Optimal Matching for Observational Studies that Integrate Quantitative and Qualitative Research

Ruoqi Yu, Dylan S. Small, David Harding, José Aveldanes, Paul R. Rosenbaum[1]

University of Pennsylvania and University of California at Berkeley

Abstract. A quantitative study of treatment effects may form many matched pairs of a treated subject and an untreated control who look similar in terms of covariates measured prior to treatment. When treatments are not randomly assigned, the inevitable concern is that individuals who look similar in measured covariates may be dissimilar in unmeasured covariates. An existing proposal entails interviewing a small subset of these pairs, hoping in a few cases to observe quite a bit more of what was not quantitatively measured. A few pairs cannot rule out subtle biases that materially affect analyses of many pairs, but perhaps a few pairs can inform discussion of such biases, perhaps leading to a reinterpretation of quantitative data, or perhaps raising new considerations and perspectives. The large literature on qualitative research contends that open-ended, ethnographic, thick, narrative descriptions of a few people can be informative. Here, we discuss and apply a form of optimal matching that supports such an integrated, quantitative-plus-qualitative study. The optimal match provides many closely matched pairs plus a small subset of exceptionally close pairs suitable for interviews. We illustrate the matching technique using data from the National Study of Youth and Religion that combined quantitative measures and some detailed interviews.

Keywords: Causal inference; optimal matching; threshold algorithms; thick description.

1 Strengthening causal inference by integrating methodologies

1.1 Effects caused by treatments

The effect of a treatment on an individual is a comparison of two potential responses, the individual's response if assigned to treatment and the individual's response to control; see Rubin (1974). This causal effect is not observed: we see the response of an individual under treatment or control, not both, so we do not see the effect of the treatment. Inference about

[1] *Address for correspondence:* Department of Statistics, The Wharton School, University of Pennsylvania, Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 USA. E-mail: dsmall@wharton.upenn.edu. 8 April 2020.

the effects caused by treatments is comparatively straightforward in randomized trials that use random numbers to divide a finite population into treated and control groups; see Fisher (1935). Inference about causal effects is much more difficult when treatments are not randomly assigned, because in the absence of randomization, treated and control groups may differ prior to treatment in terms of covariates that were not observed.

Mervyn Susser wrote that if evidence of cause and effect suffers from certain limitations that are not a consequence of a limited sample size, then we should not try to dispel these limitations by increasing the sample size or exactly replicating the study; rather, we should seek evidence that suffers from different limitations. He wrote:

> The epidemiologist ... seeks ... consistency of results in a variety of repeated tests. ... Consistency is present if the result is not dislodged in the face of diversity in times, places, circumstances, and people, as well as of research design (1987, p. 88) ... The strength of the argument rests on the fact that diverse approaches produce similar results (1973, pp. 148).

For related discussion, see Rosenbaum (2001) and Munafò and Davey-Smith (2018).

A basic approach combines qualitative and quantitative research, each having certain limitations not shared by the other. Although we believe that the integration of qualitative and quantitative research is useful, our goal here is not to argue for that claim, but rather to illustrate a new matching technique that facilitates such an integration.

1.2 Pairing many people and interviewing a few pairs

Among methods of adjustment for covariates, matching leaves people intact as people. Intact people can be met, interviewed, recorded, quoted, described in narrative terms. Do people who look comparable in measured covariates still seem broadly comparable when met, interviewed, quoted and described?

Rosenbaum and Silber (2001) argued that it is natural to strengthen a matched quantitative comparison of many pairs by adding qualitative description of a few matched pairs. Their full study matched 830 patients who had died following surgery to patients who survived, examining these case-control pairs. The matching used quantitative data from medical chart abstraction in an effort to compare patients similar prior to hospital admission. In a pilot study, they compared, for 38 pairs: (i) quantitative data obtained by chart abstraction, and (ii) a detailed reading of the medical charts for these pairs. This comparison led to improved use of the quantitative data — revised definitions of cancer and congestive heart failure — and a new match incorporating these revisions. The side-by-side reading of charts for matched pairs was critical: unremarkable, accurately abstracted charts were revealed to be poorly matched only by seeing that the paired charts described people who differed in notable respects.

For related discussion of purposeful, comparative selection of cases for qualitative comparisons, see Seawright and Gerring (2008) and Tarrow (2010).

2 Optimal matching for quantitative plus qualitative comparisons

2.1 Match many people, and produce a few very close pairs for qualitative comparison

How should we match if we want a quantitative comparison of many, say T, pairs, and a thick description of a few, say $F \geq 1$, pairs? For instance, T might be hundreds or thousands or tens of thousands of pairs, while F might be ten pairs. Because considerable time and effort will be expended upon the F pairs, we want these few pairs to be exceptionally close in terms of pretreatment covariates. We want to judge the adequacy of the observed covariates using pairs that are exceptionally close in terms of the observed covariates.

In a matched sample, covariates are balanced if the distribution of age, say, is almost the same among the T treated individuals and their T matched controls, and the same

is true of the distribution of income, and so on. In contrast, a single pair is close if the two people in the pair have nearly the same age, the same income, and so on. Modern matching methods often use propensity scores, fine balance constraints, and similar devices to balance many covariates in treated and control groups, but they do this with pairs that need not be close person by person. For instance, it is often possible to balance, say, 30 binary covariates, but these define 2^{30} or about a billion categories of people, so it is not possible to find pairs that are uniformly close on 30 binary covariates.

