Using the Cross-Match Test to Appraise Covariate Balance in Matched Pairs

Ruth HELLER, Paul R. ROSENBAUM, and Dylan S. SMALL

Having created a tentative matched design for an observational study, diagnostic checks are performed to see whether observed covariates exhibit reasonable balance, or alternatively whether further effort is required to improve the match. We illustrate the use of the cross-match test as an aid to appraising balance on high-dimensional covariates, and we discuss its close logical connections to the techniques used to construct matched samples. In particular, in addition to a significance level, the cross-match test provides an interpretable measure of high-dimensional covariate balance, specifically a measure defined in terms of the propensity score. An example from the economics of education is used to illustrate. In the example, imbalances in an initial match guide the construction of a better match. The better match uses a recently proposed technique, optimal tapered matching, that leaves certain possibly innocuous covariates imbalanced in one match but not in another, and yields a test of whether the imbalances are actually innocuous.

KEY WORDS: Multivariate matching; Observational study; Propensity score; Seemingly innocuous confounding; Tapered matching.

1. INTRODUCTION: MOTIVATING EXAMPLE; NOTATION; A MULTIVARIATE MATCH

1.1 Covariate Balance in Matched Observational Studies

In experiments, random assignment of treatments tends to create similar distributions of covariates in treated and control groups; that is, randomization tends to balance the distributions of both observed and unobserved covariates. Randomization does not yield identical treated and control groups, but rather groups which exhibit no systematic relationship with covariates. It is common in randomized trials to begin with a table showing that randomization has been reasonably effective, bringing important observed covariates into reasonable balance. Observational or nonrandomized studies of treatment effects are common in contexts where random assignment is unethical or infeasible, and in these cases, multivariate matching is often used in an attempt to balance the observed covariates. In parallel, it is common in observational studies to begin with a table showing that matching has brought observed covariates into reasonable balance. Of course, unlike randomization, matching for observed covariates cannot be expected to balance unobserved covariates whose possible imbalances must be addressed by other means, such as sensitivity analyses.

One might wish to match exactly for covariates, but when there are many covariates this is not possible. For instance, with 20 covariates, there are 2^{20} or about a million quadrants defined by the medians of the 20 covariates, so with thousands of subjects, it will typically be impossible to match a treated subject to a control who is on the same side of the median for all 20 covariates. Instead of matching exactly for covariates, balancing many observed covariates is often quite feasible; see, for instance, the work of Rosenbaum and Rubin (1985). Covariate balance refers to the distribution of observed covariates in treated and control groups, ignoring who is matched to whom; specifically, observed covariates are independent of treatment assignment. Given that exact matching is not possible, the covariate balance that would be found in a randomized experiment is a useful benchmark for appraising a matched comparison. It is, however, just a recognizable benchmark. There is no particular reason to expect that a matching algorithm will produce balance similar to a completely randomized experiment; it may produce more in easy matching problems or less in difficult ones. Nonetheless, it is useful to know where a particular matched comparison stands in relation to a recognizable benchmark.

In matching, examination of covariate balance is diagnostic. We judge diagnostics by whether they accomplish what they are intended to accomplish, in case of matching, whether they play a constructive role in obtaining a better matched comparison. As is generally true of diagnostic work, the process requires exploratory analysis and judgment, but significance tests can play a limited role, principally as an aid to appraising whether an ostensible pattern could merely reflect the play of chance. For instance, we would not reject a randomized experiment if it exhibited the degree of covariate imbalance that randomization is expected to produce. In a completely randomized experiment, we expect one covariate in twenty to exhibit an imbalance judged significant in a 0.05-level randomization test. See the articles by Hansen and Bowers (2008) and Imai, King, and Stuart (2008) for two views of the relative importance of exploratory analysis, hypothesis tests, and judgement.

1.2 Outline: Using a Balance Diagnostic to Guide Design of a Matched Comparison

In the current article, we illustrate the use of the cross-match test (Rosenbaum 2005; Heller et al. 2010) as a diagnostic in

Ruth Heller is Senior Lecturer, Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa, Israel, and Landau Fellow supported by the Taub Foundation (E-mail: *ruheller@techunix.technion.ac.il*). Paul R. Rosenbaum is Professor and Dylan S. Small is Associate Professor, Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340. Supported by grants SES-0849370 and SES-0961971 from the Measurement, Methodology and Statistics Program of the U.S. National Science Foundation and grant BSF 2008049 from the U.S.–Israel Binational Science Foundation.

 DOI: 10.1198/tast.2010.09210

appraising multivariate covariate balance. The cross-match test momentarily forgets who is treated and who is control, pairing subjects on the basis of covariates only; then, it counts the number of times a treated subject was paired with a control, that is, it counts the cross-matches. If two multivariate distributions are quite different, there will be few cross-matches. Section 2.4 discusses a new result relating the cross-match test to the propensity score. The cross-match test also provides an estimate of the magnitude of departure from covariate balance.

In a typical matched observational study, matched samples are gradually improved until an acceptable match is obtained. An acceptable match will balance observed covariates. Diagnostics play a role in judging whether the current match is acceptable or whether more effort is required. Because matching uses only covariates and treatment assignments without examining outcomes, matching is part of the design of the study. That is, the aspects of the data used in matching would be regarded as fixed predictors if a conventional Gaussian covariance adjustment model were used instead.

In statistics, as in medicine, accurate diagnosis is nice to have, but it is genuinely valuable only if it leads to effective action. To illustrate the value of a diagnostic, it is not sufficient to show that it yields correct diagnoses; rather, one must trace a path from accurate diagnosis to improved results. In matching, this means that the diagnostic must identify a problem with a first match, which leads to a second better match that the diagnostic judges unproblematic. The article is organized around one such path from an unsatisfactory initial match to a much more satisfactory final match. This path will take different forms in different observational studies depending upon the pattern of covariates and treatment assignments. In the example in the current article, the path leads to a tapered match as proposed by Daniel et al. (2008), a technique we describe in detail. In some other example with different problems, the diagnostic might lead in a different direction.

We illustrate the cross-match test in a reanalysis of a study by Cecilia Rouse (1995) which compared educational attainment at two-year and four-year colleges in the United States. In Section 1.3, her study is described. It has 20 observed covariates, and some of these are quite out of balance before matching. Although there are enough controls to match 3-to-1—that is, three students at four-year colleges to each student at a two-year college—use of the cross-match diagnostic in Section 2 strongly suggests a 1-to-1 match will balance covariates, but 2-to-1 or 3-to-1 will not. This is, of course, disappointing, and it raises the question: Is it possible to create a balanced 1-to-1 match in such a way that many controls not used in this match find some other good use? Inspection of the first, disappointing match reveals that one of the most imbalanced groups of variables is the region of the U.S., that is, the North-East, South, Midwest, and West. Two-year colleges are more common in some regions than in others; so, region is substantially out of balance. How important is it to control for region once there is control for educational test scores and socioeconomic measures? One might argue that being in a region that contains few two-year colleges discourages attendance at a two-year college, but aside from doing that it is an innocuous covariate, something that might safely be left unmatched. We answer both of the two questions in this paragraph in Section 3 using optimal tapered matching (Daniel et al. 2008) that optimally splits the potential controls to form two optimally matched control groups, one matched for all 20 covariates, the other matched for the 17 covariates other than the three region indicators. In particular, in Section 4, this matched design yields a test of the hypothesis that the imbalances in region are actually innocuous or else only seem so. To repeat, although the article follows a circuitous path from a poor initial match to a better design, our main goal is to show that the cross-match test is a useful guide along such a path. As discussed in the summary in Section 5, we repeatedly resort to the cross-match test to judge our progress toward an acceptable match.

