Who Matches? Propensity Scores and Bias in the Causal Effects of Education on Participation

John Henderson University of California, Berkeley
Sara Chatfield University of California, Berkeley

In a recent study, Kam and Palmer (2008) employ propensity score matching to assess whether college attendance causes *participation after reducing selection bias due to pre-adult factors. After matching the authors find no correlation, upending a major pillar in political science. However, we argue that this study has serious flaws and should not be the basis for rejecting the traditional view of an "education effect" on participation. We match on 766,642 propensity scores and use genetic matching to recover better matches with lower covariate imbalances. We consistently find positive effects as covariate balance improves, though no matching approach yields unbiased results. We demonstrate that selection is a serious concern in studying the participatory effects of college attendance and that balance in the covariates and robustness to sensitivity diagnostics should be the ultimate guide for conducting matching analyses.*

Understanding why some people participate when others opt out is fundamental to both the study and practice of politics. Yet, after more than 50 years of research, we still possess more well-reasoned conjectures than we do well-supported conclusions about what in fact causes political participation. One of the most robust findings in political science is that going to college is strongly associated with greater participation later in life. According to the conventional view, college attendance confers long-lasting participatory returns that reduce the costs and increase the benefits of political activity (Brady, Verba, and Schlozman 1995; Campbell et al. 1960; Jennings and Niemi 1981; Miller and Shanks 1996; Rosenstone and Hansen 1993; Wolfinger and Rosenstone 1980). Indeed, the illumination of an "education effect" may constitute one of the major contributions of political science to the general body of knowledge (Schlozman 2002).[1]

In a recent study, Kam and Palmer (2008) launch a serious and timely challenge to this consensus. The authors reject the claim that college attendance *causally* impacts participation and argue that college is really a *proxy* for the many preceding life experiences that drive both educational attainment and political engagement. At issue is the fact that selection into college is nonrandom, a problem that is difficult to resolve through the use of traditional statistical approaches in observational data (Neyman 1990; Rosenbaum 2002; Rubin 2006). To overcome this obstacle, Kam and Palmer employ a propensity score matching design to assess the effect of education on participation in the now classic Youth-Parent Socialization Panel Study (Jennings and Niemi 1981). Notably, their approach potentially provides a way to account for selection confounders using an innovative method and some of the best observational data available for this question. In a major reversal from prior results, the authors find no education effect after matching. In light of this, Kam and Palmer recommend a serious reconsideration of our existing theories about participation and education.

In spite of Kam and Palmer's striking result, we believe it is premature to reject the consensus view. We argue that there are strong reasons to question the authors' null finding, most importantly that overt bias actually *increases* after matching on their estimated propensity score. Our results show that college attendance correlates with education, even after substantially reducing (though not eliminating) imbalances in the

[1]An online appendix describing the covariates used in the analysis is available at http://journals.cambridge.org/jop, and replication data and code are available at www.jahenderson.com.

The Journal of Politics, Vol. 73, No. 3, July 2011, Pp. 646–658 doi:10.1017/S0022381611000351

ISSN 0022-3816

observed characteristics that likely influence self-selection into college. However, we maintain that even these results are not free from bias. We demonstrate that selection is a major concern in analyzing the participatory returns to college education. In fact, selection may be so problematic as to make it practically impossible to recover unbiased causal estimates using even the most sophisticated matching methods as yet available. We do make major strides in reducing overt bias through an innovative combination of propensity score and genetic matching approaches. However, we cannot be assured that our findings are wholly robust to pre-adult confounding forces. Ultimately, we show that postmatching diagnostics, specifically balance tests, sensitivity checks, and match inspections, are critical to the matching enterprise in observational settings.

In the sections below, we first consider the challenge posed by Kam and Palmer (2008) to the participation literature and then discuss the Neyman-Rubin causal framework which provides the foundation for matching approaches. Next, we explore the authors' propensity score analysis in this context. We perform balance and sensitivity diagnostics on Kam and Palmer's propensity score estimate. Then we match on 766,642 propensity scores, sampled from combinations of Kam and Palmer's model, to elicit a comparable distribution of matching statistics. We find that the propensity score matching approach is highly sensitive to minor alterations in modeling choices and assumptions. Finally, we utilize a genetic matching algorithm, using the best propensity scores from the 766,642 combinations as starting points. The resulting estimates show much improvement, but still fail balance and sensitivity tests, underscoring the difficulty that self-selection poses for researchers analyzing important effects in observational data.

Competing Models of Costly Participation

Most models of political behavior start with the finding that participating is costly, and its benefits largely immaterial (Downs 1957; Wolfinger and Rosenstone 1980). The probability that a person will have a decisive impact on the outcome of an election is virtually zero. Devoting valuable time, energy, or money participating is wasteful then from a strictly self-interested view. Yet, people still participate in spite of the act's instrumental irrationality. Some culprit is moving around the costs and benefits that motivate participation, and education is a usual suspect.

Education as Cause

People participate when they possess the right resources to afford the costs of doing so, and when the benefits (whether instrumental, expressive, solidarity, or purposive), outweigh those costs (Brady, Verba, and Schlozman 1995; Wilson 1995; Wolfinger and Rosenstone 1980). In the costly participation model, college education is a wellspring of such material and immaterial resources.

Scholars typically argue that college education makes participation less costly relative to the benefits. In college, students are exposed to a wide variety of intellectual stimuli that may improve their cognitive, research, and reasoning skills. These skills facilitate the acquisition and consumption of political information and make it easier for people to navigate politics (Brady, Verba, and Schlozman 1995; Rosenstone and Hansen 1993; Wolfinger and Rosenstone 1980). College also transforms the types of benefits people receive through participating (Brady, Verba, and Schlozman 1995; Rosenstone and Hansen 1993). Campus activities and intellectual pursuits may transmit civic values, facilitate student organization and leadership, and heighten awareness of and interest in political matters. In this way, college encourages greater political involvement by attaching a noninstrumental reward to the act of participating itself, while simultaneously reducing the cost and complexity of participatory acts (Galston 2001; Schlozman 2002; Wilson 1995).[2]

Education as Proxy

Contrary to the traditional view, Kam and Palmer (2008) argue that college education provides no additional participatory resources to those who attend, but merely reflects or "proxies" those prior experiences that increase later participation. The authors suggest that a common set of pre-adult factors strongly determines both a person's future educational and participatory choices (Jennings and Niemi 1968; Kam and Palmer 2008; Sears and Levy 2003). Thus, the same resources that enable people to participate also encourage them to seek out more education, making the observed association between college and participation causally spurious.

