# Using Linear Regression and Propensity Score Matching to Estimate the Effect of Coaching on the SAT

**Ben Domingue** **Derek C. Briggs**
University of Colorado

Using observational data from the Education Longitudinal Survey of 2002, the effect of coaching on the SAT is estimated via linear regression and propensity score matching approaches. The key features of taking a propensity score matching approach to support causal inferences are highlighted relative to the more traditional linear regression approach. A central difference is that propensity score matching restricts the sample from which effects are estimated to coached and uncoached students that are considered comparable. For those students that have taken both the PSAT and SAT, effect estimates of roughly 11 to 15 points on the math section and 6 to 9 points on the verbal are found. Only the math effects are statistically significant. We found that coaching is more effective for certain kinds of students, particularly those who have taken challenging academic coursework and come from high socioeconomic backgrounds. In the present empirical context the summary causal inference being drawn does not depend much upon whether the effect is estimated using linear regression or propensity score matching.

The SAT plays a high stakes role in the American college admissions process. As a consequence, a commercial industry has emerged to prepare students to take the SAT. According to The College Board, the SAT "tests students' knowledge of subjects that are necessary for college success" (College Board, 2009). If short-term preparation (i.e., "coaching") can be linked to substantial score increases, then the claim that the SAT measures knowledge that is developed gradually over years of schooling becomes equivocal. Moreover, if coaching services, some of which are quite expensive, give those who can afford them a sizable boost in their SAT scores, then this would further tilt an already uneven playing field when it comes to college admissions. One purpose of this study is to gauge the size of the coaching effect using data from a recent national cohort of high school students.

From an experimental design perspective, the most accurate method of estimating the causal effect of coaching would be through the use of a randomized experiment. In principle, randomization allows for relatively straightforward estimation of causal effects since, when done correctly, the process of randomization ensures that differences between experimental treatment and control groups on some outcome of interest can be attributed to the treatment, and not to some other variable. In short, a randomized experiment controls for confounding by design. Of course, randomized experiments are both expensive and hard to implement on a large scale. In this study we make use of observational data taken from the Educational Longitudinal Survey of 2002 (ELS:02). The ELS:02 data is strictly observational in the sense that students self-select to participate in test preparation programs. This presents substantial complications in drawing causal inferences since score differences between students who do and do not receive coaching may be confounded by preexisting differences on variables correlated with both coaching status and SAT performance. To account this we use linear regression (LR) and propensity score matching (PSM) in an attempt to control statistically for observable confounders in the process of estimating causal effects. A second purpose of this study is to compare and contrast the use of these two methods to support causal inferences in an observational setting.

We begin by describing the existing research on SAT coaching and highlight some methodological differences and similarities between LR and PSM. In the next section, we describe the ELS:02 data, presenting descriptive statistics for coached and uncoached students. Since propensity score matching is an umbrella term that encompasses a variety of different analytical procedures, we present the specifics of the two PSM approaches we will be invoking. Sections that focus on our empirical results, and the sensitivity of these results to our modeling assumptions follow. Finally, we compare this research to other studies in the SAT coaching literature and make a few methodological recommendations regarding the use of PSM to support causal inferences in education research contexts.

## Background

Since 1953 there have been more that 30 studies conducted to evaluate the effect of coaching on specific sections of the SAT (Briggs, 2002). While one might assume from this that the empirical effectiveness of coaching on SAT performance has been well-established, this is only somewhat true. One principal reason for this is that the vast majority of coaching studies conducted over the 40 year period between 1951 and 1991 tended to involve small samples that were not necessarily representative of the national population of high school seniors taking college admissions exams, and of the programs offering test coaching. In addition, a good number of these studies contained a variety of methodological flaws that compromised the validity of their conclusions. To date, the findings from analyses that are at both methodologically sophisticated and generalizable have indicated that the effect of coaching is about 10-20 points on the math section of the SAT and and 5-10 points on the verbal section of the exam (Powers & Rock, 1999; Briggs, 2001; Briggs, 2002).

There are two principal motivations for conducting a new study to estimate the effect of SAT coaching using data from ELS:02. First, the SAT has undergone substantial changes to its format over the years (Lawrence et al., 2004). The SAT administered to students in the ELS:02 sample in 2003-2004 was considerably different in format than the SAT administered from 1953 through 1992 (the period during which the bulk of coaching studies were conducted). In particular, the types of items thought to be most coachable in the critical reading and math sections have been replaced (antonyms and analogies in the critical reading section; quantitative comparisons in the mathematics section). Because of this it is conceivable that the SAT has become less coachable over time, and this is one hypothesis we are testing in the present study.

Second, a relatively new method of estimating causal effects from observational data—propensity score matching (PSM)—has become increasingly popular over the past decade. The studies of Briggs (2001) and Powers and Rock (1999) both illustrate the classic approach of drawing inferences from observational data using a linear regression model (although both studies did use other methods as well): A single dummy variable represents treatment status and is included in a regression alongside other variables thought to be confounders. The estimated coefficient on the treatment variable represents the causal effect of coaching. There are, however, some clear limitations to the estimation of causal effects using linear regression. The biggest is that the method generally assumes that all potentially confounding variables have been measured without error and properly included in the model's specification (for details on the assumptions of the linear regression model in the context of estimating a coaching effect, see Briggs, 2004). Beyond this, the approach makes strong parametric assumptions. When many covariates are included in the model, the analyst is often implicitly relying upon assumptions of linearity and extrapolations well beyond observed variable combinations to correct for differences between treated and control units.