The most commonly used measures of covariate balance are descriptive statistics, such as the difference in means in units of the pooled standard deviation before matching, or two-sample t-statistics computed after matching to compare with the benchmark of complete randomization. Imai, King, and Stuart (2008) proposed quantile-quantile deviations for individual covariates as more informative than t-tests, in part because their method pays attention to the entire distribution, not just the means. Hansen and Bowers (2008, section 4) suggested a single multivariate test on means similar in form to Hotelling's T^2 statistic, but with the statistic compared to a randomization distribution. In principle, the method of Hansen and Bowers comes in two versions: one compares the balance obtained by matching with the balance obtained by complete randomization; the other looks at residual imbalances in covariates within pairs beyond that expected in a randomized paired experiment. Each of these several diagnostics is likely to be sensitive to differences the others might miss; for instance, differences in means are common, and looking for one is likely to yield greater power if there is a difference in means to be found, but distributions may differ in many ways besides their means. In diagnostic work, it is helpful to have more than one diagnostic, because diagnostics yield not conclusions but an improved match, so if one is going to err it is better to err slightly on the side of excessive rather than deficient improvement.

1.3 Total Educational Attainment of Students Who Begin College at a Two-Year College

In an interesting study, Cecilia Rouse (1995) compared the educational attainment of students who began college in a two-year (or junior or community college) to that of students who began college at a four-year college. Her study used data from the *High School and Beyond* longitudinal study, which includes a good test score from high school composed from subject area tests. Although *High School and Beyond* includes students who did not attend college, all students in the analysis here had some college.

A student who sets out at a two-year or a four-year college may not end up with two or four years of college. A student who attends a two-year college may continue on to get a bachelor's degree at a four-year college, perhaps continuing on to graduate or professional education. A student who attends either a two-year or a four-year college may fail to complete the degree. It is sometimes argued that the path to a BA degree starting in a two-year college is more affordable, perhaps aided by living at home for two years, and hence perhaps easier to complete. Among students whose academic preparation would permit attendance at either a two-year or a four-year college, what is the effect of this choice on educational attainment? Rouse compared the total years of education completed by students who attended two-year and four-year colleges.

We look at students with test scores above 55, which was the median test score of students who attended a four-year college. In terms of test scores, a student with a score above 55 who attended a two-year college could plausibly have been admitted to a four-year college instead, so it is not unreasonable to ask what might have happened had she done so. There were $L = 1818$ students with test scores above 55, denoted $\ell = 1, \ldots, L$, and of these $m = 429$ attended two-year colleges, denoted $Z_\ell = 1$, and $L - m = 1389$ attended four-year colleges, denoted $Z_\ell = 0$.

Unsurprisingly, these students attending two- or four-year colleges looked quite different in high school; see Table 1. In particular, compared to students at four-year colleges, the group attending tw- year colleges had relatively fewer blacks and more Hispanics, had lower test scores (by about half a standard deviation) despite the cutoff at 55, and their parents had less education and less income. Moreover, the group attending two-year colleges had relatively more students from the West and fewer from the Midwest, fewer from an urban area, and more from high schools with a lower percentage of white students. Denote by $\mathbf{x}_\ell$ the vector of covariates in Table 1 for the ℓth of the $L = 1818$ students.

Region of the United States is out of balance in Table 1. Two-year colleges are more common in some regions than in others, and presumably the relative ease of attending a two- or four-year college affects decisions about which college to attend. An investigator might be tempted to view region of the U.S. not as a covariate, but rather as an innocuous nudge toward or away from attending a two-year college, a nudge that is ignored by many students but is decisive in some instances. There is, of course, a concern that region may not be innocuous, that it may be directly related to outcomes apart from college choices, perhaps because it is related to social and economic factors, some not measured, that vary from region to region. Mississippi and Oregon differ in the availability of two-year colleges, but they differ in other ways as well. An "innocuous covariate" is defined formally in (3) of Section 4. Our final matched sample uses region in both of its potential roles: as a covariate controlled by matching, and as a possibly innocent source of seemingly innocuous, uncontrolled variation in the availability of the treatment; see the book by Rosenbaum (2010, section 18.2). Moreover, in Section 4, there will be a statistical test of this seeming innocence, that is, a test of a logical consequence (4) of condition (3).

Table 1. Baseline covariates for students with test scores in high school of 55 or above (the median for students who attended a four-year college). The p-value is from a t-test. The pooled standard deviation (Pooled SD) is the square root of the equally weighted average of the sample variances in the two-year and four-year groups, and the standardized difference (st-dif) is the difference in means divided by this standard deviation.

Covariate	Two-year college $n = 429$ Mean	Four-year college $n = 1389$ Mean	p-value	Pooled SD	st-dif
		Student			
Female %	50	51	0.76		
Black %	6	10	0.00		
Hispanic %	14	10	0.04		
Test score	59.26	60.92	0.00	3.45	−0.48
		Dad's education			
Missing %	13	12	0.52		
Vocational school %	9	7	0.12		
Some college %	15	11	0.09		
BA degree %	21	35	0.00		
		Mom's education			
Missing %	7	4	0.03		
Vocational school %	10	9	0.54		
Some college %	16	16	0.98		
BA degree %	14	25	0.00		
		Family			
Family income 1980 $	24,303	28,265	0.00	17,181	−0.23
Family income missing %	5	6	0.43		
Own's home %	82	84	0.30		
		Neighborhood			
% White in HS	75.96	79.18	0.03	26.2	−0.12
Urban %	17	22	0.01		
		Region			
Midwest %	24	31	0.01		
South %	28	23	0.04		
West %	32	15	0.00		

1.4 Notation: Outcomes, Treatment Assignments, Observed and Unobserved Covariates

The outcome is the total number of years of education. Each student ℓ has two potential values of the outcome, $r_{T\ell}$ if the student is 'treated,' that is, attends a two-year college, and $r_{C\ell}$ if the student is 'a control,' that is, attends a four-year college; see the works of Neyman (?n22) and Rubin (1974). A student who would complete an associate's degree at a two-year college, transfer to a four-year college, and receive a bachelor's degree after two more years would have $r_{T\ell} = r_{C\ell}$ if the student would also have completed the bachelor's degree starting in a four-year college. Similarly, a student who would complete the associate's degree in two years at a two-year college and stop would have $r_{T\ell} = r_{C\ell}$ if the student would have dropped out of a four-year college after two years of study. A student who completes a college's degree program in the expected time and stops would have $r_{T\ell} + 2 = r_{C\ell}$. For student ℓ, $r_{T\ell}$ is observed if the student attends a two-year college, $Z_\ell = 1$, and $r_{C\ell}$ is observed if the student attends a four-year college, $Z_\ell = 0$, so $R_\ell = Z_\ell r_{T\ell} + (1 - Z_\ell) r_{C\ell}$ and Z_ℓ are observed, but the effect, $r_{T\ell} - r_{C\ell}$, is not observed for any student. Write $\mathcal{F} = \{(r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell), \ell = 1, \ldots, L\}$, noting that $\mathcal{F}$ does not include Z_ℓ. In a completely randomized experiment, a fair coin is independently flipped to determine the L treatment assignments. To say the coin is fair is to say that $\Pr(Z_\ell = 1|\mathcal{F})$ is constant for $\ell = 1, \ldots, L$, so $\Pr(Z_\ell = 1|\mathcal{F})$ does not vary with $(r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell)$.