According to Kam and Palmer's "education as proxy" model, precollege socialization and socioeconomic stratification are the primary social forces

[2]Educated individuals also participate in politics because elites recruit them into the process (Jackson 1996; Leighley 1995; Rosenstone and Hansen 1993; Schlozman 2002). However, this finding only requires that education be a useful signal for and not necessarily a cause of higher rates of participation.

that influence people's basic resources and attributes over time (Baker and Velez 1996; Grusky 2001; Luster and McAdoo 1996; Saunders 1990; Sears and Levy 2003). Modern societies are subject to considerable inequality in social and economic status, as well as in the distribution of natural talents. Those who start life with fewer advantages face steeper challenges in forming and pursuing certain goals, compared to those born under more favorable conditions (Grusky 2001; Saunders 1990). Moreover, social learning in childhood reinforces these differences by fostering the unequal development of skills, attitudes, and resources that are important in adulthood (Baker and Velez 1996; Luster and McAdoo 1996; Sears and Levy 2003). These processes differentiate people in *cognitive ability*, *personality*, *values and attitudes*, and *financial or other endowments*, which are most relevant in affecting both later education- and participation-seeking behavior (Baker and Velez 1996; Entwisle, Alexander, and Olson 2005; Jencks et al. 1972; Kam and Palmer 2008).[3]

This account is a sharp departure from the traditional finding in political science. Notably, it implies that whatever things an individual gains during her college years are unique to that person, and not to the experience itself—going to college contributes no additional abilities, values, or resources relevant to the calculus of costly participation. The ramifications of this argument for scholars and policymakers are considerable. Principally this suggests that social forces expressed earlier in life play a far more prominent role in driving adult political behaviors. Thus, inequalities that influence policy outcomes may take root much earlier and may be more difficult to address through investments at the university level (Brady, Verba, and Schlozman 1995). And if the causal effect of education on participation is spurious, then this calls into question a whole host of other participation inferences (e.g., income, occupation, union membership) that may also be confounded by selection due to these prior social forces.

We suggest that few interventions in life are as extensive and intensive as going to college. While college attendance is subject to considerable selection pressures that simultaneously affect educational and political outcomes, we argue that college may also cause greater rates of participation on top of these prior influences. The task at hand for scholars is (and has been) to address whether or not college has some *independent* effect on participation in addition to its role as a proxy for the confounding characteristics and experiences discussed above. We now turn to this task.

[3]See Kam and Palmer (2008) for a particularly superb review of these large and disparate literatures.

Neyman-Rubin Causal Framework

Statistical matching designs have their basis in a causal framework developed by statisticians Neyman and Rubin (Neyman 1990; Rubin 2006). The Neyman-Rubin causal framework, when tied to a powerful research design, enables researchers to draw persuasive causal inferences in nonexperimental settings where selection confounders are typically a problem for analysis. Essentially, causal inference is a missing data problem. Let Y_{i1} be the potential outcome for the *ith* individual if she receives treatment (e.g a drug regimen, college attendance), and Y_{i0} if she does not. The causal effect of treatment then is: $\pi_i = Y_{i1} - Y_{i0}$. However, Y_{i1} and Y_{i0} cannot both be observed, since we cannot both *give* and *not give* individual i the treatment. Instead, we must give the treatment to some people and not others and observe differences across the groups. How treatment is assigned will determine whether or not we can conclude that any observed differences between the groups are due to the treatment or to other factors (Rosenbaum 2002; Rubin 2006).

In an experimental design, treatment is assigned randomly. In doing so, we know that individuals assigned to the treatment group ($T_i = 1$) will be drawn from the same population as those assigned to the control group ($T_i = 0$). This means that the distribution of both observed and unobserved variables in the two groups should be balanced, that is, essentially similar (Rosenbaum 2002; Sekhon 2008). Since treatment assignment is the only thing that differs across the groups before the experiment, any differences we observe in an outcome generally can be attributed to treatment assignment alone. However, the choice to go to college is clearly not random and is usually based on the outcomes a person expects to receive from doing so. Randomizing college attendance for the purpose of analysis, moreover, is practically difficult and morally undesirable. Thus, we must turn to observational methods to study its effects.

Causal Inference in Observational Data

Drawing causal inferences from nonrandom treatments generally requires both strong assumptions about the nature of selection in the population and statistical adjustments to control for confounders. A necessary assumption is that selection into treatment depends only on X observable covariates (such as

parental income and education), and on no other observable or unobservable characteristics. We can then assume, conditional on X, that the potential outcomes of receiving treatment or control are not confounded with the particular treatment assignment, or $\{T \perp\!\!\!\perp Y_1, Y_0\}|X$. In other words, by conditioning on the set of observable factors that drive the choice to go to college, we can assess the independent effect of college on participatory outcomes.

This is typically called the Selection on Observables Assumption (SOA), which is assumed by most causal approaches in observational social science (Rubin 1974).[4] SOA requires that there are no unobserved or excluded characteristics that drive selection into treatment after conditioning on X. For example, a student's work ethic likely predicts his scholastic success and may be related to later political activity. But work ethic is extremely difficult to survey and measure, and there is no guarantee that conditioning on X observable characteristics will eliminate its confounding influence. Therefore, SOA must be assumed to hold after conditioning. In practice, there is no direct way to assess whether SOA is reasonable in a particular study. However, balance and sensitivity tests can provide information about how strong the selection assumption will be in a particular research design.

Formally i's observed outcome is: $Y_i = T_i Y_{i1} + (1-T_i)Y_{i0}$. Assuming SOA holds, two reasonable causal estimators are the Average Treatment Effect on the Treated (ATT):

$$\pi_{ATT} = E\{E(Y_i|T_i = 1, X_i) - E(Y_i|T_i = 0, X_i)|T_i = 1\}$$

And the Average Treatment Effect on the Controlled (ATC):

$$\pi_{ATC} = E\{E(Y_i|T_i = 1, X_i) - E(Y_i|T_i = 0, X_i)|T_i = 0\}$$

In other words, for this analysis the ATT is the difference in the average rates of participation for college attenders as compared to the average rates for nonattenders, conditioning on the X covariates in the college population. Conversely, the ATC is simply this difference in participation rates taken after conditioning on values in X for the noncollege population.[5]

[4]A second assumption that must be made is the Stable Unit Treatment Value Assumption (SUTVA). This assumption states that one individual's treatment assignment should have no effect on the outcome that results for another individual being assigned into treatment or control. See Rosenbaum (2002).