Rosenbaum and Rubin (1983) originally developed the idea and theoretical justification for PSM. The method tends to be consistent with an underlying framework for causal inference described by Holland (1986) as "Rubin's Causal Model." Recently the use of PSM within the framework of Rubin's Causal Model has become more visible in the education research literature. Hong and Raudenbush (2005, 2006) used PSM to evaluate the effect of kindergarten retention policies on academic achievement. Morgan (2001) used PSM to analyze the effect of Catholic schools on learning. Ruhm and Waldfogel (2007) used ratio matching techniques to estimate the effect of prekindergarten on later performance. A complete literature review would be beyond the scope of this paper, but these articles offer some indication of the range of educational questions that have been addressed using PSM techniques. Details for the particular analysis used here are given in the next section; a more general survey of PSM techniques is given in Caliendo and Kopeinig (2008). While PSM methods still make the assumption of "selection on observables," they typically relax the parametric assumptions associated with regression-based techniques, and perhaps most importantly, they focus the researcher's attention on the comparability of treatment and control units. Some subjects receiving an experimental treatment are simply not comparable to subjects receiving a control and vice-versa. Under a PSM approach subjects that are not comparable are excluded from the analysis and not used to estimate a causal effect.

The PSM approach has prompted strong claims from its proponents: "With flexible matching routines increasingly available, will regression adjustment for observational studies soon be obsolete?" (Hansen, 2004, p. 617). In fact, Hansen raised this question after performing a new analysis using the data from the Powers and Rock (1999) study on coaching effectiveness. Powers and Rock had used PSM methodology alongside regression to estimate the effect of SAT coaching, and the two methods had yielded similar results. In contrast, Hansen's estimates were roughly 5-8 points higher in math and 6 points lower in verbal, and he concluded that they were more defensible than those found by Powers and Rock[1]. In this paper, we revisit this conclusion in the context of comparing the LR and PSM approaches to estimating the effect of coaching on SAT performance.

## ELS Variables

Our analysis is based upon data from the ELS:02 survey conducted by the National Center for Educational Statistics. The survey followed a longitudinal cohort of high school sophomores in 2002 through their senior years (2004) and beyond (2006). It is designed to be representative of this national cohort of American high school students. The ELS:02 data contains basic demographic information about students, as well as more specific information about their academic achievement, attitudes, and opinions on a variety of subjects related to their schooling experiences. Information in the dataset on students' grades, test scores, and course-taking are based on official high school transcripts and test reports from the test makers rather than being self-reported.

The "treatment" of interest in this study is defined by the responses to a set of questions that asked students whether and how they prepared for the SAT. Students were able to indicate if they had prepared through the use of school courses, commercial courses, tutoring, or a variety of preparatory materials. In what follows, a coached student is defined as one that reported participating in a commercial preparatory course. Of the 16,197 students in the ELS:02 data, we restricted the sample to those students who had a 10th grade transcript available, responded to both the 2002 (grade 10) and 2004 (grade 12) surveys, and took the PSAT and SAT. We refer to these students as the "POP1" sample (N = 1,644). In contrast, we run separate analyses for a "POP2" sample (N = 2,549) that represents those students who took the SAT but did not take the PSAT[2]. There are some distinct differences between the types of students who take the PSAT and those who do not. Students that take the PSAT are much more likely to be college-bound and motivated to perform well on the SAT. Splitting the sample into the POP1 and POP2 groupings allows for explicit comparison to the results from Briggs (2001), where the same groups of students were defined from an earlier survey of a longitudinal student cohort from 1988 to 1992 (NELS:88). However, from the perspective of drawing unbiased causal inferences, the POP1 sample is clearly preferable to the POP2 sample because PSAT scores (available for the former but not the latter) are well correlated with both coaching status and subsequent SAT performance.

Descriptive statistics for the variables used in this analysis are given in Table 1 (an index of the variables used in this study is given in the Appendix). What sorts of variables that may be correlated with SAT performance serve to distinguish students that do and do not participate in commercial coaching? As elaborated in previous work (Briggs, 2002) these sorts of variables fall into roughly three groups: demographic characteristics of students, variables that proxy for academic achievement and aptitude, and motivational variables. In Table 1 we can see to what extent coached and uncoached students differ with respect to these variables. In many cases the differences are significant. For example, relative to uncoached students, coached students in POP1 score 2.3 points better on the PSATM and 1.6 points better on the PSATV (or 23 and 16 points when expressed on the SAT scale shown in Table 1). For students in the POP2 sample such comparisons with respect to PSAT obviously cannot be made, however, students participating in the ELS base year survey (in grade 10) were administered standardized tests in both math (BYMATH) and reading (BYREAD). These tests have strong positive correlations with the PSAT, so for students in the POP2 sample they serve as a substitute. However, it seems clear that they are an imperfect substitute. For the POP1 sample, in contrast to the mean differences observed on PSAT scores, there is no significant difference in mean BYMATH and BYREAD scores between coached and uncoached students. Hence it is likely that the BYMATH and BYREAD variables used for the POP2 sample do not fully capture the differences in prior ability to perform well on high stakes tests that is captured by the PSAT variable.