We seek a match for T pairs that balances covariates, but additionally we seek a small number F of exceptionally close pairs for interviews. The method uses a threshold technique from Rosenbaum (2017a), which develops an idea of Garfinkel (1971). The method and the example are in an `R` package `thickmatch`.

There are T treated individuals, $\tau_1, \ldots, \tau_T$, and $C \geq T$ potential controls, $\gamma_1, \ldots \gamma_C$. Each treated individual will be paired with a different control, yielding T matched sets consisting of $2T$ distinct individuals. There is a distance $\delta_{tc} \geq 0$ between treated individual τ_t and potential control γ_c, based on their observed covariates, so $\delta_{tc} = 0$ if these two people look identical on all covariates. The exact form of the distance does not matter in the discussion that follows. Commonly, this distance is some form of Mahalanobis distance focused on important covariates, with a caliper on the propensity score computed from all covariates, and it may incorporate other considerations. An optimal or minimum-distance match $\mu(\cdot)$ assigns each τ_t to a different control, specifically to control $c = \mu(\tau_t)$, forming T pairs, $\left\{\tau_t, \gamma_{\mu(\tau_t)}\right\}$, in such a way that the sum of the T within-pair distances, $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$, is minimized. Various algorithms (e.g., Bertsekas 1981) implemented in various `R` packages solve this problem. There are $C!/(C-T)!$ possible pairings $\mu(\cdot)$ — an enormous number — but algorithms exist that can find the best $\mu(\cdot)$ in $O\left(C^3\right)$ arithmetic steps; so, it is entirely practical.

Commonly, there are many covariates, including continuous covariates, with the consequence that the $T \times C$ distances δ_{tc} are all distinct. For instance, if one of the covariates had a Normal distribution and the Mahalanobis distance was used, then the probability that at least two δ_{tc} take the same numerical value is zero. The procedure is slightly easier to describe if the δ_{tc} are untied, so we assume this, discussing the minor consequences of a few ties in §2.3.

Let $\mathcal{I} \subseteq \{\tau_1, \ldots, \tau_T\}$ be a subset of the treated individuals, perhaps most commonly all of them, $\mathcal{I} = \{\tau_1, \ldots, \tau_T\}$. For thick description, we want F treated individuals in $\mathcal{I}$ to be exceptionally well matched, with small distances to their matched controls, and subject to doing that, we want to minimize the total of all T within-pair distances, $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$. This means that the total distance $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$ will typically be a little larger than the minimum in the previous paragraph, but these F pairs will be very close.

Any match $\mu(\cdot)$, good or bad, pairs the individuals in $\mathcal{I}$ to distinct controls, and we will interview or otherwise thickly describe the $2F$ individuals in these F closest pairs. So, for any match $\mu(\cdot)$, good or bad, let Δ_μ be the Fth largest within-pair distance, $\delta_{t,\mu(\tau_t)}$, for the treated individuals in $\tau_t \in \mathcal{I}$, if we use match $\mu(\cdot)$. So Δ_μ is the Fth quantile of distances $\delta_{t,\mu(\tau_t)}$ individuals $t \in \mathcal{I}$ if we opt for match $\mu(\cdot)$. If $\mathcal{I} = \{\tau_1, \ldots, \tau_T\}$ and $F = 10$, then Δ_μ is the single largest of the $F = 10$ smallest distances in the match, the largest distance for the $F = 10$ pairs that will be thickly described. We want a match $\mu(\cdot)$ that minimizes this quantile Δ_μ, and among all matches that do that, we want a match that minimizes the total distance over all pairs, $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$. This is a constrained version of the original optimization problem: minimize $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$, but subject to the constraint that Δ_μ is as small as possible.

2.2 The algorithm

The solution first determines the smallest possible Δ_μ by a threshold algorithm similar to that of Garfinkel (1971). Then, knowing the best Δ_μ, the algorithm solves the optimal matching problem with a revised, penalized distance.

We start with a guess κ of Δ_μ, perhaps quite a poor guess. We create new distances $\delta^{'}_{tc}$ where $\delta^{'}_{tc} = 1$ if $\delta_{tc} > \kappa$ and $\delta^{'}_{tc} = 0$ if $\delta_{tc} \leq \kappa$. We find an optimal matching to minimize $\sum_{t=1}^{T} \delta^{'}_{t,\mu(\tau_t)}$. If this minimum is $\sum_{t=1}^{T} \delta^{'}_{t,\mu(\tau_t)} \leq T - F$, then we found at least F pairs at distance of at most κ, so we revise κ to be a smaller number and try again. If this minimum is $\sum_{t=1}^{T} \delta^{'}_{t,\mu(\tau_t)} > T - F$, then it is impossible to find F distinct pairs with a distance of at most κ, so we revise κ to be a larger number and try again. A binary search quickly finds a close upper bound for Δ_μ.

Let β be a large number, called a penalty. Define new distances $\delta^{''}_{tc} = \delta_{tc}$ if $\delta_{tc} \leq \Delta_\mu$ and $\delta^{''}_{tc} = \delta_{tc} + \beta$ if $\delta_{tc} > \Delta_\mu$. We then obtain our desired match $\mu(\cdot)$ by finding a minimum distance match for the new distances $\delta^{''}_{tc}$. For large enough β, the algorithm finds a match that minimizes the original distance $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$ subject to the constraint that we have F pairs with distances of at most Δ_μ, where Δ_μ is as small as possible; see Rosenbaum (2017a, Proposition 1).