[5]Generally, these estimators will not be equal since conditioning on X selection confounders for the treatment group may produce different samples than when doing so for the control group.

Matched sampling has emerged as a popular way to condition on observed covariates to estimate these average treatment effects. This is so primarily because matching makes no functional form or distributional assumptions about the relationship between treatment and outcome, but also because interpreting its results is relatively straightforward. A simple example of exact matching illustrates this point. Suppose that individuals select into college based only upon their high school GPA and their gender. One could then match a college-bound woman who earned a B-average in high school with a B-average woman who did not attend college and observe their levels of political participation later in life. We could then do the same for the remaining female B-students, and so on, until each male and female individual is exactly matched. This procedure identifies the causal effect of college attendance on participation by matching each individual to her best counterfactual, that is, those who are identical on all relevant confounders.

Propensity Score and Genetic Matching

In reality, a variety of categorical and continuous factors influence people's choices to attend college, making exact matching of this sort impossible. Given multivariate and continuous confounders, two matching techniques are frequently utilized: matching on a propensity score and matching on multivariate distances, through a genetic algorithm or some other procedure. In propensity score matching, researchers assign a conditional probability (p-score) of receiving treatment to each individual based on the set of relevant observed covariates in X. Matching on such a propensity score will eliminate all overt bias in the sample, but *only* if we know the correct propensity score or the right model to estimate it, and thus the exact probability that each observation received treatment (Rosenbaum and Rubin 1983). In practice, we do not know this propensity score and must instead estimate it using traditional models. But, how do we know if an estimated propensity score is appropriate?

A typical approach to estimating a propensity score is to include all the potentially confounding covariates into a logit regression model (Galiani, Gerfler, and Schargrodsky 2005; Kam and Palmer 2008). However, there is no strong reason, a priori, to prefer this propensity score over any other when matching (Dehejia and Wahba 1999; Ho et al. 2007; LaLonde 1986; Smith and Todd 2001). Choosing variables that influence selection or that are theoretically relevant may be a good place to start, but the ultimate test of a good set of matches must be *balance*

in the observed characteristics of the treated and control groups and *robustness* to sensitivity checks, and not model specifications (Ho et al. 2007; Rosenbaum 2002; Rubin 2006).[6] This claim is counterintuitive and runs against the logic of parametric modeling, which typically fails in the case of omitted variables. However, in estimating a propensity score, including too many variables, even those that correlate with treatment, can actually induce overt bias in the matched samples. The reason for this is simple. Including a large number of intercorrelated variables in logit estimation may separate the propensity scores of attenders and nonattenders, thereby reducing their overlap in values at the extremes (Lesaffre and Albert 1989). Matching on this propensity score may result in a highly biased inference since a tiny fraction of controls will be matched to a very large proportion of treated observations.

An alternative to propensity score matching is genetic matching, a technique developed by Sekhon (forthcoming) that uses a genetic optimizer (Sekhon and Mebane 1998) to match individuals based on their weighted Malahanobis distances in multivariate space.[7] A form of multivariate matching is to match treateds to controls with the smallest weighted Malahanobis distances between them, given the best choice for each covariate weight. Picking the best weights and matches, however, is an extremely difficult optimization problem. The fundamental advance in genetic matching is the use of a genetic algorithm to find the weights that improve covariate balance when matching on reweighted Malahanobis distances. It works by finding improvements in the most imbalanced variables, gradually improving balance over a sequence of successive 'generations'. From a practical standpoint, genetic matching tends to improve overall balance as compared to propensity score matching. Although this technique does not necessarily guarantee that acceptable levels of balance will be achieved, nor does it guarantee that SOA will be satisfied, genetic matching does provide a more efficient way to search the interminable space to find matches that get us closer to the experimental benchmark.

[6]Good balance is important to ensure that the groups being compared are good counterfactuals for one another, at least on characteristics that we can observe. Sensitivity tests help assess whether unobserved characteristics could be driving the observed result. See Rosenbaum (2002) and Rubin (2006).

[7]Following Diamond and Sekhon (2005), let i and j be individuals and their Malahanobis distances in k-dimensional space be $d(X_i, X_j) = \left\{ (X_i - X_j)' \left(S^{-1/2}\right)' W S^{-1/2} (X_i - X_j) \right\}^{1/2}$, where W is a $k \times k$ positive definite weight matrix, with w_p (for $p = 1,2,\ldots,k$) along the diagonal elements and zeros in the off-diagonal cells, and $S^{-1/2}$ is the Cholesky decomposition of S, which is the variance-covariance matrix of X. If all w_p equal 1, then $d(X_i,X_j)$ is the k-dimensional distance between two observations standardized by the variance-covariance in X. If different weights are supplied to each w_p, then $d(X_i,X_j)$ is re-weighted to emphasize the distance between individuals along some covariates more than others. See Diamond and Sekhon (2005) for technical details on the genetic matching algorithm. Also see Ladd and Lenz (2009) and Jennings and Stoker (2008) for some recent empirical applications of genetic matching.

Sensitivity Analysis in Matching Designs

All matching designs must make choices over a variety of specifications used to find and assess "good" matches, and there is no universally agreed upon way to choose among these alternatives. For instance, researchers must decide how many control observations to match to each treated observation, which covariates to match and test balance on, and which loss function to use when running genetic matching, among others. Different choices will typically result in different sets of matches and levels of observed and unobserved bias. This suggests that caution and judgement should be used when interpreting any matching estimate as causal.

Sensitivity analysis can help reduce this problem, however, by assessing whether a resulting estimate is robust to bias due to remaining imbalances in any observed or excluded covariates after matching (Rosenbaum 2002). Specifically, the test analyzes how large the difference in the underlying probability of receiving treatment would have to be in order to change the substantive interpretation of a matching estimate. In the Rosenbaum test, different levels are set for Γ, the log odds of receiving treatment. The test then assesses the lower and upper bounds of a matching estimate when one observation in a matched pair is allowed to have a higher probability of receiving treatment because of confounding observed or unobserved factors. If an estimate remains bounded above zero, at say $\Gamma = 3$, then one observation in a matched pair could be three times as likely to have received treatment without eliminating the observed effect of that treatment. Typically, these bounds are estimated for increasing levels of Γ. If the bounds include zero at low levels of Γ, then the estimate should be considered highly sensitive to bias. We consider these tests crucial in any matching research design since they provide information about the quality of matches regardless of the choices made to obtain them.

