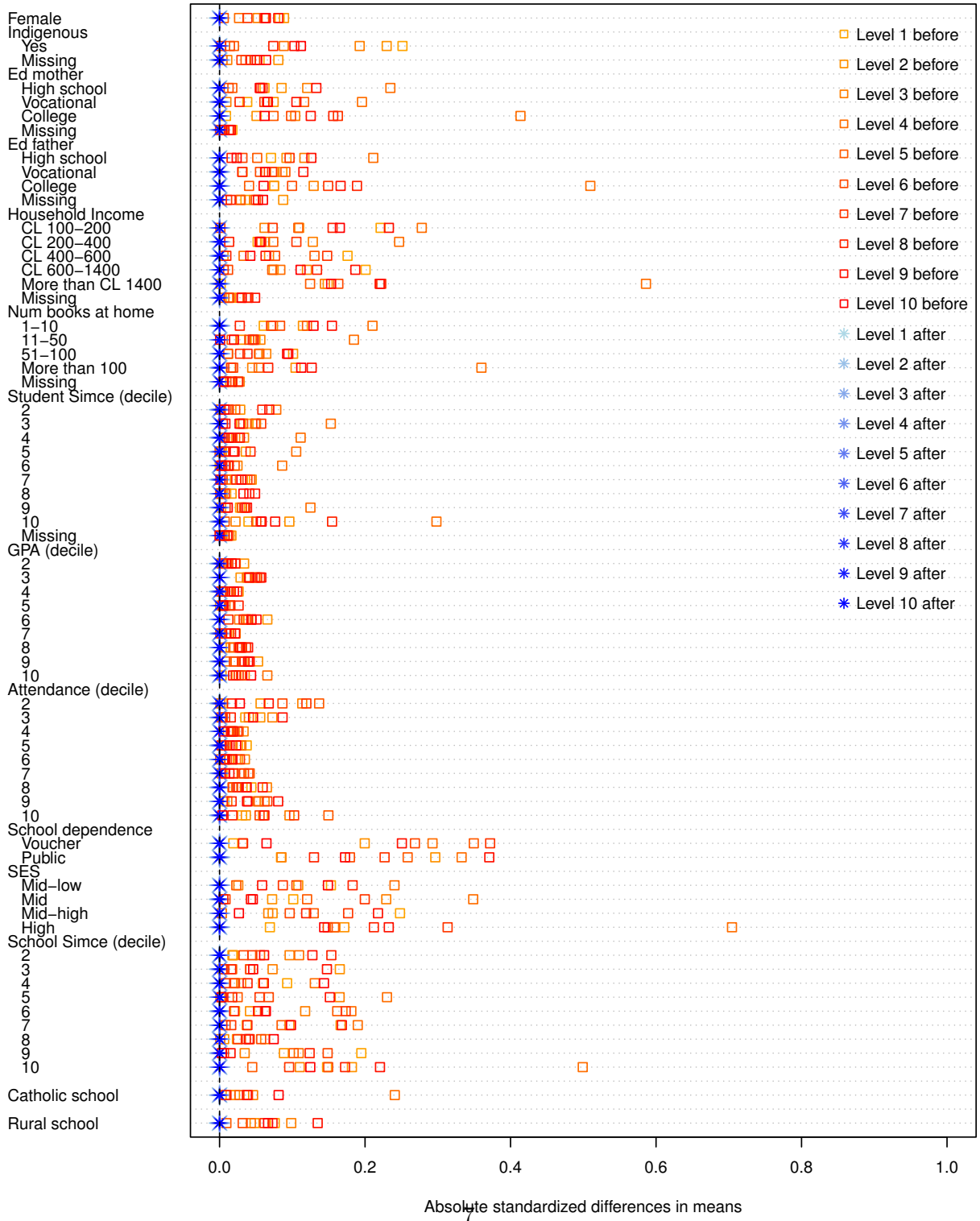
Appendix E: Effect estimates

Table 6: Effect estimates and 95% confidence intervals for different levels of exposure to the earthquake. The point estimates are contrasts with respect to exposure level 1. The 95% confidence account for multiple comparisons.

(a) 3 exposure levels

Exposure level	Attendance (%)	PSU score
2	-1.00	11.00
	[-2.01,0.00]	[1.90,18.50]
3	-11.85	8.00
	[-12.35,-11.00]	[1.50,14.10]

(b) 5 exposure levels

Exposure level	Attendance (%)	PSU score
2	-0.00	6.00
	[-1.00,1.00]	[-0.60,14.00]
3	-2.55	6.50
	[-4.00,-1.99]	[-0.10,13.50]
4	-4.00	4.50
	[-5.01,-3.00]	[-3.00,11.00]
5	-12.95	8.50
	[-13.70,-12.09]	[0.90,15.10]

(c) 10 exposure levels

Exposure level	Attendance (%)	PSU score
2	0.00	4.00
	[-0.01,1.01]	[-3.50,11.60]
3	0.00	9.00
	[-1.00,1.01]	[0.90,17.10]
4	-1.00	10.00
	[-2.00,0.01]	[2.00,17.60]
5	-2.00	9.00
	[-2.16,-0.99]	[0.90,16.50]
6	-3.00	7.50
	[-4.01,-2.00]	[-0.50,14.50]
7	-2.00	-3.50
	[-3.01,-1.69]	[-3.60,11.00]
8	-6.10	-4.00
	[-7.66,-5.00]	[-4.10,11.10]
9	-11.80	9.00
	[-12.85,-10.80]	[1.50,17.50]
10	-13.75	10.50
	[-14.56,-12.94]	[3.0,18.60]