To speak about what happens in large samples, $L \to \infty$, it is convenient to assume that the L vectors $(r_{T\ell}, r_{C\ell}, Z_\ell, \mathbf{x}_\ell)$ were independently sampled from an infinite population, and to let the omission of a subscript, say $\mathbf{x}$, signify that reference is made to the distribution of a quantity in that population. One consequence of random assignment is that the probability distributions of covariates are balanced in treated and control groups, $\Pr(\mathbf{x}|Z = 1) = \Pr(\mathbf{x}|Z = 0)$, but Table 1 strongly suggests $\Pr(\mathbf{x}|Z = 1) \neq \Pr(\mathbf{x}|Z = 0)$ in this nonrandomized comparison. The propensity score $e(\mathbf{x})$ is the conditional probability of treatment given the observed covariates, $e(\mathbf{x}) = \Pr(Z = 1|\mathbf{x})$, and conditioning on $e(\mathbf{x})$ balances the observed covariates $\mathbf{x}$ in the sense that $\Pr\{\mathbf{x}|e(\mathbf{x}), Z = 1\} = \Pr\{\mathbf{x}|e(\mathbf{x}), Z = 0\}$, although it cannot be expected to balance an unobserved covariate u; see the article by Rosenbaum and Rubin (1983). Treatment assignment is said to be ignorable given $\mathbf{x}$ if $\Pr(Z = 1|r_T, r_C, \mathbf{x}) = \Pr(Z = 1|\mathbf{x})$ with $0 < \Pr(Z = 1|\mathbf{x}) < 1$ for all $\mathbf{x}$, and in this case: (i) matching exactly for the high-dimensional $\mathbf{x}$ suffices to estimate expected treatment effects, such as $E(r_T - r_C|Z = 1)$, but (ii) so does matching on the scalar propensity score, $e(\mathbf{x})$; see, again, the article by Rosenbaum and Rubin (1983). Because the propensity score depends on Z and $\mathbf{x}$, it can be estimated from observed data, perhaps with the aid of a model such as a logit model for $\Pr(Z = 1|\mathbf{x})$.

2. TESTING COVARIATE BALANCE USING THE CROSS-MATCH TEST

2.1 Three Layered Matched Samples

For the 429 students attending a two-year college, we construct three nonoverlapping matched control groups of students attending four-year colleges, each matched control group containing 429 students. The control groups are layered: the first control group is an optimal pair matching; the second is an optimal pair matching from the unused controls; the third is an optimal pair matching from the still unused controls. Together, the three control groups include $3 \times 429 = 1287$ controls or $1287/1389 = 93\%$ of the available controls. As in the article by Smith (1997), we examine the degree of covariate imbalance with one, two, or three matched controls.

The matched control groups were formed using calipers of 0.2 standard deviation on an estimated propensity score based on a logit model, one standard deviation on the test score, and optimal matching within calipers using the Mahalanobis distance. See the works of Bergstralh, Kosanke, and Jacobsen (1996), Bertsekas (1981), Hansen and Klopfer (2006), Hansen (2007), Rosenbaum and Rubin (1985), Rosenbaum (1989), and Rubin (1980) for discussion of various aspects of such a match, and see the book by Rosenbaum (2010, chapter 8) for an overview.

Table 2 and Figure 1 describe the three resulting matched samples. In Table 2 and Figure 1, the first match is C-1, the second is C-2, and the third is C-3; each contains 429 controls. Viewed informally, the first match appears to be quite successful at balancing the observed covariates, and the third match is terrible. For the third match, the difference in mean test scores in high school is 80% of the standard deviation before matching, with a t-statistic of -12.4: the C-3 controls had much higher test scores than the students in two-year colleges. Also, the C-3 controls had wealthier, better educated parents. In the final panel of Figure 1, the upper quartile of the estimated propensity score in the third control match is well below the lower quartile in the treated group, so in a multivariate sense these groups barely overlap.

It is useful to pause for a moment to think about the value added, if any, by the third control match, C-3, in Table 2 and Figure 1, and in particular to connect our technical thoughts about this subject with our everyday experiences with colleges and college admissions in the United States. Compared to the students in two-year colleges, the C-3 controls have much higher test scores in high school and parents with more education and more income. Think about the U.S. in all its complexity, think about these two groups of students, their childhoods at home, the colleges they attended. It is easy to imagine certain students thoughtfully deciding between a two-year and a four-year college, while it is very difficult to imagine certain other students spending even a moment on the decision. Presumably, a student with ample financial resources who attended Harvard or Stanford or MIT spent very little time considering the possibility of attending a two-year college instead, and had such a student attended a two-year college she would have stood out as quite unusual in several respects. Would such a C-3 student, with her high test scores and ample finances, play a useful role in estimating the effect of two-versus-four-year colleges? If one could have total faith in the extrapolations of a parametric regression model, such as a Gaussian linear model, then yes, of course, she would help us fit that model, and the model would predict what would have happened if she, an MIT undergrad, had instead attended a two-year college, even though the model

Table 2. Covariates in three layered matched comparisons. For continuous covariates, both the mean and the mean difference in units of the pooled standard deviation (st-dev) are given using the standard deviation before matching from Table 1.

	Two-year college	Four-year match C-1	Four-year match C-2	Four-year match C-3	Match C-1	Match C-2	Match C-3
	$n = 429$	$n = 429$	$n = 429$	$n = 429$	2-sample t-statistic		
Covariate	Mean	Mean	Mean	Mean	t	t	t
Student							
Female %	50	52	49	52	−0.6	0.2	−0.5
Black %	6	5	8	16	0.6	−1.5	**−4.9**
Hispanic %	14	14	9	8	−0.1	**2.1**	**2.7**
Test score (mean)	59.26	59.36	59.93	62.03	−0.5	**−3.1**	**−12.4**
Test score (st-dif)		−0.03	−0.19	−0.80			
Dad's education							
Missing %	13	12	11	14	0.6	0.8	−0.5
Vocational school %	9	9	10	3	0.2	−0.3	**3.9**
Some college %	15	16	13	7	−0.4	0.6	**3.8**
BA degree %	21	22	25	46	−0.5	−1.5	**−7.9**
Mom's education							
Missing %	7	6	5	2	0.7	0.9	**3.8**
Vocational school %	10	9	10	8	0.3	−0.1	1.0
Some college %	16	16	16	18	0.0	0.1	−0.7
BA degree %	14	16	15	33	−0.8	−0.2	**−6.6**
Family							
Family income (mean)	24,303	23,641	26,346	31,194	0.6	−1.8	**−5.4**
Family income (st-dif)		0.04	−0.12	−0.40			
Family income missing	5	5	5	7	0.0	−0.2	−1.4
Own's home %	82	84	84	85	−0.6	−0.7	−1.0
Neighborhood							
% White in HS (mean)	76	76	79	81	−0.2	−1.7	**−2.6**
% White in HS (st-dif)		−0.01	−0.11	−0.18			
Urban %	17	13	23	30	1.3	**−2.2**	**−4.6**
Region							
Midwest %	24	26	33	35	−0.6	**−3.0**	**−3.5**
South %	28	30	29	14	−0.7	−0.3	**5.2**
West %	32	30	14	3	0.8	**6.4**	**12.0**

has never seen such a student attend a two-year college, and so is extrapolating its parametric form into regions where there are no data. If one had less than complete faith in the extrapolations of a parametric model, then the contribution of a C-3 student to the study of two-year colleges is, at best, less clear. Matching attempts to compare people who received one treatment to other people who received a different treatment but otherwise look similar in terms of observed covariates. Matching diagnostics—the elementary ones in Table 2 and the cross-match test in the current article—raise objections when an attempt is made to compare groups that are visibly very different prior to treatment.