Nonexperimental Treatments

Finally, since controlled manipulation is impossible in observational analysis, it is important at an initial

stage to clearly define the treatment that will be measured and studied. In a classic statement, Rubin (1986) argues that there is "no causation without manipulation," which strongly implies that treatments that cannot be manipulated, at least theoretically, cannot be studied in the causal framework (Holland 1986; Rubin 1986). This logic originates in the fundamental insight of the potential outcomes approach. If we cannot imagine realistic situations in which person i could potentially receive either treatment or control, then either Y_{i1} or Y_{i0} is undefined, and π_i does not exist.

Considering some education choices as manipulations is problematic. College *graduation*, for example, would theoretically be difficult to manipulate, and its effects difficult to interpret. In order to graduate, students must invest time and effort passing classes and meeting degree requirements. It is generally impossible to randomly manipulate how a student allocates their activity in achieving this degree. Further, randomizing graduation certificates directly is very different from randomizing the event of actually graduating. Unlike graduation, we argue that college *attendance* can more easily be analyzed in the context of a randomized experiment. For instance, a researcher could randomly assign a college scholarship or admissions counselor to high school seniors and observe later participation outcomes. Alternatively, we could observe a natural experiment in which an admissions board decides to randomly admits students who tie on test scores and other admissions qualifications.[8] These experiments are not perfect analogies for studying the effects of college attendance through observational designs, but they illustrate that we are able to define treatment and control groups such that non-attenders could theoretically have attended college (and vice versa) under the assignment

Assessing the "Education as Proxy" Hypothesis

The main task in this analysis is to identify the set of matches that contain the least amount of bias in observed covariates as the basis for evaluating whether college education is primarily a proxy for or a cause of political participation. If we observe that these matches correspond to consistently zero findings when estimating the "education effect," then indeed we agree that the consensus view should be revisited. Alternatively, if Kam and Palmer's null estimate is found to be subject to serious bias, and if positive effects are consistently recovered when overt bias is reduced or eliminated, then we may reject the claim that education is merely a proxy for pre-adult factors.

In this vein, we start by replicating and extending Kam and Palmer's original analysis of college attendance among high school seniors in the Youth-Parent Socialization Panel Study (Y-PSPS). Due to its panel structure and extensive set of survey instruments, this data is ideal for research designs using matching to analyze the causal effects of college attendance on later political participation (Jennings et al. 2004).[9] We evaluate Kam and Palmer's propensity score analysis from the 1965 and 1973 waves of Y-PSPS, making as few deviations from their research design as possible.[10] Treatment is a dichotomous measure of college attendance by the 1973 survey, and the outcome is an additive index of political participation variables taken in 1973. Specifically, the outcome index consists of voting in the 1972 election, attending a campaign rally or meeting, wearing a campaign button, working on a campaign, donating money, contacting a representative, going to a demonstration, and participating in local politics.

In their analysis, Kam and Palmer perform 1 to 3 matching to estimate the ATT, which takes each college attender and matches him to the three nearest neighbors in the nonattending control group based on the predicted probabilities estimated in the propensity score model.[11] The model includes the 81 covariates that Kam and Palmer consider most relevant to selection into college, each transformed into factors. After matching, they find

[8]A variation on this would be a regression discontinuity design to study participation outcomes for students just above and below the point cutoffs used in admissions at many large universities.

[9]Y-PSPS contains interviews from students and their parents in 1965 and then reinterviews with the same respondents in 1973, 1982, and 1997 that measure college attendance, political participation, and a battery of demographic, social, and attitudinal questions. Given the panel structure, we can be assured that the baseline covariates in 1965 are pre-treatment, as they were all measured prior to any students having attended college. Baseline covariates are particularly rich in this dataset and tap key attributes (e.g., cognitive ability, personality, attitudes, and endowments) that likely affect education choices.

[10]We follow Kam and Palmer (2008) in eliminating students with no parental data and those who dropped out of the survey by the second wave. For those students who had interview data from both parents, we randomly selected one parent to include in our dataset. We focus our attention on the authors' model for the 1973 Y-PSPS survey. Our substantive findings do not change when we replicate Kam and Palmer's analysis for the 1982 survey, and so we do not report these figures here.

[11]We utilize 1 to 3 matching in our analysis to ensure comparability between Kam and Palmer's results and our own. In tests, we did not find significant differences between 1 to 1 and 1 to 3 matching.

a zero effect of college attendance on the participation of college attenders (See Table 1 below).

After replicating Kam and Palmer's original 81-covariate propensity score model, we recover their estimate for the college treatment effect on participation. We then test for balance in the matched population on 109 covariates transformed into 355 dummies.[12] Surprisingly, we find that overall balance actually gets much worse. Before matching, 41% of the 355 dichotomized covariates are balanced at the $p \geq 0.1$ level.[13] After matching, only 25% of the covariates are similarly balanced. This raises the concern that overt bias has actually *increased* after matching on the propensity score. However, achieving balance on some covariates may be more important than on others, in particular, those that are better predictors of college attendance. For instance, we might believe that parents' income strongly influences the educational and participatory outcomes of their children, but worry less about a student's participation in high school sports. Does the propensity model achieve balance on more important baseline covariates?[14]

While balance does improve on some important covariates after matching on Kam and Palmer's propensity score, most of the crucial ones still remain imbalanced. As Table 2 shows, after p-score matching, imbalances remain in both key parental (e.g. PARTICIPATION, EMPLOYMENT, EDUCATION, INCOME, and PARTY ID) and student (e.g. HIGH SCHOOL GPA, GENDER, and PARTY ID) covariates. This covariate imbalance, especially in the most important predictors, gives us strong reasons to suspect that Kam and Palmer's resulting ATT estimates are seriously biased and that major selection confounders remain uncontrolled (and may have even gotten worse) after matching on the propensity score. In the absence of extremely strong assumptions, we cannot interpret this particular null finding as the correct causal inference of education on participation.

We perform additional robustness tests to assess the sensitivity of the ATT estimate after dropping college and noncollege observations that may be difficult to match due to their distance from possible matches at the extremes of the propensity score. In logit estimation, both predicted probabilities and linear predictors tend to cluster observations at the tails of the resulting distribution, especially as greater numbers of intercorrelated variables are added to the logit model. Kam and Palmer do test for this problem by applying a caliper of 0.25 standard deviations ($\sigma = 0.3512$) during matching, which deletes all college-attenders who have predicted probabilities that are are farther than ± 0.0878 from the nearest non-attenders. They also delete observations with the 5% highest and 5% lowest propensity scores. Yet, these tests are insufficient as each college-attender has *at least* one nonattender within the respective bounds.