Major differences also exist between these two groups in terms of socio-economic status (SES), GPA (especially in POP2), group percentage of Asian students, percentages in private and rural schools, percentages of remedial course takers, percentages of ESL students, and percentages taking college preparatory curriculum and doing more than 10 hours per week of homework. In contrast, coached and uncoached students have fairly similar numbers of math credits and attend urban schools in similar percentages (especially in POP1). Student motivation is classic example of a plausible confounding variable in the context of coaching studies. Since the SAT is viewed by students as a crucial piece of their college application, they may have far greater motivation to perform well on this test than previous

**Table 1**. Variable Means for POP1 and POP2 by Coaching Status

| | POP1[a] | | | POP2[b] | | |
|---|---|---|---|---|---|---|
| Variable & Brief Description | Coached | Uncoached | p value | Coached | Uncoached | p value |
| PSATM*10[c] | 542 | 519 | 0 | NA | NA | NA |
| PSATV*10[c] | 524 | 508 | 0 | NA | NA | NA |
| BYMATH-ELS Math Test | 57.6 | 57.2 | 0.19 | 56.2 | 55.8 | 0.19 |
| BYREAD-ELS Reading Test | 57.2 | 56.9 | 0.25 | 55.2 | 55.1 | 0.46 |
| AGE/12[c] | 17.8 | 17.9 | 0.26 | 17.9 | 17.8 | 0.10 |
| SES Index | 0.68 | 0.38 | 0 | 0.56 | 0.3 | 0 |
| GPA | 3.09 | 3.05 | 0.14 | 3.08 | 2.96 | 0 |
| MCRD-# of math credits | 3.84 | 3.81 | 0.28 | 3.77 | 3.76 | 0.44 |
| Below variables are categorical, means expressed as percents. | | | | | | |
| FEMALE | 53 | 56 | 0.15 | 57 | 50 | 0 |
| ASIAN | 24 | 12 | 0 | 28 | 16 | 0 |
| BLACK | 12 | 8 | 0.01 | 15 | 14 | 0.36 |
| NATIVE | 3 | 3 | 0.48 | 4 | 5 | 0.18 |
| HISPANIC | 6 | 9 | 0.06 | 12 | 11 | 0.22 |
| PRIVATE | 56 | 43 | 0 | 36 | 25 | 0 |
| RURAL | 5 | 12 | 0 | 10 | 18 | 0 |
| URBAN | 39 | 39 | 0.50 | 42 | 33 | 0 |
| AP-Taken an AP course | 58 | 52 | 0.01 | 64 | 48 | 0 |
| REM_ENG-Remedial English | 7 | 6 | 0.33 | 6 | 7 | 0.21 |
| REM_MATH-Remedial Math | 8 | 7 | 0.18 | 6 | 8 | 0.17 |
| COLL_PREP-HS curriculum | 78 | 75 | 0.12 | 76 | 72 | 0.05 |
| HW->10 hours/wk homework | 40 | 25 | 0 | 34 | 22 | 0 |
| ESL | 23 | 13 | 0 | 23 | 16 | 0 |
| EDU_AFTER_HS[d] | 96 | 88 | 0 | 92 | 86 | 0 |
| COLLEGE_INFO[d] | 6 | 11 | 0 | 6 | 10 | 0 |
| PRNTS_DISC_PREP[d] | 34 | 19 | 0 | 37 | 22 | 0 |
| PRNTS_DISC_SCH[d] | 39 | 33 | 0.02 | 45 | 37 | 0 |
| NERVES.M[d] | 1 | 2 | 0.05 | NA | NA | NA |
| NERVES.V[d] | 2 | 2 | 0.34 | NA | NA | NA |
| UNDERPERFORM.M[d] | 14 | 14 | 0.44 | 13 | 16 | 0.03 |
| UNDERPERFORM.V[d] | 13 | 15 | 0.22 | 14 | 16 | 0.08 |
| N | 357 | 1195 | | 448 | 1941 | |

a. All students who responded to 2002 and 2004 surveys, had 10th grade transcripts, and took the PSAT and SAT.

b. All students who responded to 2002 and 2004 surveys, had 10th grade transcripts, and took the SAT (but not the PSAT).

c. PSATM, PSATV, and AGE are all transformed for the table, but the untransformed variables were used in the analysis.

d. See Appendix.