Our threshold algorithm for quantile-constrained optimization has been described for matched pairs, but there are many other matched designs, such as matching two controls to each treated individual. A parallel approach may be taken when matching with multiple controls and related designs, such as the designs discussed by Hansen (2004) and Pimentel et al. (2015). In this case, one seeks to interview F distinct treated subjects and one very close control for each of them in F distinct match sets that may also contain other controls. The `thickmatch` package in `R` implements this extension of the method, and the example in §3 matches to two controls. See §2.4 for additional features of the `thickmatch` package

useful in other problems. For a different application of threshold matching algorithms in observational studies, see Yu et al. (2020).

2.3 Tied distances

Ties among the distances, δ_{tc}, require a slightly more precise statement of the optimization problem that the algorithm solves. When ties are few, the two problems differ negligibly from a practical point of view. With or without ties among the δ_{tc}, the algorithm maximizes the number of pairs with distances at most equal to the Fth quantile, Δ_μ, and minimizes $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$ subject to that constraint. If the distances are never tied, then there are, by definition, exactly F distances below the Fth quantile, Δ_μ. If there are ties at the Fth quantile, then the algorithm might produce $F+1$ or $F+2$ pairs below the Fth quantile, Δ_μ, so that several pairs have distances of exactly Δ_μ.

Suppose that the investigator wanted to interview $F = 10$ pairs, but some distances δ_{tc} were equal. The algorithm might produce 8 untied and 3 tied pairs with $\delta_{tc} \leq \Delta_\mu$, where three pairs were tied with $\delta_{tc} = \Delta_\mu$. So far as covariate distances are concerned, these three pairs are equivalent. The investigator would then interview all 8 pairs with $\delta_{tc} < \Delta_\mu$ and two of the three pairs with $\delta_{tc} = \Delta_\mu$, making ten pairs in total.

2.4 Aspects of implementation in `thickmatch` relevant to other data sets

The `thickmatch` package in `R` has quite a few additional standard features of modern matching. We mention these only briefly as they are reviewed elsewhere (Rosenbaum 2020a). It is important that the `thickmatch` package integrates these features, because forcing a few pairs to be extremely close for qualitative comparison must be part of, must be compatible with, the entire match that is intended for quantitative analysis.

Constraints may be imposed on a matching algorithm to ensure that the match has

certain desired properties. Often, much of the work in achieving a satisfactory match is done by the constraints, not by closely pairing individuals in terms of a distance, δ_{tc}. For qualitative comparison of a few pairs, small distances are helpful for those few pairs. However, for quantitative comparison of treated and control groups, it is the comparability of the groups, not close pairs, that matters most, and constraints often target comparable groups.

Constraints may be imposed either as hard or soft constraints. If a hard constraint cannot be achieved, then the problem is infeasible, and no match is produced. If a soft constraint is imposed instead, then a match is always produced, but if necessary it may slightly violate the constraints. For instance, the `thickmatch` package implements both exact and near-exact matching for nominal covariates. Exact matching for gender would fail if the number of control women was one woman short of what is needed for exact matching, while near-exact matching would produce a match with one pair in which a man was matched to a woman. Soft constraints are typically implemented by adjusting the objective function, say $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$, so that violations of a constraint are severely penalized by dramatically increasing the objective function's value. Also, the `thickmatch` package provides the option of a caliper, implemented as a soft constraint, on a score such as the propensity score, thereby minimizing the number of times the caliper is violated.

Fine balance for a nominal covariate occurs if the marginal distribution of the nominal covariate is the same in the treated and control groups, ignoring whether treated individuals are paired to controls for this nominal covariate. Fine balance may be implemented as a hard constraint, but may then generate infeasibility. Near-fine balance comes as close to fine balance as the data will allow, and it is implemented as a soft-constraint. The `thickmatch` package implements near-fine balance for pair matching and for matching with a fixed number of controls. To implement fine balance or near-fine balance with a

variable number of controls, the method of Pimentel et al. (2015) should be used instead.

The `thickmatch` package was conceived to implement modern matching techniques while generating a few, F, pairs that are exceptionally close, so that qualitative comparison of these few pairs might be particularly informative. However, nothing in the implementation requires F to be small or the corresponding pairs to be exceptionally close; rather, F can be most or all of the treated group, and the `thickmatch` package can then avoid large distances for many or all pairs rather than produce a few very close pairs; see Rosenbaum (2017a).

There are several soft constraints in the `thickmatch` package, each implemented as a penalty on the objective function. The package gives the user control over the relative size of these penalties, so the user sets the priorities when some constraints cannot be satisfied. For instance, the user can say that the caliper on the propensity score is more important than fine balance.

3 An example: substance use by working and nonworking adolescents

3.1 Background: Is working good for adolescents?

There has been controversy about whether substantial adolescent employment is healthy. Greenberger and Steinberg (1986) raised concerns that employment during high school has negative consequences; in particular, they found that adolescent workers used more illicit substances than adolescent nonworkers, including alcohol and marijuana. Longest and Shanahan (2007) conducted an observational study of the effect of adolescent work on substance use that controlled for sociodemographic features. They used data from the National Survey of Youth and Religion, a nationally representative telephone survey of 3290 U.S. teenagers, aged 13-17 years, and their parents conducted from July 2002 to August 2003 (Smith and Pearce, 2019). Longest and Shanahan found that after controlling for

sociodemographic features (family income, family structure (two biological parents, two parents including a nonbiological parent, single parent), highest parents' education in the household, gender, age and race/ethnicity), adolescent working was associated with higher substance use.