The problem, however, is that there is *only* one nonattender within this bound for a large number of the attenders to be matched to, especially for attenders near the extremes of the propensity score. Figure 1 displays the densities of Kam and Palmer's propensity scores for attenders and nonattenders. In examining these, we observe that clustering around 1 is so pronounced that the p-scores for the top 5% of college-attenders range in value from .9998174 to .9999998. In fact, over half of all attenders have propensity scores greater than .9 and over a quarter have scores greater than .99. In contrast, only 14 nonattenders have propensities greater than .9 (3%) and only 5 (1%) have propensities greater than .95, with the highest propensity for any nonattender being .9889.

As a consequence of this pattern of clustering at the tails, an extremely small number of nonattenders are matched to a large proportion of college-attenders. In Kam and Palmer's matched sample for 1973, the five nonattenders with the highest predicted probabilities (.989, .982, .974, .964, and .953) are matched to about 44% of the college-attenders (respectively, 287 or 11.7%, 316 or 12.9%, 335 or 13.7%, 70 or 2.8%, and 70 or 2.8%). Further, three of these five nonattenders are big outliers in terms of their participatory behavior. The average rates of participation in the 1973 survey are 2.79 participatory acts for all college attendees and 1.43 acts for all noncollege respondents. Individuals who attended college and who also have predicted propensities greater than .95, engaged in 3.26 participation deeds on average. Yet, for the five nonattenders with predicted probabilities above .95, the average participation rate is 4.00 activities. Since these five outlier nonattenders get matched to almost half of the attenders in the data, it is no surprise that a null effect is recovered. Indeed, dropping just these five (1.3%) nonattenders

[12]This includes the 81 covariates used in estimating the propensity score, plus 28 additional covariates included for comprehensiveness. Generally, it is wise to check balance on as many relevant covariates as are available in the data, but ultimately adding the 28 additional covariates and dichotomizing them does not change the balance results presented here.

[13]These are t-test p-values, since the variables are all dichotomous. The $p \geq 0.1$ level is chosen out of convention and not due to any deeper justification.

[14]We follow Kam and Palmer's identification of 16 key covariates for the purposes of comparability across papers. These covariates are also theoretically important as discussed in the secondary literature and generally correlate with college attendance, participation levels, or both.

TABLE 1 Matching Estimates of the Causal Effect of College Attendance on Participation

	KP P-Score		GenMatch	
	ATT	*Outliers Dropped*	ATT	ATC
Estimate	0.023	1.340	1.020	0.626
AI Standard Error	0.469	0.519	0.166	0.154
P-Value	0.961	0.007	0.000	0.000

from the analysis results in a positive college treatment estimate (see the second column of Table 1), undermining the robustness of Kam and Palmer's finding.

Large Sample Distribution of Propensity Score Models

The above discussion shows that the particular propensity score matches Kam and Palmer use to estimate college treatment effects fail both balance diagnostics and robustness checks. Given these vulnerabilities, should we reject the "education as proxy" hypothesis altogether? Would we still recover a null effect after matching on a p-score model that produces better balance and is robust to perturbations in model specifications? More generally, what sorts of estimates do we recover as covariate balance improves?

We address these questions by matching on 766,642 propensity score models, sampling from all linear combinations of the original 81-variable model, to recover a more general distribution of propensity score estimators and balance statistics.[15] Most importantly, the resulting distribution allows us to assess whether or not a zero college treatment effect is a rare or robust finding when balance in the covariates improves. This distribution may also give us some insight on the utility of propensity score matching for this question and may illuminate how troubling selection into college is in studying its effects on participation.

In running these models, we again deviate as little as possible from Kam and Palmer's basic research design, changing only which combination of the 81 covariates gets included in the logit model. Thus, we again factorize each variable and conduct 1 to 3 matching on the predicted probability to estimate the ATT for college attendance on the participation outcome index. For each of the models, we record the resulting estimate, and the associated p-value and balance statistics (the proportion of the 355 dichotomized covariates and the 16 'key' variables balanced at the $p \geq 0.1$ level), making these findings directly comparable to Kam and Palmer's propensity score approach.

The results in Figure 2 allow us to put the "education as proxy" hypothesis in larger context. A sizeable majority (76.4%,) of all the propensity models produce an estimate that is positive and significant. Kam and Palmer's estimate (indicated by Δ) accompanies about a quarter of the models in recovering the less frequent (though not rare) null effect. However, over 98.8% of the propensity matches produce better *overall* levels of balance compared to Kam and Palmer's model, and 32 of the propensity score models (indicated by ●) actually recover monotonic improvements on *all* of the 16 key covariates, though only three of these report significant estimates. In terms of balance, Kam and Palmer's model does well on a few difficult but important variables (e.g., STUDENT POLITICAL KNOWLEDGE, STUDENT COLLEGE PLANS, and PARENTAL POLITICAL PERSUASION), yet performs relatively poorly on the rest.

Overall, the best propensity score models only achieve 63% covariate balance. Moreover, until balance generally climbs above 50%, it seems that any possible result (positive, negative, zero) may be

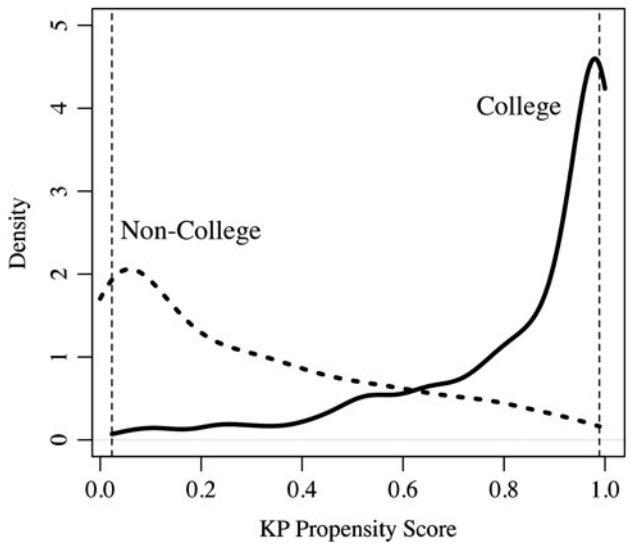
[15]Sampling from all combinations of 81 covariates poses two challenges. First, the total number of all possible combinations is extremely large: $2^{81} = 2.42 \times 10^{24}$. Second, we suspect that the sheer number of variables included in propensity score estimation may impact the resulting matches. Uniformly sampling from all possible combinations either would require an enormous sample size or would exclude all combinations at most discrete levels. For instance, with a sample size of 1 million and uniform probability, the expected number of samples from the $\binom{81}{n}$ combinations is 2 or less, for all $n = (1,2, \ldots, 20, 61, 62, \ldots, 81)$. To get around these problems, we employ a stratified uniform sampling method. In this approach, we sample 10,000 variable combinations at each level taken uniformly within that level to construct our population of propensity models. Therefore, for the outcome in 1973, we sample 10,000 combinations (or all of the combinations if less than 10,000) each taken uniformly at the $\binom{81}{1}$ level, the $\binom{81}{2}$ level, the $\binom{81}{3}$ level and so on, all the way up to the $\binom{81}{80}$ level. This produces a total sample population of 766,642 propensity score models of the effect of college attendance on the 1973 political participation index.