standardized tests (such as the PSAT or the ELS base year tests) or in their classes (as expressed by their GPA). Variables such as whether or not a student plans to continue his/her education after high school (EDU_AFTER_HS) and whether or not a student has sought out information about college (COLLEGE_INFO) capture some aspects of student motivation. We also created two new variables as proxies for motivation: UNDERPERFORM and NERVES. Students with a low GPA (less than 3.0) but a math or verbal score greater than the mean (for POP1 or POP2) on the ELS base year test were given a 1 on the variable UNDERPERFORM. We reasoned that such students have greater academic ability than their GPA suggests, and may sense a greater need to perform well on the SAT as the deadlines for college admissions approach. Should these students elect to get coached in preparation for the SAT, part of any score increase may be due to their new-found motivation rather than coaching. Those students who did substantially worse on the PSAT than we would have predicted given their performance on the ELS base year tests get a value of "1" on the variable NERVES[3]. Such students, doing worse on the PSAT than they may have expected, may be quite likely to sign up for coaching. We may falsely attribute a later score gain on the SAT for these students to coaching when we are in fact merely seeing a regression to the mean. The second variable relied upon a student's PSAT score, so we created it only for those students in POP1. As can be seen in Table 1, we found significant differences between coached and uncoached students on some of these variables as a function of SAT test subject and POP1 or POP2 membership.

One complication in using the ELS data is that there is a substantial amount of missing data. To simply exclude the missing cases would not only eliminate a large percentage of our data, but would also necessitate either the "missing at random" or "missing completely at random" assumptions (Rubin, 1976) which may be difficult to support. In examining the variables we found that certain dichotomous variables contain the bulk of the missing data[4]. Rather than simply throwing these cases out, we included missingness as an additional level of these variables[5]. This strategy, also followed by Hansen (2004), allowed us to include most of the POP1 and POP2 samples in our analysis. Only students with missing data on our continuous variables were removed from the subsequent analysis. This decreased the POP1 sample from 1,644 students to 1,552 and the POP2 sample from 2,549 to 2,389. While this is not the only approach available for dealing with missing data (imputation techniques would also be an option), this did allow us to retain most of our sample without a substantial increase in the complexity of our analysis. One final complication was that each student who failed to respond to the question regarding being Black also failed to respond to the question about being of Native American origin. Since this led to linearity between these variables, they were aggregated into a variable called RACE.

## Method

Because the approach typically used in a regression analysis is well understood, we will focus here on the methods used in our PSM analysis. Similar to Caliendo & Kopeinig (2008), we break a PSM analysis into five separate steps:

1. Estimating the Propensity Score
2. Implementing a Matching Algorithm
3. Assessing the Balance after Matching
4. Computing an Effect Estimate
5. Sensitivity Analysis

We shall discuss these five steps as we implement them in our analysis that follows.

The propensity score is at the core of the PSM methodology. It is the estimated probability of the unit of analysis receiving the treatment given the observed covariates, typically computed using logistic regression. Unbiased estimation of causal effects relies upon selection into treatment being a function of only those covariates used in the estimation of propensity scores. What to include in the selection function, the function which predicts treatment status, and how to choose its functional form are aspects of the methodology about which there is still some confusion in the literature. However, one agreed upon aspect in the PSM literature is that the success of the selection function should be the "balance" it generates in the distribution of covariates among treatment and control groups that have been matched according to their propensity scores.

Researchers have used the estimated propensity scores to compare treatment and control groups in several ways—inverse propensity weighting and kernel matching being two alternatives (see Frank et al., 2008 for an example of the first and Callahan et al., 2009 for an example of the second)—but in the present students we focus on two matching approaches that appear commonly in education research applications: subclassification and optimal pair matching. In the subclassification approach, units in the common support (the area of overlap on the estimated propensity score between the coached and uncoached groups) are split into subclasses based upon the quantiles of the distribution of the estimated propensity scores for the coached students. In the process we eliminate those students from the analysis who lack directly comparable counterparts in terms of their propensity scores. As noted earier, this constitutes a key distinction of PSM relative to LR. (As it turns out, in this study there is good overlap between coached and uncoached students on the estimated propensity scores, so we do not lose substantial portions of the POP1 or POP2 groups.) The optimal pair match is a one-to-one matching algorithm, meaning that each coached student is matched to a single uncoached student. A consequence of pair matching is that a larger number of uncoached students will be excluded from the analysis relative to subclassification. Optimal matching is done such that we obtain the lowest possible mean difference across all of the matches, hence the use of the term “optimal”. We used the R statistical computing package (R Development Core Team, 2008) as well as specialized matching software (Ho et al., 2004) to perform the propensity score estimation and matching.

Propensity scores are a means to an end. They are used to match treatment and control units such that after the units have been matched, their covariate distributions will be equivalent. For example, prior to matching, coached students may have higher mean PSAT scores than uncoached students. After matching, the PSAT means (and SDs) for coached and uncoached students should be about the same. Balance is a necessary condition for unbiased estimation via PSM. We apply two approaches to evaluate balance: the reduction in standardized mean differences after matching and an omnibus test for balance (Hansen & Bowers, 2008). Standardized differences are differences in coached versus uncoached covariate means relative to a weighted combination of the standard deviations across matched units. The omnibus test for balance is designed to answer the following question. Is the degree of difference between the treatment and controls groups consistent with that which would be expected between two groups randomly created from a single sample, as in an experiment? The test statistic we use, computed via the software of Bowers et al. (2008), simultaneously tests a Fisher Randomization Hypothesis for all covariates.