To illustrate our methodology, we conduct an observational study that builds on Longest and Shanahan's, using the National Survey of Youth and Religion and controlling for the same sociodemographic variables as they did.

3.2 Matching to combine quantitative and qualitative comparisons

We compared adolescents who worked at least 20 hours per week (briefly treated, $n = 268$) to those who did not work at all (briefly controls, $n = 2548$). We matched 1-to-2, yielding 268 matched sets, 268 treated and $2 \times 268 = 536$ matched controls. We matched for family income, whether one or two parents were present at home, the highest education of the parents in the household (<HS, HS, BA, or >BA), gender age, and race/ethnicity (black, Hispanic, other) of the adolescent, whether he or she had dropped out of school, indicators for missing income and education, and a propensity score built from these variables.

As in §2, we demanded ten exceptionally close pairs for thick description, and for the rest we did a minimum distance match within calipers on the propensity score. Because of ties in the distances, we obtained 12 rather than 10 close pairs, as listed in Table 1. Notably, these 12 pairs are extremely close on the covariates controlled by matching, so it would be of interest to compare them in direct interviews.

Figures 1 and 2 show the balance on covariates for all three groups, the 268 treated, the 536 matched controls and the $2548 - 536 = 2012$ unmatched controls. Generally, groups T and C were fairly close on covariates after matching. The unmatched group U was typically a few years younger, was more often Hispanic, contained fewer dropouts, and

somewhat more missing family incomes.

3.3 Comparing outcomes, alcohol and marijuana use

Figure 3 depicts self-reported alcohol and marijuana use for matched working (T) and nonworking groups (C). Neither group reports heavy use, but the working group reports considerably more use than nonworking group.

Could the difference in substance use seen in Figure 3 readily be explained as a small imbalance in an unobserved covariate that matching did not control? A small imbalance would not produce Figure 3, as we now explain in quantitative terms. There are two outcomes, alcohol and marijuana, so a comparison must control for testing more than one hypothesis. We used the method in Rosenbaum (2020b) to control the family-wise error rate when combining the two scaled outcomes with equal weights while simultaneously considering arbitrary comparisons using Scheffé projections. The planned, equally weighted comparison would reject the hypothesis of no effect, with a P-value of ≤ 0.03, even in the presence of a substantial bias from an unobserved covariate that increased the odds of substance use five-fold and increased the odds of working 20+ hours a week by three-fold ($\Gamma = 2$, corresponding with $\Delta = 5$ and $\Lambda = 3$ via $\Gamma = (\Delta\Lambda + 1) / (\Delta + \Lambda)$; see Rosenbaum (2017b, Table 9.1). Still controlling the family-wise error rate, we may employ closed-testing (Marcus et al. 1976) to follow rejection in the joint test by individual testing of the two separate outcomes. Taken separately, the effects on alcohol and marijuana use are sensitive to somewhat smaller biases, but are insensitive to an unobserved covariate that would triple the odds of greater use and triple the odds of working 20+ hours ($\Gamma = 1.75$, corresponding with $\Delta = 3.4$ and $\Lambda = 3$). (Specifically, these comparisons used the `comparison`, `planScheffe` and `amplify` functions, with default settings, in the `sensitivitymult` package in `R`.)

4 Thick description of a few pairs

To illustrate thick description of a few pairs, we compared drug use by interviewed working and nonworking adolescents in the National Survey of Youth and Religion (NSYR). Our matching method is intended to pick pairs to interview, but that is not possible with NSYR whose interviews are complete. So, just for illustration, we created a separate pairing of individuals with completed interviews, not the pairs in Table 1.

Table 2 excerpts parts of the conversation between paired working/nonworking respondents (R) and an interviewer (I). In particular, Table 2 excerpts discussions about drug use. Because this is only an illustration of a statistical technique, Table 2 is very brief; however, the interviews are extensive, permitting extended comparisons of the ten selected pairs.

All four workers say they have used either drugs or alcohol or both, whereas two of the nonworkers say this. This is not inconsistent with the quantitative analysis, but little can be said about relative frequencies based on a few thickly described pairs.

The adolescents in Table 2 express, at times hesitantly, a range of opinions about drug use, and a range of experiences. The nonworker in pair #4 admits to selling drugs, and so makes one think about the meaning of the label 'nonworker'. One might reasonably regard someone who is selling drugs as "working," albeit in the underground economy. Like other workers, adolescents who sell drugs have a source of money besides their relatives, and they have an activity to earn money that competes with school-work. Although only a few pairs are thickly described, we know for all pairs that reclassifying "selling drugs" as "working" would change the working and nonworking groups to enlarge the involvement with illegal drugs in the working group, and shrink it in the nonworking group, thereby enlarging, rather than calling into doubt, the difference observed in the quantitative analysis.