TABLE 2 Covariate Balance of College Attenders and Nonattenders Before and After Matching

	Baseline	KP P-Score	GenMatch	
		ATT	ATT	ATC
Student GPA	0.000	0.000	0.005	0.141
Student Gender	0.000	0.000	0.000	0.891
Student Race	0.134	0.326	0.437	0.229
Student Replublican Party ID	0.000	0.000	0.000	0.324
Student Knowledge Index	0.000	0.000	0.000	0.093
Student College Plans	0.000	0.306	0.000	0.000
Parent Vote Participation	0.000	0.148	0.151	0.154
Parent Political Persuasion	0.000	0.833	0.062	0.854
Parent Participation Index	0.000	0.000	0.037	0.137
Parent Employment	0.003	0.079	0.150	0.670
Head of Household Education	0.000	0.000	0.000	0.179
Wife's Education	0.000	0.000	0.000	0.140
Parent Income	0.000	0.000	0.000	0.000
Parent Homeownership	0.000	0.000	0.175	0.105
Parent Republican Party ID	0.000	0.000	0.001	0.078
Parent Knowledge Index	0.000	0.000	0.000	0.001

obtained, though the range of coefficients gets larger as overall balance declines. Figure 2 (a) shows the relationship between levels of covariate balance achieved by the propensity score models and the corresponding ATT estimates. One observable trend is that as balance increases, the estimates tend to be increasingly bounded above zero and convergent towards one—that is, for individuals who attended college, the effect of that attendance converges on an average of one additional participatory act as compared to those who did not attend college. When balance is greater than 0.55, *all* of the ATT estimates of college attendance on participation are positive and significant as shown in Figure 2 (b). However, the key finding is that none of the 766,642 propensity models comes close to recovering the type of balance we would expect from randomization.

FIGURE 1 Density Plot of Kam and Palmer (2008) Propensity Score

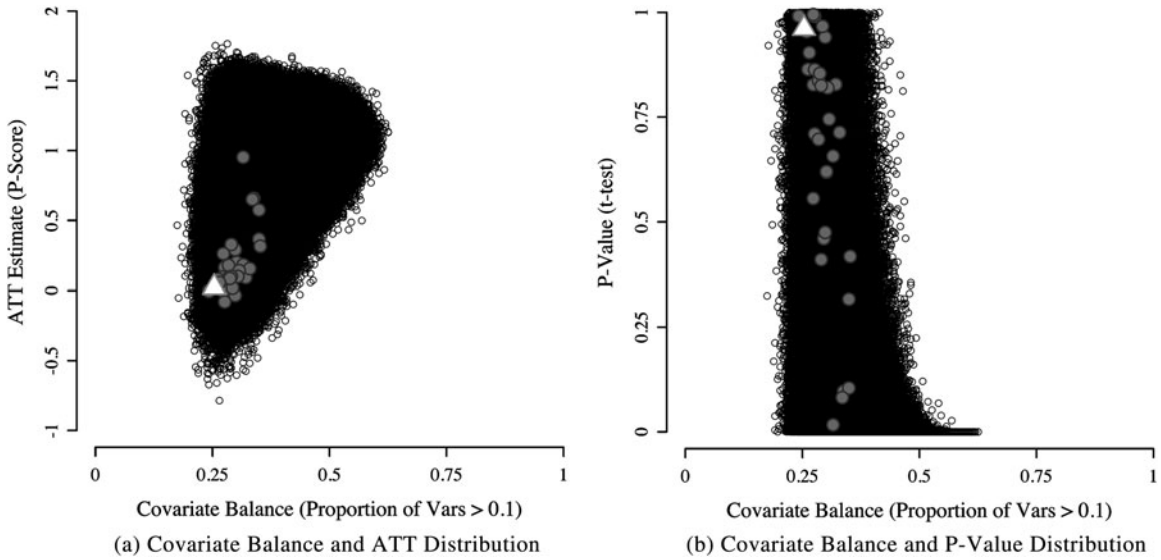


The critical shortcoming to this sampling approach is that it is impossible to know what matching estimates would emerge from models that could achieve balance greater than 0.63. In other words, the propensity score balance distribution is truncated from above. The source of this truncation could be that we happened to sample all the wrong propensity score models. After all, this is an astronomically small fraction of all the possible combinations we could have estimated. Further, we could have explored other modeling specifications (e.g., not factorizing the variables, using interactive terms). We argue that limitations in propensity score matching and the deeply rooted problem of selection are likely the biggest culprits impeding the ability to get good matches. Thus, we turn to genetic matching to obtain matches with better covariate balance.

Genetic Matching to Reduce Overt Bias

Propensity score matching, as a general approach, falls short as a way to control selection confounders

FIGURE 2 **Sampling Distribution of Propensity Score Matching ATT Estimates and Covariate Balance**



(a) Covariate Balance and ATT Distribution (b) Covariate Balance and P-Value Distribution

that affect both participatory and educational outcomes, at least in the Y-PSPS data. As discussed above, genetic matching allows us to control for confounding factors more efficiently than propensity score matching when estimating college treatment effects. Thus, it may allow us to overcome the selection problem by finding better matches that improve covariate balance as compared to that obtained in any propensity score model.