If the matched data looks as though it could have come from randomization, the simplest approach to computing an effect estimate is to compare differences in group means. If selection into treatment is solely a function of the observable data, then this will be an unbiased estimate of the treatment effect (Rosenbaum & Rubin, 1983). This is the basic idea behind the approach we use to estimate the coaching effect under the subclassification PSM approach. In contrast, for the optimally pair matched data, we depart from this simple approach because clear imbalances remain on important covariates (the PSAT for the POP1 sample and the ELS base year tests for the POP2 sample), even after matching. Hence we estimate the coaching effect after adjusting for remaining PSAT score differences using a regression model. We estimate standard errors using the Huber-White correction to adjust for the clustering of students at the school level.

As with the LR approach, the validity of the PSM-based estimates for the coaching effect depend most fundamentally upon the availability of all the relevant variables that predict whether or not a student is likely to be coached. Rosenbaum (2002) outlines a procedure which allows us to check the robustness of our results to certain deviations from this assumption. In particular, we assume that there exists a hidden variable with a known relationship to treatment status. Using Keele’s (2008) software, we are able to examine the degree to which our effect estimates may change if such a confounder were to exist.

## Results

### *Linear Regression*

In presenting the results, we first discuss the results from the regression analysis. Table 2 shows the coefficient estimates for each of four regressions: both sections of the test for both POP1 and POP2 samples. The estimated effect of coaching on the math section is 11 and 22 points for the POP1 and POP2 samples respectively, both statistically significant at the .01 level. For the verbal section, the effects were 6 points for POP1 and 8 points for POP2, only the second of which was statistically significant at the .05 level.