When asked about drug use by others, two of the nonworkers are critical, saying they

avoid drug users. In contrast, two workers express greater tolerance of drug use by others. These two workers describe drug use as harmful to the user but not morally wrong, while these two nonworkers depict drug users as people to be avoided. This suggests questions that might be investigated further. Does working make it more difficult for an adolescent to avoid drug users? If drug users are not rare at work, do workers find that they must become tolerant of views that they cannot avoid if they are to continue working?

Not shown in Table 2, workers and nonworker interviews were also compared in terms of remarks about money, having enough, how it was used. The most consistent pattern we saw was that workers much more often spoke of their cars and using their wages to pay the expenses associated with owning a car. So, perhaps access to money and a car are imbalanced in Figure 3, and perhaps either or both affect an adolescent's access to alcohol and illegal drugs.

5 Discussion: ensuring variety, fine balance

The method in §2 simultaneously produces a conventional matched sample and an exceptionally close subsample of F pairs for interviews or other qualitative studies. The goal is to match in a way that facilitates integration of quantitative and qualitative research. In part, we judge the observed covariates with the aid of much more information about F pairs of individuals who are exceptionally close on the observed covariates. In the example, Table 1 shows that F exceptionally close pairs were indeed found. A few extensions of the matching method are described below.

Ensuring variety in the interviews: We might wish to ensure that the F close pairs that will be interviewed are not too similar to one another. This is easily accomplished: apply the method in §2.2 several times in several subpopulations defined by covariates. In the example in §3, most survey respondents were white, and consequently all F pairs in

Table 1 were white. Instead of selecting the $F = 10$ closest pairs, one could ensure variety by, for instance, selecting $F_1 = 6$ closest pairs in the white subpopulation, the $F_2 = 2$ closest pairs in the black subpopulation, and the $F_3 = 2$ closest pairs in the nonblack Hispanic subpopulation.

Additional matching tools: There are a variety of common extensions of pair matching that may be combined with the method in §2.2. Fine balance is a technique that insists that some nominal variable, perhaps one with many levels, is perfectly balanced; see, for instance, Pimentel et al. (2015) or Yu et al. (2020). Fine balance is a constraint on the match, and this constraint is indifferent to whether pairs are matched for the nominal covariate. It is possible to impose both a fine balance constraint and also the constraint of F exceptionally close pairs. Details are given in Rosenbaum (2017a). The `R` package `thickmatch` implements fine balance.

Alternative matching algorithms: The algorithm in §2.2 used a threshold method to optimize a quantile Δ_μ of the within-pair distances, then imposed that quantile as a constraint, minimizing the total distance subject to that constraint using the optimal assignment algorithm. The first two steps may be retained as is, with the third step replaced by a different algorithm. For instance, the threshold method may be applied to nonbipartite matching, as discussed by Lu et al. (2011), or to matching using mixed integer programming, as discussed by Zubizarreta (2012). In this way, a qualitative component may be added to a variety of matching techniques.

6 Acknowledgement

The National Study of Youth and Religion, http://youthandreligion.nd.edu/, whose data were used by permission here, was funded by Lilly Endowment Inc., under the direction of Christian Smith, of the Department of Sociology at the University of Notre Dame and

Lisa Pearce, of the Department of Sociology at the University of North Carolina at Chapel Hill.

References

Bertsekas, D. P. (1981), "A new algorithm for the assignment problem," *Mathematical Programming*, **21** 152-171.

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

Garfinkel, R. S. (1971), "An improved algorithm for the bottleneck assignment problem," *Operations Research*, 19, 1747-1751.

Gerring, J. and Cojocaru, L. (2016), "Case-selection: A diversity of methods and criteria," *Sociological Methods and Research*, 45, 392-423.

Greenberger, E. and Steinberg, L. D. (1986). *When Teenagers Work: The Psychological and Social Costs of Adolescent Employment.* NY: Basic Books.

Hansen, B. B. (2004), "Full matching in an observational study of coaching for the SAT," *Journal of the American Statistical Association*, 99, 467, 609-618.

Longest, K. C. and Shanahan, M. J. (2007), "Adolescent work intensity and substance use," *Journal of Marriage and Family*, 69, 703-720.

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

Marcus, R., Peritz, E., and Gabriel, K. R. (1976), "On closed testing procedures with special reference to ordered analysis of variance," *Biometrika*, 63, 655-660.

Munafò, M. R. and Davey-Smith, G. (2018), "Repeating experiments is not enough," *Nature*, 553, 399-401.

Pimentel, S. D., Yoon, F. and Keele, L. (2015), "Variable-ratio matching with fine balance in a study of the Peer Health Exchange," *Statistics in Medicine* 34, 4070-4082.

Rosenbaum, P. R. (2001), "Replicating effects and biases," *American Statistician*, 55, 223-227.

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

Rosenbaum, P.R. (2017a), "Imposing minimax and quantile constraints on optimal matching in observational studies," *Journal of Computational and Graphical Statistics*, 26, 66-78.

Rosenbaum, P.R. (2017b), *Observation and Experiment: An Introduction to Causal Inference*, Cambridge, MA: Harvard University Press.

Rosenbaum, P. R. (2020a), "Modern algorithms for matching in observational studies," *Annual Review of Statistics and Its Application*, 7, 143-176.

Rosenbaum, P. R. (2020b), "Combining planned and discovered comparisons in observational studies," *Biostatistics*, to appear, doi:10.1093/biostatistics/kxy055.

Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66, 688-701.