Descriptive statistics and informal examination of t-statistics for the 20 covariates viewed one at a time suggest the first layer is balanced. Nonetheless, we should ask: Could it be that the distributions of the 20-dimensional $\mathbf{x}$ in Table 2 are different in treated and control groups, though the marginal means look similar? Conversely, the second layer exhibits a few large t-statistics among 20 t-statistics. With 20 t-statistics testing covariate balance in a completely randomized experiment, it would not be surprising to see one or two t's significant at the 0.05 level by chance alone. Would a single test applied to all 20 covariates reinforce the view that the second layer exhibits more imbalance than would be expected in a completely randomized experiment? In Section 2.3, the cross-match test will provide an answer to these questions.

2.2 Missing Values for Some Covariates

In Table 2 and in matching generally, missing values of an observed covariate are viewed as an observable aspect of the covariate, to be balanced in treated and control groups along with other observables. That is, a missing value of mother's education is an observable category of mother's education, which is in reasonable balance for the C-1 controls in Table 2 and substantially out of balance for the C-3 controls. For the continuous variable, 'family income,' there is a supplemental binary indicator covariate, 'family income missing,' which is also in balance for the C-1 controls at 5% in both treated and control groups. Obviously, balancing the observable pattern of missing data does not imply that the unobservable missing data are also balanced, but matching is targeted at observables, and should be judged by what it can realistically be expected to do. See the work of Rosenbaum and Rubin (1984, appendix) and Rosenbaum (2010, section 9.4) for details and specifics. The cross-match test handles missing covariate values in the same way,

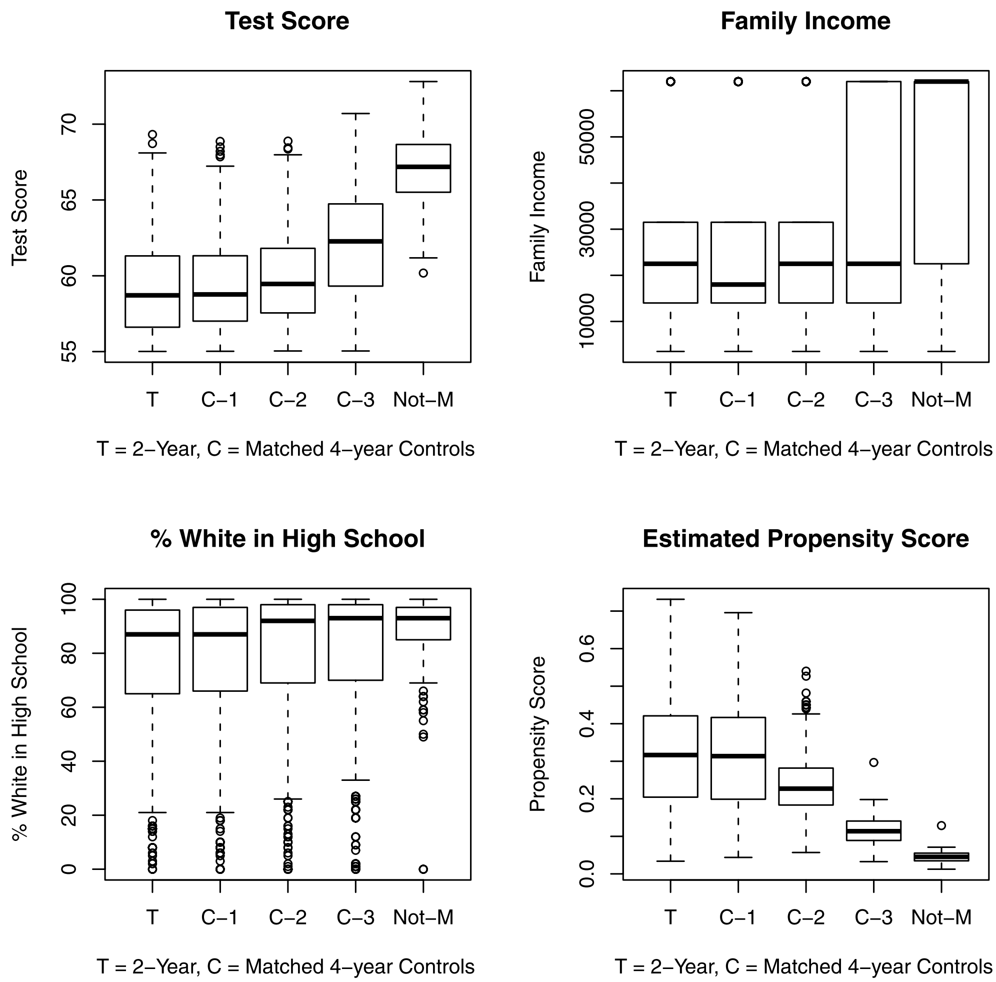

Figure 1. Boxplots of "continuous" covariates for the treated group (T) of 429 students in two-year colleges and three layered matched control groups of 429 students in four-year colleges (C-1 = first, C-3 = last), and 102 unmatched potential controls (Not-M). Family income is given in seven levels, which is the reason for the gaps in the boxplot.

judging whether observable covariate values and patterns of missing covariate values are in balance in treated and control groups.

2.3 Can the Treated and Control Groups Be Rediscovered From the Covariates Alone?

Suppose that we ignored who is treated and who is control, and who is matched to whom, and suppose that we paired subjects based on the covariates alone. Would we tend to pair treated subjects to treated subjects and controls to controls? Or would the pairing be unrelated to the treatment group? In a completely randomized experiment, treatment assignment is independent of covariates, so apart from chance, pairing subjects based on covariates would fail to identify the treatment group. If the covariate distributions were very different in treated and control groups, then the pairing would, more often than chance, pair individuals in the same group.

The cross-match test pairs subjects based on covariates and takes as the test statistic A_1 the number of times a treated subject was paired with a control, rejecting the hypothesis of equal distributions for small values of the statistic; see the article by Rosenbaum (2005). As in that article, a rank-based Mahalanobis distance is computed between every pair of subjects, and subjects are divided into pairs to minimize the total distance within pairs, using Derigs's (1988) algorithm, as made available in R as nbpMatching by Lu et al. (?Lu09). An R package <ref:Lu09?> crossmatch to perform the cross-match test is available from the first author's web page or CRAN; it calls the R package by Lu et al. If $858 = 429 + 429$ subjects are paired into 429 pairs, then $\mathrm{E}(A_1) = 214.75$ cross-matches are expected by chance when the distributions of covariates are the same, with variance $\mathrm{var}(A_1) = 107.38$, and $\{A_1 - \mathrm{E}(A_1)\}/\sqrt{\mathrm{var}(A_1)}$ converges in distribution to the standard Normal as the sample size increases; see propositions 1 and 2 in the article by Rosenbaum (2005).