**Table 2**. Coefficient Estimates with Huber-White Standard Errors for Linear Regression Analysis

| | SATM | | | | SATV | | | |
|---|---|---|---|---|---|---|---|---|
| | POP1 | | POP2 | | POP1 | | POP2 | |
| Variable | Coeff | SE | Coeff | SE | Coeff | SE | Coeff | SE |
| (Intercept) | 41.7 | 51.7 | 63.0 | 55.5 | 48.5 | 45.4 | 29.8 | 61.9 |
| **COACH** | **11.3** | **3.1** | **21.5** | **3.0** | **5.8** | **3.6** | **7.9** | **3.8** |
| PSATM | 5.0 | 0.3 | NA | NA | 0.5 | 0.3 | NA | NA |
| PSATV | 1.0 | 0.3 | NA | NA | 6.4 | 0.3 | NA | NA |
| BYMATH | 3.3 | 0.4 | 7.6 | 0.3 | 0.8 | 0.3 | 2.3 | 0.3 |
| BYREAD | 0.0 | 0.3 | 0.6 | 0.2 | 1.8 | 0.3 | 5.3 | 0.3 |
| AGE | -0.2 | 0.3 | -0.4 | 0.2 | -0.2 | 0.2 | -0.1 | 0.3 |
| SES | 1.4 | 2.3 | 10.2 | 2.1 | 7.1 | 2.1 | 18.1 | 2.4 |
| FEMALE | -13.0 | 2.7 | -18.0 | 3.0 | 0.2 | 2.9 | -6.5 | 2.9 |
| ASIAN1 | 0.9 | 5.6 | 10.5 | 4.7 | -2.2 | 5.4 | -13.4 | 4.5 |
| ASIANNA | -41.6 | 31.6 | -71.0 | 70.9 | 19.4 | 27.9 | -57.2 | 49.8 |
| RACENATIVE | 6.8 | 6.7 | -0.2 | 5.9 | 1.8 | 10.1 | -0.3 | 6.7 |
| RACEBLACK | -11.4 | 4.0 | -16.1 | 4.9 | 8.2 | 5.4 | -4.5 | 5.2 |
| RACEBOTH | -28.9 | 11.6 | -9.5 | 13.5 | 0.4 | 12.3 | -4.7 | 10.7 |
| RACEBOTHNA | 31.3 | 31.2 | 72.9 | 70.4 | -21.8 | 26.7 | 52.2 | 49.4 |
| HISPANIC1 | -9.8 | 4.3 | -17.0 | 4.9 | 0.5 | 7.4 | -2.9 | 6.5 |
| HISPANICNA | 2.3 | 18.2 | 12.4 | 13.2 | 21.2 | 16.4 | -17.2 | 9.4 |
| PRIVATE | 1.6 | 3.2 | 8.4 | 3.7 | 2.7 | 3.5 | 16.5 | 4.3 |
| RURAL | -4.7 | 4.9 | -2.1 | 3.9 | -4.3 | 4.2 | 2.5 | 4.6 |
| URBAN | -6.5 | 3.4 | -1.0 | 3.5 | -10.4 | 3.7 | -2.4 | 3.9 |
| AP | 12.5 | 3.2 | 27.0 | 3.2 | 12.6 | 3.4 | 33.7 | 3.5 |
| REM_ENG1 | 12.1 | 9.1 | 17.3 | 8.0 | 2.7 | 9.3 | -3.5 | 8.7 |
| REM_ENGNA | -5.7 | 14.4 | 2.4 | 12.2 | 7.7 | 13.3 | -10.3 | 13.6 |
| REM_MATH1 | -12.4 | 9.0 | -23.6 | 7.6 | -0.9 | 7.2 | 7.6 | 9.1 |
| REM_MATHNA | 11.9 | 15.8 | 5.1 | 12.3 | -1.3 | 12.8 | 16.4 | 12.7 |
| COLL_PREP | -1.0 | 3.4 | -0.6 | 2.8 | -1.4 | 3.2 | 6.7 | 3.0 |
| MATH_CRD | 0.5 | 1.8 | 2.8 | 1.4 | 0.4 | 1.8 | -2.2 | 1.5 |
| HW | 0.6 | 3.3 | 8.5 | 2.8 | 2.2 | 3.0 | 3.7 | 3.2 |
| ESL | 17.8 | 4.3 | 8.0 | 4.0 | 2.3 | 4.5 | -8.6 | 4.9 |
| EDU_AFTER_HS | 1.5 | 4.2 | 0.5 | 3.5 | -0.4 | 4.4 | -6.5 | 4.1 |
| COLLEGE_INFO1 | -4.4 | 5.8 | 1.8 | 4.5 | 3.1 | 4.4 | 4.0 | 4.9 |
| COLLEGE_INFONA | 4.5 | 6.6 | 16.5 | 7.3 | -0.1 | 8.1 | 12.8 | 7.1 |
| PRNTS_DISC_PREP1 | 1.9 | 3.8 | 5.8 | 3.4 | 5.3 | 3.6 | -0.8 | 3.8 |
| PRNTS_DISC_PREPNA | -9.7 | 9.2 | -21.0 | 10.0 | 13.0 | 9.6 | -6.0 | 11.8 |
| PRNTS_DISC_SCH1 | -7.6 | 3.0 | -8.1 | 2.8 | -5.9 | 3.1 | -3.4 | 3.1 |
| PRNTS_DISC_SCHNA | 16.3 | 10.1 | 18.4 | 10.8 | -19.2 | 9.8 | -8.2 | 12.4 |
| GPA | 14.4 | 3.8 | 23.3 | 3.4 | 7.9 | 3.5 | 24.6 | 3.4 |
| UNDERPERFORM.M | 3.0 | 5.7 | 9.5 | 4.6 | 1.9 | 4.7 | 1.3 | 5.4 |
| UNDERPERFORM.V | -2.1 | 4.6 | -10.5 | 4.2 | -3.2 | 4.3 | -10.0 | 4.9 |
| NERVES.M | 25.9 | 13.3 | NA | NA | 7.9 | 11.8 | NA | NA |
| NERVES.V | 12.5 | 10.2 | NA | NA | 25.0 | 11.4 | NA | NA |
| N | 1552 | | 2389 | | 1552 | | 2389 | |
| $R^2$ | 0.77 | | 0.73 | | 0.77 | | 0.67 | |

Since the SAT scale has been internalized by many who have been educated in the United States, the magnitude of our causal effect has inherent meaning to those who recall taking the test as a high school student. However, there are two additional ways of contextualizing this effect. We can first compare our effect estimates to the unadjusted difference in SAT means for coached and uncoached students to assess the impact of adjusting for confounding variables. On the math section, there were unadjusted differences of 31 and 36 points in POP1 and POP2 respectively. After adjustment using LR, those differences fall to 11 and 22 points respectively. On the verbal section, initial differences of 23 and 22 points were reduced to 6 and 8 points for POP1 and POP2. Aggregating the effects across both parts of the test, we find that of the initial 54 point difference between coached and uncoached students in POP1, only 17 points can be attributed to coaching. For POP2, we can only attribute 30 of the 58 point initial difference to coaching. We can also express the coaching effects as effect sizes. Using the standard deviation of the uncoached students, the effect sizes of coaching on the math part were 0.11 and 0.20 for POP1 and POP2 samples. On the verbal section, the effect sizes were .06 and .07 for POP1 and POP2 samples respectively.

*Propensity Score Matching*

As a first step in our PSM analyses, we ran a logistic regressionsusing the full set of covariates shown in Table 2. Logistic regression coefficients for the subclassification and pair matching approaches are shown in Table 3. We distinguish between the logistic regression coefficients for subclassification and optimal pair matching because in each case, different samples of students were included or excluded in the matching procedure depending upon where they fell within the area of common support (or overlap) on the estimated propensity score, and whether (in the case of the optimal matching approach) an uncoached student could be matched to a specific coached student. In the opimal pair matching apprpoach, we only excluded those coached students from matching who were more likely to be coached than any uncoached student. In the subclassification appoach, we also excluded students who were more like to be uncoached than any coached student. Once students were excluded, the logistic regression was run again using the restricted sample. Matches in the optimal pair match were made using logit units rather than the actual propensity scores. The difference between similar propensity scores on the high or low end of the propensity score scale (near 0 or 1) will increase when logit units are used instead. Since the goal is to have the smallest global difference, using the logit forces better matches at the high and low end of the propensity score scale. Following the lead of Rosenbaum & Rubin (1983), we use quintiles of the propensity score distribution to form our subclasses[6].