Seawright, J. and Gerring, J. (2008), "Case selection techniques in case study research: A menu of qualitative and quantitative options," *Political Research Quarterly*, 61, 294-308.

Smith, C. and Pearce, L. (2019), *National Study of Youth and Religion*, https://youthandreligion.nd.edu/

Susser, M. (1973), *Causal Thinking in the Health Sciences: Concepts and Strategies in Epidemiology*, New York: Oxford University Press.

Susser, M. (1987). Falsification, verification and causal inference in epidemiology: Reconsideration in the light of Sir Karl Popper's philosophy. In: Susser, M., ed., Epidemiology, Health and Society: Selected Papers, pp. 82–93, New York: Oxford University Press.

Tarrow, S. (2010), "The strategy of paired comparison: toward a theory of practice," *Comparative Political Studies*, 43, 230-259.

Yu, R., Silber, J. H., and Rosenbaum, P. E. (2019), "Matching methods for observational studies derived from large administrative databases," *Statistical Science*, to appear.

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.

Table 1: Covariates for the close pairs that will be thickly described. The treatment, Worked, is hours worked per week, 0 or 20+. Income is family income. Parents is the number present at home. Ed is education is the highest education of parents. The propensity score is p.

id	Pair	Worked	Income	Parents	Ed	Female	Black	Hispanic	Age	Dropout	p
1	1	20+	45000	1	HS	1	0	0	16.5	0	0.15
2	1	0	45000	1	HS	1	0	0	16.5	0	0.15
3	2	20+	45000	2	HS	1	0	0	17.6	0	0.35
4	2	0	45000	2	HS	1	0	0	17.6	0	0.35
5	3	20+	35000	1	HS	0	0	0	17.4	0	0.31
6	3	0	35000	1	HS	0	0	0	17.4	0	0.31
7	4	20+	25000	2	BA	0	0	0	17.2	0	0.25
8	4	0	25000	2	BA	0	0	0	17.2	0	0.26
9	5	20+	105000	2	BA	0	0	0	17.3	0	0.24
10	5	0	105000	2	BA	0	0	0	17.2	0	0.24
11	6	20+	45000	2	HS	0	0	0	14.4	0	0.02
12	6	0	45000	2	HS	0	0	0	14.4	0	0.02
13	7	20+	35000	1	HS	0	0	0	17.2	0	0.26
14	7	0	35000	1	HS	0	0	0	17.2	0	0.26
15	8	20+	25000	1	HS	1	0	0	18.1	0	0.50
16	8	0	25000	1	HS	1	0	0	18.1	0	0.50
17	9	20+	35000	1	HS	1	0	0	17.5	0	0.34
18	9	0	35000	1	HS	1	0	0	17.5	0	0.34
19	10	20+	25000	2	HS	1	0	0	16.0	0	0.09
20	10	0	25000	2	HS	1	0	0	15.9	0	0.09
21	11	20+	45000	2	HS	0	0	0	18.0	0	0.45
22	11	0	45000	2	HS	0	0	0	18.0	0	0.45
23	12	20+	75000	2	BA	1	0	0	17.3	0	0.26
24	12	0	75000	2	BA	1	0	0	17.2	0	0.26

Table 2. Bits of Conversations about Drugs in Selected Matched Pairs

	Nonworker	Worker
Pair 3	I: Okay. Okay. Um, are there any types of people you're not comfortable being friends with or that you don't really want to associate with and why would that be? R: Um, people who do drugs and drink. I: Okay. And because those are things you don't do? R: Yeah.	I: Okay, and do you drink alcohol, or smoke pot, or do other drugs? R: Um, no. … Not now … I: Is there any reason …? R: I don't drink because I crashed my car driving drunk. … I could have hurt somebody so I stopped drinking. I don't smoke weed because … it just, I don't know, it's not bad or anything, I mean I have before, and I used to a lot, um … I just don't. I don't know why I just don't. I just don't see any reason to.
Pair 4	I: ... do you still do drugs? R: No not really, I do, I smoke weed every once in a while, you know … you know, I've tried to quit dealing you know cause I know I'm gonna get caught if I don't quit you know. I: Dealing, drugs? R: Yeah. I mean, I don't deal but I mean, like — I: You sell. R: Yeah ...	I: Okay. Do you ever drink alcohol, or … smoke pot, or do other drugs? R: I have done alcohol, I have done pot, I have not done anything besides that … . I: And you still do smoke marijuana? R: Um, I will every once in a while, but not so much anymore. I: What's every once in a while? R: Maybe half a year. I: Okay. Do you think that drinking or drugs um, is morally wrong? R: I don't think it's morally wrong. It's up to the individual if he wants to screw himself over or not. I don't find anything to be morally wrong.
Pair 7	I: Okay. Do you drink alcohol, smoke pot, or do other drugs? R: No. I: Okay. But you said you had at one point? R: Yes, I have before. Several times. … I: Okay what about pot? R: Same thing. Well actually, I used to do it a lot.	I: Do you drink alcohol, smoke pot, do drugs or anything? R: I drink alcohol occasionally. … I don't smoke cigarettes, I don't smoke pot, I don't do any, that's like the only thing that I do otherwise I drink coffee, but otherwise than that I'm straight as far as that goes. I: … do you [think] that drinking is morally wrong or not? Or doing drugs …? R: … I couldn't see anything morally wrong with drinking or I guess drugs, I just think that it's more of a lack of respect for yourself that's, that's the problem there.
Pair 8	I: Okay. And, ah, are there any people you're not friends with? R: … just a bunch of kids who like to drink and do drugs and stuff. … I: … what are examples of things … that would be morally wrong? R: Well drinking. Drugs.	I: Okay. So do you drink alcohol, smoke pot or do other drugs? R: I have. … I've drank, I've smoked pot. I: Anything else? R: Um, I've done coke. I: Okay. R: I've done ice. … I've rolled (giggles)