In our analysis, we use the best two propensity scores from the 766,642 models as starting points for genetic matching. Both of these sets of matches resulted in 63% covariate balance and an ATT estimate of 1.13 additional participatory acts for college attenders, significant to the $p < .001$ level. After running the genetic matching algorithm with these two top propensity scores as starting points, things improve with 71% of all covariates balanced. The estimate recovered from this genetic matched sample is 1.02 additional participatory acts for college attenders, again significant to the $p < .001$ level. (See the third column of Table 1 and Table 2 for the genetic matching ATT estimate and balance statistics.) Genetic matching reduces overt bias, but does not eliminate it here, since nearly 30% of the covariates still remain imbalanced. After conducting Rosenbaum sensitivity tests for this set of matches, we observe that the estimate is no longer bounded above zero when $\Gamma = 3$. Although our finding has the lowest overt bias of any ATT estimate of college attendance in this analysis, it still remains fragile due to remaining overt and hidden bias.

A potential difficulty in estimating the ATT of college on participation in the Y-PSPS data is that there are twice as many attenders as nonattenders in the sample. It may be easier to obtain good matches when estimating the ATC instead of the ATT, simply because there are more potential counterfactuals to choose from when matching. The ATC quantity is also theoretically interesting both from a scholarly and a policy standpoint. For instance, governments or universities may wish to invest resources to encourage nonattenders to go to college for the purposes of reducing political or participatory inequality.

We repeat our above analysis for the ATC, similarly running 766,642 propensity scores on the same set of randomly selected combinations of the 81 covariates to match three attenders to each nonattender. We again select the best two propensity scores from this analysis as starting points for genetic matching and present these results in the fourth columns of Table 1 and Table 2. In our final ATC estimate, 90% of all the covariates are balanced, which is a vast improvement over the levels of balance achieved in estimating the ATT, especially for difficult to balance covariates. In addition, 11 of 16 key covariates are balanced, including STUDENT GPA, STUDENT GENDER, WIFE'S EDUCATION, and HEAD OF HOUSEHOLD EDUCATION. The estimate of the effect of college for nonattenders is 0.63 additional participatory acts, significant at the $p < 0.001$ level. Finally, we examine the sensitivity of this estimate to remaining bias using Rosenbaum bounds and again find that our result fails the test at low levels

($\Gamma = 2$). This indicates that despite significant reductions in overt bias, hidden bias remains a serious problem in recovering a causal estimate of the education effect on participation.

Conclusion

College is by definition selective. Attending college requires students be able to afford its costs, withstand its rigors, and value its purposes. Naturally then, people who are born into favorable circumstances and further develop certain attitudes and abilities throughout pre-adulthood will possess great advantages over others when deciding to apply and being admitted to college. It is hard to deny that these differences play a major role in determining who attends college.

On the other hand, participation in democracy is inclusive. In the United States, with the exception of felony restrictions, suffrage is universal for all adult citizens. And most other forms of participation, like writing letters or protesting, are open to all who reside in the country. Nonetheless, stark differences between participants and nonparticipants persist, and these correlate strongly with educational attainment. Certainly many of the skills, resources, and values necessary to navigate the political arena are developed before reaching college age. However, political scientists traditionally agree that college also contributes unique resources and benefits that motivate individuals to participate.

Kam and Palmer (2008) present a major challenge to this conventional view. After matching on a propensity score, the authors find that college does not correlate with participation. Although their matching approach does control for some confounders, it ultimately makes overall overt bias worse. As a consequence, we show here that the extensive theoretical and empirical work on this relationship should not be upended by one flawed study. Kam and Palmer provide a rich theoretical account that demonstrates the plausibility that education serves, at least partially, as a proxy for earlier life experiences. However, they are ultimately unable to convincingly show that the ''education as proxy'' hypothesis is sufficient to explain the observed correlations between college and participation. Our analysis consistently recovers this positive correlation after significantly reducing levels of overt bias. In addition, our investigation of 766,642 propensity score models indicates that as balance improves a pattern of positive findings seems to emerge. Although this study cannot conclusively show that a positive causal effect exists because of remaining overt and hidden bias, we do believe that these results caution against a rejection of the conventional causal theoretical model on the basis of Kam and Palmer (2008).

One major reason that we were unable to recover a robust causal estimate is the degree and complexity of selection into college. Kam and Palmer significantly advance work in this area by reigniting the debate over selection confounders in estimating the effects of education on participation. We too find that even when observed covariates are reasonably well balanced in the ATC matches, selection on the basis of unobserved covariates remains a serious problem. Therefore, interpreting *any* matching estimate as the unbiased causal inference requires strong assumptions, most significantly that persistent imbalances in the covariates are uncorrelated with the true effects of college on participation *and* that unobserved factors are not influencing self-selection into college attendance. Such assumptions are difficult to sustain in light of our theories and expectations about college education. Given the considerable selection forces inherent in educational attainment, we conclude that robustness and sensitivity checks are essential in studying its effects. Thus, caution should be used in interpreting the results of any matching analyses that do not employ these tests where selection into treatment is not random.

We argue that the best guide to judging a matching design is not a defense of the specific matching technique used but rather balance in the covariates and robustness to sensitivity checks after matching. When applying these techniques to other observational data, we suggest that future studies examine a large number of propensity score and genetic matching specifications, ideally through a semiautomated process similar to that developed in this paper. Genetic and propensity score matching can be combined using the best propensity scores as starting points in the genetic matching algorithm, which may improve the efficiency and quality of the final results. Integrating these methods may allow researchers to significantly reduce the confounding effects of pretreatment covariates in biasing their estimates, especially as compared to parametric modeling or either matching approach alone.

The degree of self-selection into college makes directly investigating the education effect problematic, even when employing sophisticated matching techniques on some of the best data currently available. This difficulty suggests a return to research design, theory building, and data collection. Because the direct

effect of college education on participation is difficult to identify, it is important to extend existing theory to develop new observable implications and novel empirical tests. For instance, scholars might focus on the chain of linkages hypothesized by researchers and break apart each component into a unique, testable form, such as the connection between education and civic values or increased cognitive abilities. Another possibility would be to investigate different features of college that could give us a grip on which aspects of this experience might lead to different sorts of participation. For instance, an early trajectory of coursework in the social sciences versus the hard sciences might differentiate students in their levels of civic or participatory values, but not necessarily cognition. Finally, researchers might look for natural experiments that allow them to gain leverage on this question, such as the admissions point system example discussed above. Ultimately, we argue that education is more than a proxy, but that additional work in this vein is needed to conclusively pin down the precise education effect and its implications for costly participation in policy and theory.