The focus in PSM on comparing only those units who are directly comparable on the estimated propensity score is important, so we draw some attention to the students who were "unmatchable" and hence, unlike in LR, were excluded as a basis for the subsequent estimation of a causal effect. Under subclassification, both coached and uncoached students who fall outside the region of common support were excluded. In POP1, this led to the exclusion of 54 uncoached students and 4 coached students. The coverage of the common support was even better in POP2, only necessitating the exclusion of 4 students, of which only one was coached. In the optimal pair match we only removed 4 coached students from POP1 and 1 coached student from POP2. After matching, 842 of the uncoached students from POP1 were not matched to coached counterparts, and 1,494 uncoached students from POP2 were not matched, an illustration of the restrictive nature of the pair matching approach.

As a first empirical evaluation of the extent to which balance has been achieved, Figure 1 shows plotted density curves of propensity score distributions for both POP samples under each matching algorithm. As we would expect, the top row of figures indicates that there are higher percentages of coached students who were likely to be coached (i.e., had higher propensity scores) on the basis of our predictor variables. The relative paucity of uncoached students in the higher range of the estimated propensity score indicates that the causal estimates for these subclasses, especially the highest subclass (notice that the vertical lines represent the subclass divisions), will rest upon comparisons of many coached students to few uncoached students.

Balance requires more than similar distributions for the estimated propensity scores among coached and uncoached groups; it is also necessary for the distributions of each relevant covariate to be similar. Figures 2 and 3 illustrate graphically the improvement in standardized mean differences after matching. In all four cases, there are generally big improvements in balance after matching. Balance appears to be strongest under the subclassification approach, where differences in covariates between the groups are

**Table 3**. Logistic Regression Coefficients for Subclassification

| | POP1 | | POP2 | |
|---|---|---|---|---|
| Variable | Coeff | SE | Coeff | SE |
| (Intercept) | -2.84 | 3.01 | -5.25 | 2.38 |
| PSATM | 0.02 | 0.01 | NA | NA |
| PSATV | 0.00 | 0.01 | NA | NA |
| BYMATH | -0.03 | 0.02 | -0.01 | 0.01 |
| BYREAD | 0.00 | 0.01 | -0.02 | 0.01 |
| AGE | 0.00 | 0.01 | 0.02 | 0.01 |
| SES | 0.59 | 0.11 | 0.50 | 0.09 |
| FEMALE | -0.07 | 0.15 | 0.20 | 0.12 |
| ASIAN1 | 0.78 | 0.24 | 0.70 | 0.17 |
| ASIANNA | 0.11 | 1.37 | -0.20 | 0.32 |
| RACENATIVE | -0.56 | 0.51 | -0.11 | 0.30 |
| RACEBLACK | 0.60 | 0.24 | 0.45 | 0.18 |
| RACEBOTH | 2.50 | 0.81 | -0.17 | 0.64 |
| RACEBOTHNA | 0.02 | 1.30 | NA | NA |
| HISPANIC1 | -0.32 | 0.33 | 0.43 | 0.22 |
| HISPANICNA | -0.13 | 0.83 | 0.19 | 0.53 |
| PRIVATE | 0.56 | 0.16 | 0.40 | 0.14 |
| RURAL | -0.65 | 0.30 | -0.18 | 0.19 |
| URBAN | -0.33 | 0.15 | 0.05 | 0.13 |
| AP | 0.01 | 0.17 | 0.54 | 0.14 |
| REM_ENG1 | -0.16 | 0.46 | -0.28 | 0.39 |
| REM_ENGNA | -0.23 | 0.90 | -0.41 | 0.66 |
| REM_MATH1 | 0.34 | 0.42 | -0.03 | 0.37 |
| REM_MATHNA | 0.63 | 0.89 | -0.20 | 0.65 |
| COLL_PREP | -0.04 | 0.17 | 0.05 | 0.13 |
| MATH_CRD | 0.04 | 0.09 | -0.04 | 0.06 |
| HW | 0.46 | 0.15 | 0.25 | 0.13 |
| ESL | 0.48 | 0.23 | 0.16 | 0.18 |
| EDU_AFTER_HS | 1.10 | 0.33 | 0.29 | 0.20 |
| COLLEGE_INFO1 | -0.28 | 0.27 | -0.42 | 0.24 |
| COLLEGE_INFONA | -0.48 | 0.40 | 0.53 | 0.30 |
| PRNTS_DISC_PREP1 | 0.62 | 0.17 | 0.49 | 0.13 |
| PRNTS_DISC_PREPNA | -0.64 | 0.62 | -0.11 | 0.45 |
| PRNTS_DISC_SCH1 | 0.02 | 0.16 | -0.05 | 0.13 |
| PRNTS_DISC_SCHNA | 1.29 | 0.62 | -0.53 | 0.48 |
| GPA | -0.10 | 0.16 | 0.10 | 0.14 |
| UNDERPERFORM.M | 0.12 | 0.25 | -0.05 | 0.22 |
| UNDERPERFORM.V | -0.22 | 0.25 | 0.19 | 0.22 |
| NERVES.M | -0.12 | 0.68 | NA | NA |
| NERVES.V | 0.33 | 0.52 | NA | NA |
| N | 1494 | | 2385 | |