Table 3 presents the cross-match test comparing the treated group to each of the three control groups, and comparing the control groups to each other. Although comparisons in terms of individual covariates in Table 2 are essential, Table 3 sharpens these comparisons, making it clear that the imbalances in the second layer are not artifacts of having performed 20 comparisons, and also providing no sign of a multivariate imbalance in the first layer hiding amid balance on the marginal means of the 20 covariates.

Table 3. Cross-match test results for the layered match comparing matched groups two at a time. In a completely randomized experiment with two groups of equal size $\Upsilon = 1/2$ with smaller values indicating greater separation of the covariate distributions.

Match	Cross-matches A_1	Estimate of Υ $A_1/429$	p-value
T versus C-1	219	0.51	0.66
T versus C-2	177	0.41	0.00013
T versus C-3	93	0.22	3.6×10^{-32}
C-1 versus C-2	195	0.45	0.028
C-1 versus C-3	107	0.25	1.3×10^{-25}
C-2 versus C-3	127	0.30	1.2×10^{-17}

The cross-match test may be applied to compare the treated group to the union of several layered control groups. For instance, if it is applied to the union of the treated group and the union of the three layered matched control groups, it produces 295 cross-matches when 321.94 are expected by chance, yielding a p-value of 0.0071.

The largest imbalances in the second layer refer to region of the United States. Two-year colleges are more common in some regions than in others. Perhaps imbalances in region are not so worrisome as imbalances in educational or socioeconomic covariates. Might the second layer be used in some manner ignoring the imbalances in region? If the cross-match test is applied to the second layer for just the 17 covariates excluding region, there are 187 cross-matches, with 214.75 expected by chance, yielding a deviate of -2.68 and a p-value of 0.0037, so the remaining 17 covariates in the second layer are more imbalanced than would have been expected if the treated group and the second layer had been formed by complete randomization. Of these 17 covariates, most worrisome for college success is the imbalance in Table 2 in test score from high school.

Guided by these comparisons, another match is constructed in Section 3.

2.4 The Cross-Match Test and the Propensity Score

With $\pi = \Pr(Z = 1)$, define the quantity

$$\Upsilon = 2 \int \frac{\pi(1-\pi)\Pr(\mathbf{x}|Z=1)\Pr(\mathbf{x}|Z=0)}{\pi\Pr(\mathbf{x}|Z=1) + (1-\pi)\Pr(\mathbf{x}|Z=0)}\, d\mathbf{x}. \quad (1)$$

The parameter Υ is discussed by Henze and Penrose (1999, theorem 2); it is a transformation of one of Györfi and Nemetz's measures of distributional separation. Clearly, $\Upsilon = 2\pi(1-\pi)$ if $\Pr(\mathbf{x}|Z=1) = \Pr(\mathbf{x}|Z=0)$. By Bayes's theorem,

$$\Upsilon = 2\mathrm{E}\big[e(\mathbf{x})\{1 - e(\mathbf{x})\}\big]. \quad (2)$$

So Υ has the following simple interpretation: if a value of $\mathbf{x}$ is picked at random and two subjects are sampled with this value of $\mathbf{x}$, then Υ is the probability that one subject will be treated and the other control, so that they might be paired to form a treatment-versus-control pair. In a completely randomized experiment with $\pi = 1/2$, the probability is $\Upsilon = 2\pi(1-\pi) = 1/2$.

The quantity A_1/I in Table 3 is an estimate of Υ; see the article by Rosenbaum (2005, section 3.4 where $N \doteq 2I$). More precisely, matching alters the distribution $\Pr(\mathbf{x}|Z=0)$ of observed covariates $\mathbf{x}$ among controls with $Z = 0$, and Table 3 is estimating Υ for this altered distribution. When Υ is computed for treated/control matched pairs, success or covariate balance is $\Upsilon = 1/2$, and failure is Υ much less than $1/2$. In Table 3, the treated group and third control group exhibit substantial separation: pick an $\mathbf{x}$ at random from the matched distribution of $\mathbf{x}$ and then pick two subjects at random with that $\mathbf{x}$, and it is estimated that 78% of the time they will come from the same group, both treated or both control.

3. A TAPERED MATCH

In an optimal tapered match, a single control group is optimally divided and optimally paired with treated subjects so that each treated subject is paired with two controls which meet different matching criteria in such a way that the total distance within pairs is minimized. Optimal tapered matching for two or more controls was proposed by Daniel et al. (2008) who proved that the simple steps described later in the current paragraph produce the optimal tapered match. Here, one level of the taper (C-1) will match essentially as in Section 2.1 for all 20 covariates, the other level (C-2) will match for 17 covariates excluding region, with the algorithm optimally dividing the controls among levels to minimize the total covariate distance across both matches. The distances were essentially the same as before, except one distance used 20 covariates, the other distance used 17 covariates, and there were two propensity scores, one with 17 covariates, the other with 20 covariates, with only the former used in the 17-covariate match, and both scores used in the 20-covariate match. In addition, some of the caliper widths were adjusted. Call these two distance matrices for 17 and 20 covariates d17 and d20; each matrix has one row for each treated subject and one column for each potential control. The standard optimal assignment algorithm pairs rows and columns of a distance matrix to minimize the total distance within pairs (e.g., Bertsekas 1981, 1991; Cook et al. 1998; Dell'Amico and Toth 2000). In R, the pairmatch(·) function of Hansen's (2007) optmatch package solves the optimal assignment problem. The algorithm of Daniel et al. (2008) produces the optimal tapered match by solving this familiar optimal assignment problem for an augmented distance matrix. The augmented distance matrix has two rows for each treated subject and one column for each potential control, and one of the two rows for a treated subject records the first distance for 20 covariates, the other records the second distance for 17 covariates; in R, the augmented distance matrix is rbind(d17,d20). So in R, having defined d17 and d20, you install and load optmatch, and obtain the optimal tapered match as pairmatch(rbind(d17,d20)). Given the structure of the augmented distance matrix, that optimal assignment will pair each treated subject to two different controls, one selected for proximity on the first distance, the other selected for proximity on the second. So the steps required are easy to describe, and only a little more work is required to prove that these steps do indeed produce an optimal tapered match; see the article by Daniel et al. (2008). Also, with very minor changes, there can be more than one control selected at each level of the taper, and there can be more than two levels of the taper; again, see the work by Daniel et al. (2008). For a very different approach to matching with more than one matching criterion, see the article of Rubin and Stuart (2006).

Table 4. Imbalance in region in the tapered match.

	Two-year college	Four-year match 1	Four-year match 2	Four-year unmatched	Match C-1	Match C-2	Match Unmatched
	$n = 429$	$n = 429$	$n = 429$	$n = 429$	2-sample t-statistic		
Covariate	Mean	Mean	Mean	Mean	t	t	t
Midwest %	24	29	32	32	−1.5	**−2.6**	**−2.7**
South %	28	31	17	21	−1.0	**3.8**	**2.6**
West %	32	29	8	9	1.0	**9.4**	**9.2**

The C-1 match intended to balance all 20 covariates, while the C-2 match intended to allow the three regional covariates to be imbalanced while balancing the remaining 17 covariates. Did this happen? Table 4 shows that the C-1 match is fairly well balanced for region, but the C-2 match is not. Table 5 applies the cross-match test to all 20 covariates, to the 17 covariates excluding region, and to groups of covariates. The C-2 match is clearly very different from the treated group in terms of region, but otherwise the covariates look balanced. The C-1 controls look balanced except perhaps for some imbalance in the family variables. Figure 2 depicts the imbalances in four continuous covariates. Unlike Figure 1, the C-2 match appears acceptable for the covariates in Figure 2. Figure 3 compares the layered and tapered matches for 20 covariates and 17 covariates—in the tapered match, imbalances in the second group of controls are largely confined to the three region indicators.