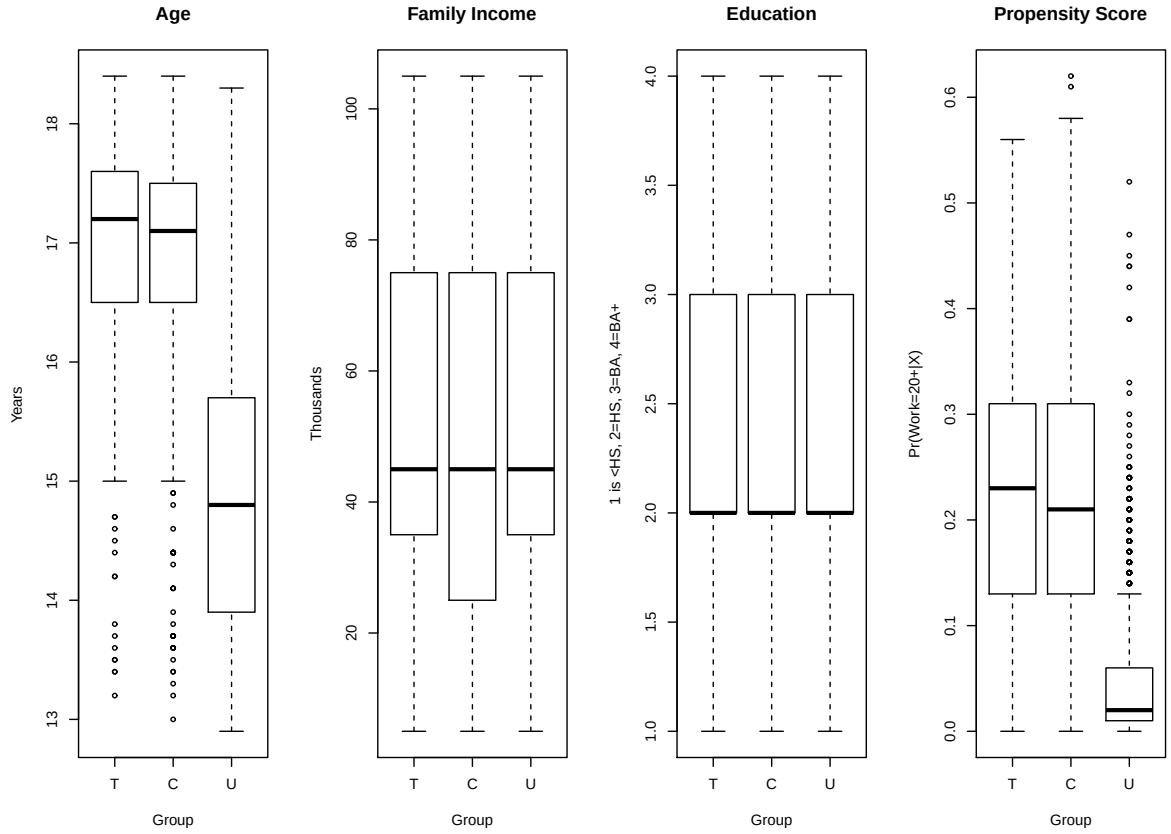

Figure 1: Boxplots of four covariates in 1-to-2 matched sets. T=treated, worked 20+ hours per week (n=268), C=matched control, worked 0 hours (n=536), and U=unmatched, worked 0 hours (n=2012). Income is missing for 4.1% of T, 4.2% of C, and 6.6% of U. Mean income is 51.9 thousand dollars for T, 51.2 for C, and 54.3 for U. Education is the highest parent's income, with mean 2.45 for T, 2.44 for C and 2.49 for U.