Acknowledgement

We thank Jas Sekhon, Eric Schickler, Henry Brady, Claudine Gay, Jonathan Wand, Rocio Titiunik, and Erin Hartman for their invaluable advice over the course of this project. We also thank Cindy Kam and Carl Palmer for generously making code and data available, and our anonymous reviewers for their thoughtful comments.

References

Baker, Therese, and William Velez. 1996. "Access to and Opportunity in Postsecondary Education in the U.S.: A Review." *Sociology of Education* 69: 82–101.

Brady, Henry, Sidney Verba, and Kay Schlozman. 1995. "Beyond SES: A Resource Model of Political Participation." *American Political Science Review* 89 (2): 271–294.

Campbell, Angus, Philip Converse, Warren Miller, and Donald Stokes. 1960. *The American Voter*. Chicago: University of Chicago Press.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94: 1053–1062.

Diamond, Alexis, and Jasjeet Sekhon. 2005. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." Working Paper. Available at http://www.sekhon.berkeley.edu/.

Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York: Harper and Row.

Entwisle, Doris, Karl Alexander, and Linda Olson. 2005. "First Grade and Educational Attainment by Age 22: A New Story." *American Journal of Sociology* 110 (5): 1458–1502.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky. 2005. "Water for Life: The Impact of the Privatization of Water Services on Child Mortality." *Journal of Political Economy* 113 (1): 83–120.

Galston, William. 2001. "Political Knowledge, Political Engagement, and Civic Education." *Annual Review of Political Science* 4: 217–34.

Grusky, David, ed. 2001. *Social Stratification: Class, Race, and Gender in Sociological Perspective*. Boulder Co: Westview Press.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15 (3): 199–236.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 954–60.

Jackson, Robert. 1996. "A Reassessment of Voter Mobilization." *Political Research Quarterly* 49: 331–49.

Jencks, Christopher, Marshall Smith, Henry Acland, and Mary Jo Bane. 1972. *Inequality: A Reassessment of the Effect of Family and Schooling in America*. New York: Basic Books.

Jennings, M. Kent, and Richard Niemi. 1968. "The Transmission of Political Values from Parent to Child." *American Political Science Review* 62 (1): 169–184.

Jennings, M. Kent, and Richard Niemi. 1981. *Generations and Politics: A Panel Study of Young Adults and Their Parents*. Princeton, NJ: Princeton University Press.

Jennings, M. Kent, Gregory B. Markus, Richard G. Niemi, and Laura Stoker. 2004. "Youth-Parent Socialization Panel Study, 1965-1997: Four Waves." Ann Arbor: University of Michigan, Center for Political Studies/Survey Research Center. Available at http://www.icpsr.umich.edu/.

Jennings, M. Kent, and Laura Stoker. 2008. "Another and Longer Look at the Impact of Higher Education on Political Involvement and Attitudes." Presented at the annual meeting of the Midwest Political Science Association.

Kam, Cindy, and Carl Palmer. 2008. "Reconsidering the Effects of Education on Political Participation." *Journal of Politics* 70: 612–31.

Ladd, Jonathan, and Gabriel Lenz. 2009. "Exploiting a Rare Communication Shift to Document the Persuasive Power of the News Media." *American Journal of Political Science* 53 (2): 394–410.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *The American Economic Review* 76 (4): 604–20.

Leighley, Jan. 1995. "Attitudes, Opportunities and Incentives: A Field Essay on Political Participation." *Political Research Quarterly* 48 (1): 181–209.

Lesaffre, Emmanuel, and Adelin Albert. 1989. "Partial Separation in Logistic Discrimination." *Journal of the Royal Statistical Society* 51 (1): 109–16.

Luster, Tom, and Harriette McAdoo. 1996. "Family and Child Influences on Educational Attainment: A Secondary Analysis of the High/Scope Perry Preschool Data." *Developmental Psychology* 32 (1): 26–39.

Miller, Warren, and J. Merrill Shanks. 1996. *The New American Voter*. Cambridge, MA: Harvard University Press.

Neyman, Jerzy. 1990. ''On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.'' *Statistical Science* 5 (4): 465–72.

Rosenbaum, Paul R. 2002. *Observational Studies*. 2nd ed. New York: Springer.

Rosenbaum, Paul, and Donald Rubin. 1983. ''The Central Role of the Propensity Score in Observational Studies for Causal Effects.'' *Biometrika* 70 (1): 41–55.

Rosenstone, Steven, and John Hansen. 1993. *Mobilization, Participation, and Democracy in America*. New York: Longman Publishing.

Rubin, Donald. 1974. ''Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies.'' *Journal of Educational Psychology* 66: 688–107.

Rubin, Donald B. 1986. ''Statistics and Causal Inference: Which Ifs Have Causal Answers.'' *Journal of the American Statistical Association* 81: 961–62.

Rubin, Donald B. 2006. *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press.

Saunders, Peter. 1990. *Social Class and Stratification*, London: Routledge.

Schlozman, Kay. 2002. ''Citizen Participation in America.'' In *Political Science: State of the Discipline*, ed. Ira Katznelson and Helen Milner. New York: W.W. Norton, 433–6.

Sears, David, and Sheri Levy. 2003. ''Childhood and Adult Political Development.'' In *Oxford Handbook of Political Psychology*, ed. David Sears, Leonie Huddy, and Robert Jarvis. Oxford: Oxford University Press, 447–56.

Sekhon, Jasjeet. 2008.''The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods.'' In *The Oxford Handbook of Political Methodology*, ed. Janet Box-Steffensmeier, Henry Brady, and David Collier. Oxford: Oxford University Press, 271–99.

Sekhon, Jasjeet. Forthcoming. ''Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R.'' *Journal of Statistical Software*. Available at http://sekhon.berkeley.edu/matching/.

Sekhon, Jasjeet, and Walter R. Mebane. 1998. ''Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models.'' *Political Analysis* 7: 189–213.

Smith, Jeffrey A., and Petra E. Todd. 2001. ''Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods.'' *The American Economic Review* 91 (2): 112–18.

Wilson, James. 1995. *Political Organizations*. New York: Basic Books.

Wolfinger, Raymond, and Steven Rosenstone. 1980. *Who Votes?* New Haven, CT: Yale University Press.

John Henderson is a Ph.D. Candidate at the University of California, Berkeley, Berkeley, CA 94720.

Sara Chatfield is a Ph.D. Candidate at the University of California, Berkeley, Berkeley, CA 94720.