4. IS REGION INNOCUOUS?

Write $\mathbf{x} = (\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ where $\overline{\mathbf{x}}$ contains the covariates controlled at both levels of tapered matching, and $\widetilde{\mathbf{x}}$ contains the covariates controlled at the first level of the taper but not the second. In Section 3, $\widetilde{\mathbf{x}}$ contains the three region indicators and $\overline{\mathbf{x}}$ contains the remaining 17 covariates. Dawid (1979) wrote $A \perp\!\!\!\perp B|C$ for "A is conditionally independent of B given C," and he made a general argument that scientific assumptions are often best expressed in terms of conditional independence rather than in terms of parametric models which may have scientifically extraneous features. In that spirit, we say $\widetilde{\mathbf{x}}$ is innocuous given $\overline{\mathbf{x}}$ if $\widetilde{\mathbf{x}}$ is related to treatment assignment Z but not to response (r_T, r_C) given $\overline{\mathbf{x}}$—that is, in Dawid's (1979) notation, if

$$(r_T, r_C) \perp\!\!\!\perp (Z, \widetilde{\mathbf{x}})|\overline{\mathbf{x}}. \tag{3}$$

If treatment assignment were ignorable given $\mathbf{x} = (\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, and if $\widetilde{\mathbf{x}}$ were innocuous, then treatment assignment would be ignorable given $\overline{\mathbf{x}}$ alone, that is, $\Pr(Z = 1|r_T, r_C, \mathbf{x}) = \Pr(Z = 1|\mathbf{x})$ with $0 < \Pr(Z = 1|\mathbf{x}) < 1$ and (3) together imply $\Pr(Z = 1|r_T, r_C, \overline{\mathbf{x}}) = \Pr(Z = 1|\overline{\mathbf{x}})$ with $0 < \Pr(Z = 1|\overline{\mathbf{x}}) < 1$. In this case, either or both of the C-1 and C-2 matches in Section 3 would provide consistent estimates of treatment effects.

Importantly, in a tapered match which controls $\mathbf{x} = (\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ at one level of the taper and only $\overline{\mathbf{x}}$ at the other, condition (3) together with ignorable assignment given $\mathbf{x}$ has a testable consequence; it implies

$$r_C \perp\!\!\!\perp \widetilde{\mathbf{x}}|(\overline{\mathbf{x}}, Z = 0), \tag{4}$$

so in the C-1 versus C-2 pairs matched for $\overline{\mathbf{x}}$ with $Z = 0$, the observable distribution of responses r_C to control among the C-1 and C-2 controls is unaffected by also matching for $\widetilde{\mathbf{x}}$. If (3) were true, then among controls matched for $\overline{\mathbf{x}}$, differences in $\widetilde{\mathbf{x}}$ would not predict the response r_C among controls $Z = 0$.

Expressed in a different way, if one thought the regional indicators were innocuous, one might estimate the treatment effect by the average difference in education between the treated subjects (T) and the average of their two matched controls (T versus the average of C-1 and C-2), whereas if one doubted that the regional indicators were innocuous, one would estimate the effect by the mean of difference between the treated subjects

Table 5. Cross-match test results for the tapered match.

Match	Covariates (number of covariates)	Cross-matches A_1	Estimate of Υ $A_1/429$	p-value
T versus C-1	All 20	219	0.51	0.66
T versus C-2	All 20	165	0.38	0.00000079
T versus C-1	17 without region	203	0.47	0.13
T versus C-2	17 without region	203	0.47	0.13
T versus C-1	Student (4)	221	0.52	0.73
T versus C-2	Student (4)	215	0.50	0.51
T versus C-1	Parents education (8)	217	0.51	0.59
T versus C-2	Parents education (8)	229	0.53	0.92
T versus C-1	Family (3)	197	0.46	0.043
T versus C-2	Family (3)	209	0.49	0.29
T versus C-1	Neighborhood (2)	207	0.48	0.23
T versus C-2	Neighborhood (2)	211	0.49	0.36
T versus C-1	Region (3)	203	0.47	0.13
T versus C-2	Region (3)	175	0.41	0.000063

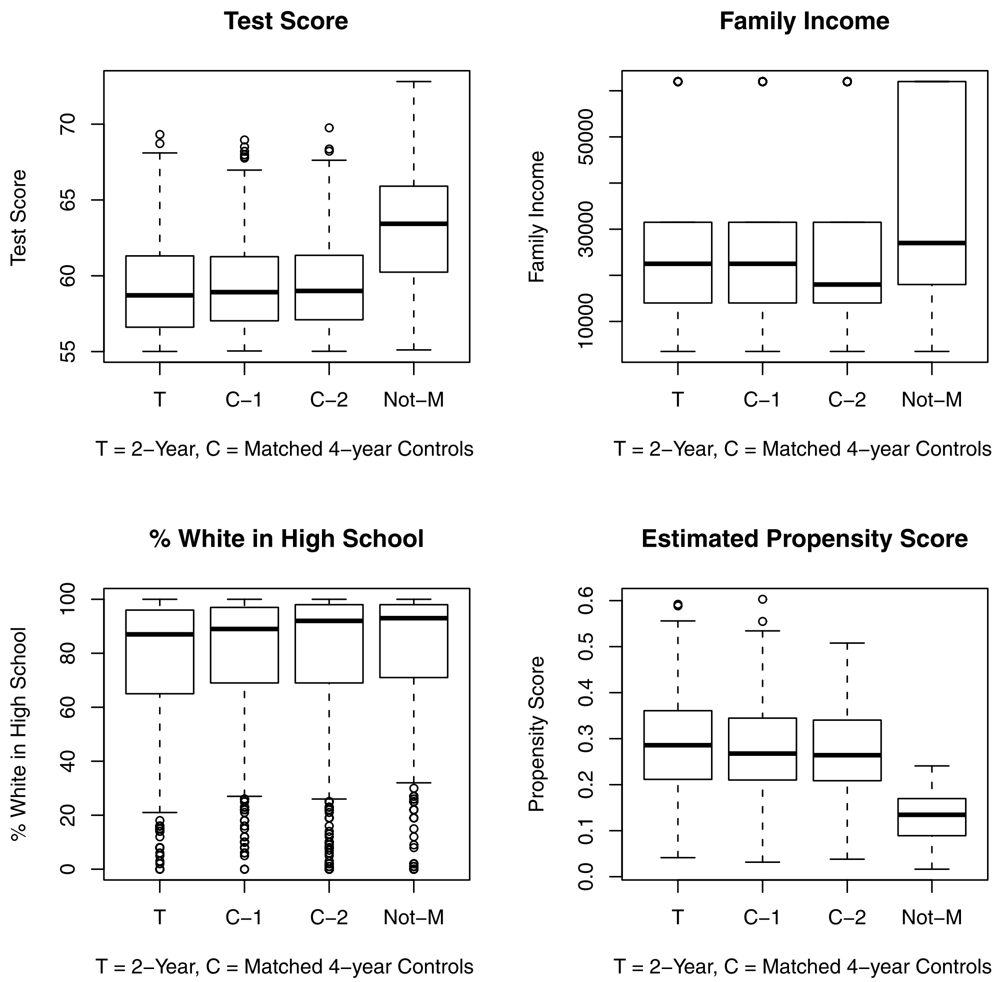

Figure 2. Boxplots of continuous covariates for the tapered match. Control group C-1 is matched for all 20 covariates, while control group C-2 is matched for 17 covariates excluding the three region indicators. The unmatched controls are Not-M. The treated group and the C-2 match differ substantially in terms of region, but not in terms of other covariates. The match uses two propensity scores, but only the 17 covariate score is displayed. Not seen here, but as expected, the propensity score with 20 covariates looks similar for the C-1 controls, but different for the C-2 controls.

and their first controls (T versus C-1) matched for all of **x**. The difference of these two estimates is the basis for the simplest form of a Hausman (1978) test, and it is proportional to the difference between the means of the two matched controls (C-1 versus C-2). In a Hausman test, an assumption is tested by the difference in two parameter estimates, where only one of the estimates requires the assumption for consistency.