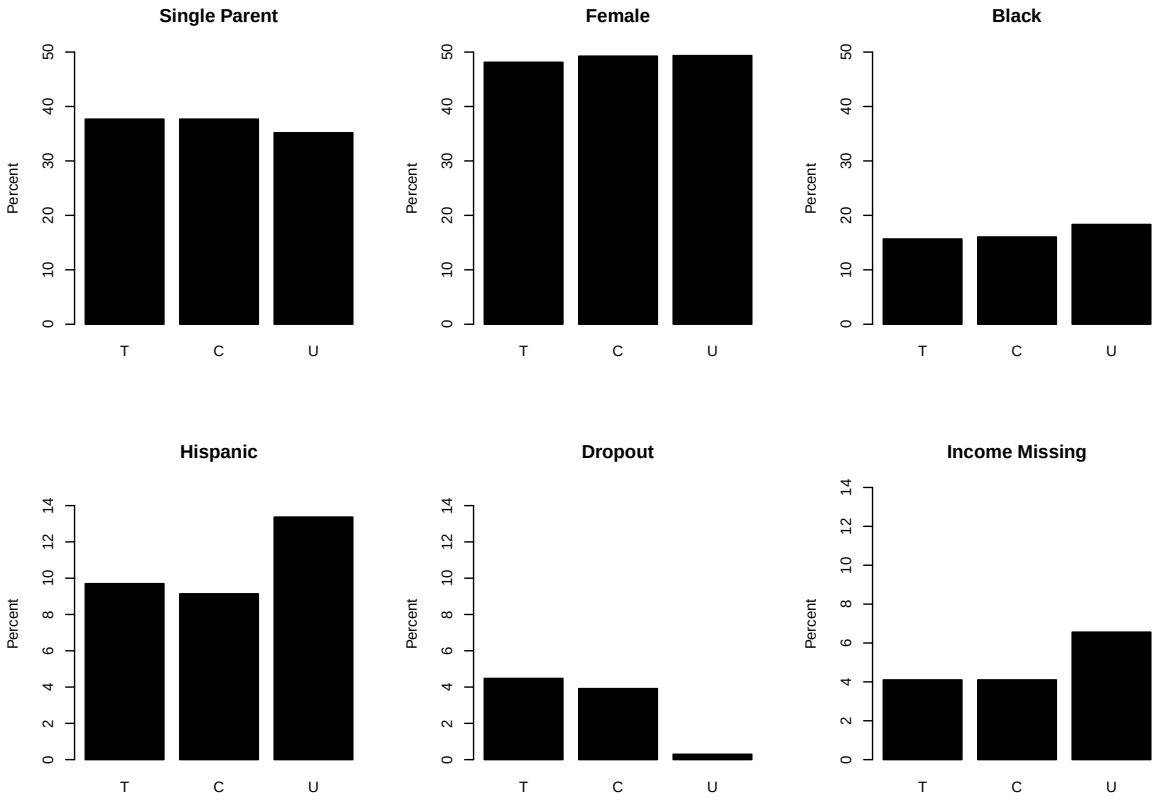

Figure 2: Percents for six binary covariates in 1-to-2 matched sets. T=treated, worked 20+ hours per week (n=268), C=matched control, worked 0 hours (n=536), and U=unmatched, worked 0 hours (n=2012).

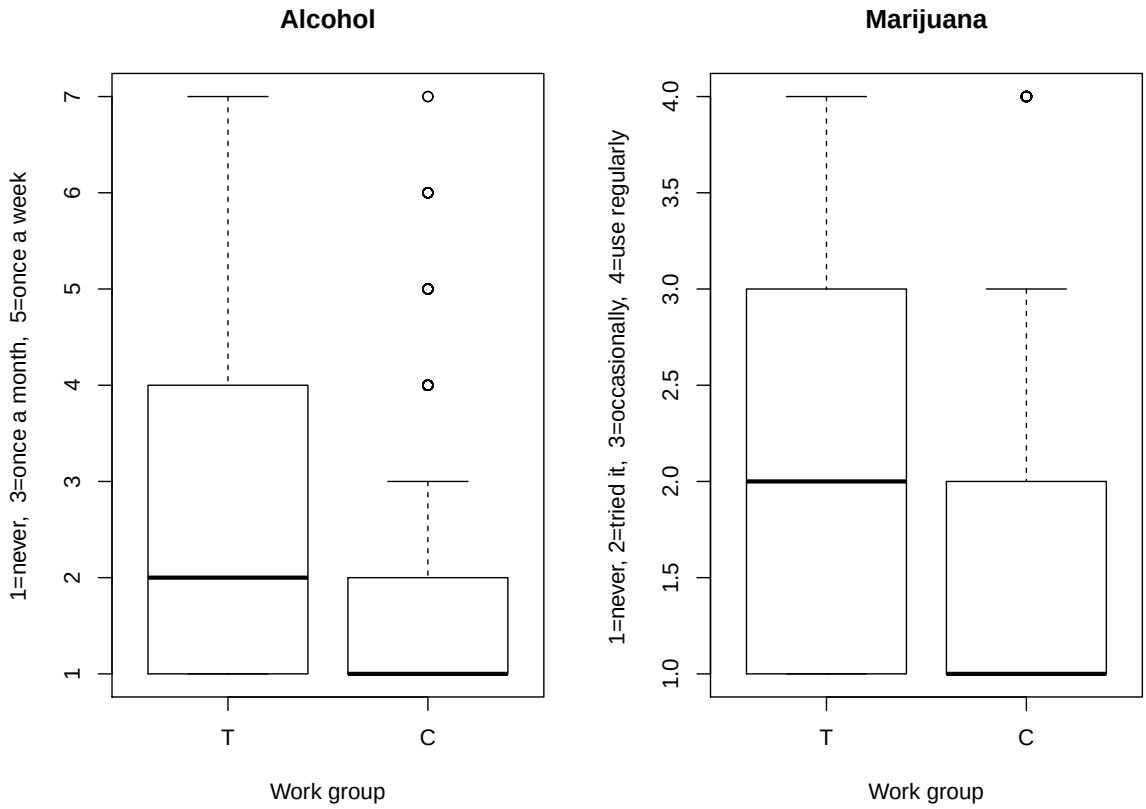

Figure 3: Outcomes, self-reported alcohol and marijuana use in 268 matched sets. T=treated, worked 20+ hours per week (n=268), C=matched control, worked 0 hours (n=536).