Figure 4 shows the results. As one might anticipate, the median years of education is 14 years for a two-year college and 16 for a four-year college, but there is considerable variation. The median difference, two-year versus four-year college, is -1 year of education, and a quarter of the students attending two-year colleges had at least as many years of education as their matched controls at four-year colleges. The C-1 and C-2 controls look similar in terms of years of education, so one obtains similar estimates of effect whether one restricts attention to comparisons within the same region or compares ostensibly similar students in regions that differ in terms of the availability of two-year colleges.

The attraction of the C-1 controls is that ostensibly similar students in the same region are compared. However, we do not know why, in the same region, two ostensibly similar students made different college choices. The attraction of the C-2 controls is that part of the variation in college choice presumably reflects the differing availability of two- and four-year colleges in different regions, and perhaps that source of variation in college choice is innocuous, that is, not much related to important unmeasured attributes of the students. However, the C-2 controls do not resemble the treated group in terms of region. In Figure 4, the two controls, C-1 and C-2, give similar impressions of the treatment effect, perhaps somewhat reducing the reasonable concerns about each group on its own.

5. SUMMARY: THE CROSS-MATCH TEST AS A GAUGE OF PROGRESS

The use of the cross-match test in appraising covariate balance has been illustrated. In a preliminary analysis, the cross-match test suggested that covariate balance on all 20 observed

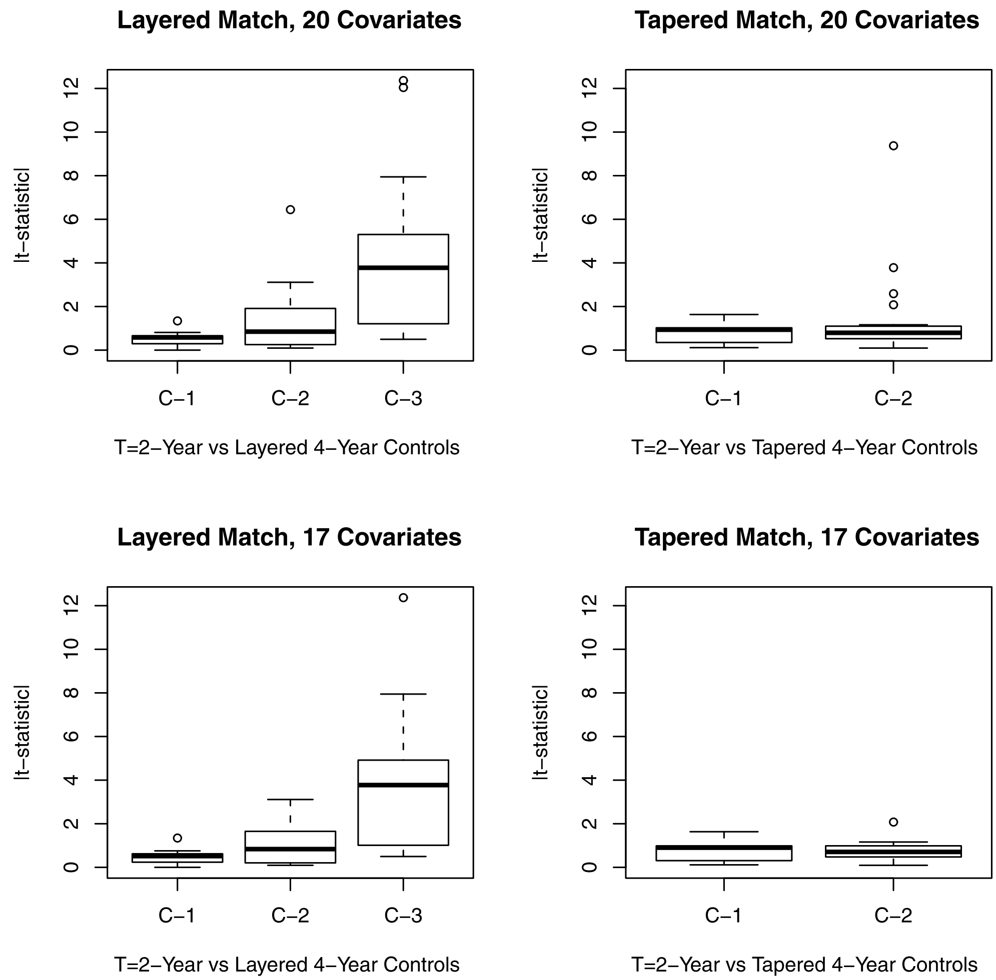

Figure 3. Comparison of absolute t-statistics in the layered and tapered matched comparisons, for all 20 covariates and for the 17 covariates excluding Region. Only the C-1 controls in the layered match are balanced with respect to observed covariates. In the tapered match, the C-1 controls are balanced with respect to all 20 observed covariates, and the C-2 controls are balanced for the 17 covariates excluding Region.

covariates was possible with 1-to-1 matching, but not with 1-to-2 matching. Tapered matching then created a 1-to-1 match for all 20 covariates, and an additional 1-to-1 match for 17 of the 20 covariates, the latter permitting the possibly innocuous 'region of the U.S.' to remain unmatched. The cross-match test indicated the first tapered control group had created reasonable balance on the 20 observed covariates, while the second control group had balanced all observed covariates except region, with region substantially out of balance. It seems reasonable to conjecture that the availability of two-year colleges in different regions was one aspect of the college choices in the second control group. In the example, similar estimates of effect were obtained from comparisons within and between regions.

Again, diagnostics are judged by what diagnostics are intended to do, in the case of matching, to produce a better matched design. Arguably, the second tapered match is a better use of the available data than any of the layered matched designs, and the cross-match test played a useful role in the steps leading to an improved design.

[Received October 2009. Revised August 2010.]

REFERENCES

Bergstralh, E. J., Kosanke, J. L., and Jacobsen, S. L. (1996), "Software for Optimal Matching in Observational Studies," *Epidemiology*, 7, 331–332. SAS code is available at *http://www.mayo.edu/hsr/sasmac.html*. [4]

Bertsekas, D. P. (1981), "A New Algorithm for the Assignment Problem," *Mathematical Programming*, 21, 152–171. [4,7]

——— (1991), *Linear Network Optimization*, Cambridge, MA: MIT Press. [7]

Cook, W. J., Cunningham, W. H., Pulleyblank, W. R., and Schrijver, A. (1998), *Combinatorial Optimization*, New York: Wiley. [7]

Daniel, S., Armstrong, K., Silber, J. H., and Rosenbaum, P. R. (2008), "An Algorithm for Optimal Tapered Matching, With Application to Disparities in Survival," *Journal of Computional and Graphical Statistics*, 17, 914–924. [2,7]

Dawid, A. P. (1979), "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society, Ser. B*, 41, 1–31. [8]

Dell'Amico, M., and Toth, P. (2000), "Algorithms and Codes for Dense Assignment Problems: The State of the Art," *Discrete Applied Mathematics*, 100, 17–48. [7]

Derigs, U. (1988), "Solving Nonbipartite Matching Problems by Shortest Path Techniques," *Annals of Operations Reseach*, 13, 225–261. [6]

Hansen, B. B. (2007), "Optmatch: Flexible, Optimal Matching for Observational Studies," *R News*, 7, 18–24. [4,7]

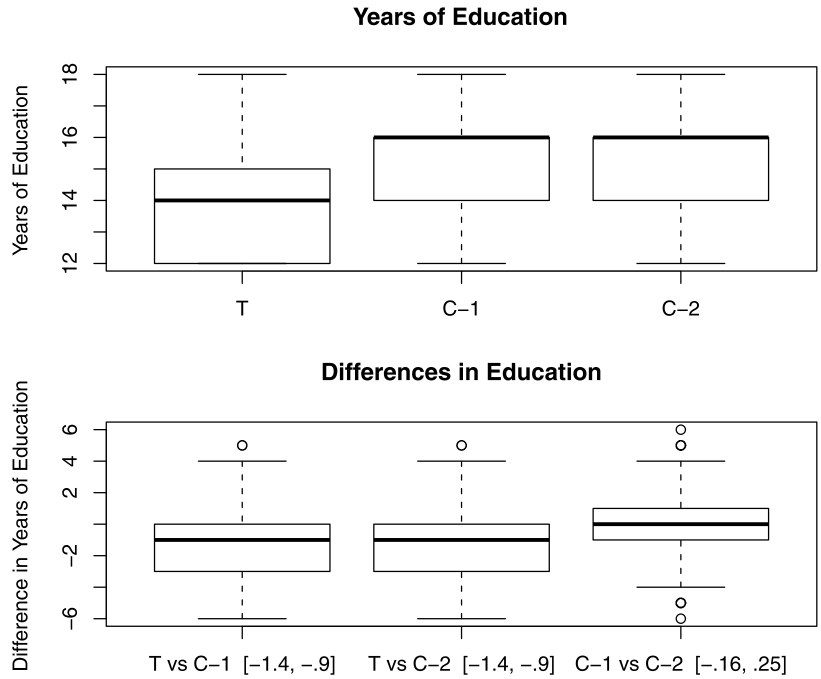

Figure 4. Years of education in the treated group (T), who started at a two-year college, and two tapered matched control groups, who started at a four-year college. The C-1 controls were matched for 20 covariates, including Region, while the C-2 control were matched for 17 covariates, excluding Region. The 95% confidence interval for the mean difference appears in brackets and is based on the paired t-statistic.

Hansen, B. B., and Bowers, J. (2008), “Covariate Balance in Simple, Stratified and Clustered Comparative Studies,” *Statistical Science*, 23, 219–236. [1,2]

Hansen, B. B., and Klopfer, S. O. (2006), “Optimal Full Matching and Related Designs via Network Flows,” *Journal of Computational and Graphical Statistics*, 15, 609–627. [4]

Hausman, J. (1978), “Specification Tests in Econometrics,” *Econometrica*, 46, 1251–1271. [9]

Heller, R., Jensen, S. T., Rosenbaum, P. R., and Small, D. S. (2010), “Sensitivity Analysis for the Cross-Match Test With Applications in Genomics,” *Journal of the American Statistical Association*, 490, to appear. [1]

Henze, N., and Penrose, M. D. (1999), “On the Multivariate Runs Test,” *The Annals of Statistics*, 27, 290–298. [7]

Imai, K., King, G., and Stuart, E. A. (2008), “Misunderstandings Among Experimentalists and Observationalists: Balance Test Fallacies in Causal Inference,” *Journal of the Royal Statistical Society, Ser. A*, 171, 481–502. [1,2]

<uncited> Lu, B. (2005), “Propensity Score Matching With Time-Dependent Covariates,” *Biometrics*, 61, 721–728.

<uncited> Lu, B., Greevy, R., Xu, X., and Beck, C. (2010), “Optimal Nonbipartite Matching and Its Statistical Applications,” *The American Statistician*, to appear.

<uncited> Neyman, J. (1923), “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9,” reprinted in *Statistical Science*, 5, 465–480.

<uncited> R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation. Available at http://www.R-project.org.

Rosenbaum, P. R. (1989), “Optimal Matching in Observational Studies,” *Journal of the American Statistical Association*, 84, 1024–1032. [4]

——— (2005), “An Exact, Distribution Free Test Comparing Two Multivariate Distributions Based on Adjacency,” *Journal of the Royal Statistical Society, Ser. B*, 67, 515–530. [1,6,7]

——— (2010), *Design of Observational Studies*, New York: Springer. [3-5]

Rosenbaum, P., and Rubin, D. (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–55. [4]

——— (1984), “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score,” *Journal of the American Statistical Association*, 79, 516–524. [5]

——— (1985), “Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score,” *The American Statistician*, 39, 33–38. [1,4]

Rouse, C. E. (1995), “Democratization or Diversion? The Effect of Community Colleges on Educational Attainment,” *Journal of Business & Economic Statistics*, 13, 217–224. [2]

Rubin, D. B. (1974), “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology*, 66, 688–701. [4]

——— (1980), “Bias Reduction Using Mahalanobis Metric Matching,” *Biometrics*, 36, 293–298. [4]

Rubin, D. B., and Stuart, E. A. (2006), “Affinely Invariant Matching Methods With Discriminant Mixtures of Proportional Ellipsoidally Symmetric Distributions,” *The Annals of Statistics*, 34, 1814–1826. [7]

Smith, H. (1997), “Matching With Multiple Controls to Estimate Treatment Effects in Observational Studies,” *Sociological Methodology*, 27, 325–353. [4]

META DATA IN THE PDF FILE

Following information will be included as pdf file Document Properties:

Title : Using the Cross-Match Test to Appraise Covariate Balance in Matched Pairs
Author : Ruth Heller, Paul R. Rosenbaum, Dylan S. Small
Subject : The American Statistician, Vol.64, No.4, 2010, 1-11
Keywords: Multivariate matching, Observational study, Propensity score, Seemingly innocuous confounding, Tapered matching

THE LIST OF URI ADRESSES

Listed below are all uri addresses found in your paper. The non-active uri addresses, if any, are indicated as ERROR. Please check and update the list where necessary. The e-mail addresses are not checked – they are listed just for your information. More information can be found in the support page: *http://www.e-publications.org/ims/support/urihelp.html*.

--- mailto:ruheller@techunix.technion.ac.il [2:pp.1,1] Check skip
200 http://www.amstat.org [2:pp.1,1] OK
200 http://dx.doi.org/10.1198/tast.2010.09210 [2:pp.1,1] OK
302 http://pubs.amstat.org/loi/tas [2:pp.1,1] Found
403 http://www.mayo.edu/hsr/sasmac.html [2:pp.9,9] Forbidden
200 http://www.R-project.org [2:pp.11,11] OK