consistently less than one tenth of an SD. Next we applied the omnibus test described by Hansen & Bowers to evaluate whether the degree of balance that we observe is close enought to that which would be expected from the creation of two samples from a single population via random assignment. The result in each case was a test statistic with a very low p-value, which indicates that the balance within subclasses was similar to what one would expect had coached and uncoached students been randomly assigned (this was true for matches with both POP1 and POP2 samples). Interestingly, however, while the omnibus test suggests that both approaches result adequate balance, under the optimal matching

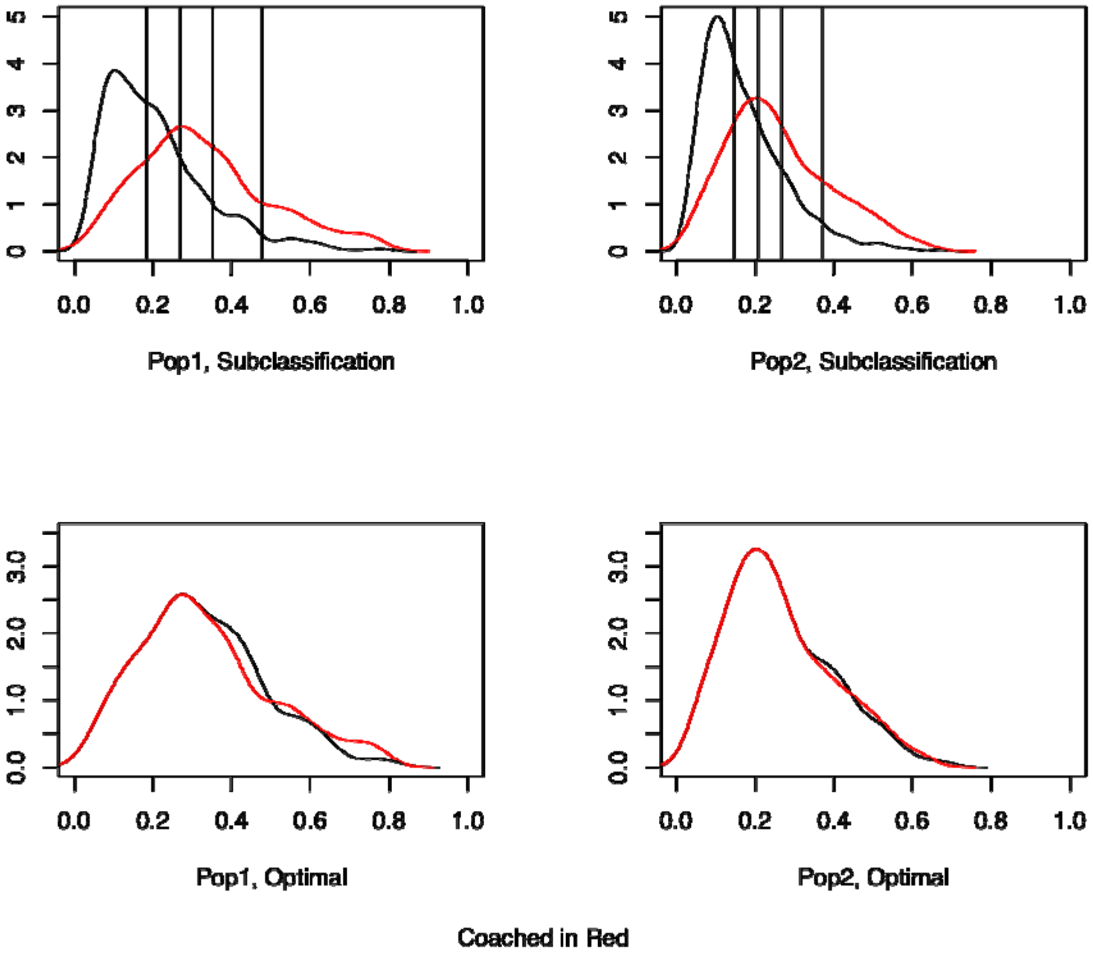

**Figure 1**. Density Curves for the 4 Matched Data Sets

approach there are clearly a number of variables for which important differences between the two groups remain. Most noteably for the POP1 samples, coached students continue to have higher PSAT scores than their uncoached counterparts. For the POP2 samples, scores on the ELS base year tests in math and reading (BYMATH and BYREAD) actually become somewhat less balanced between coached and uncoached students after matching. These imbalances point to potential sources of bias that would still need to be taken into account when estimating coaching effects, and they also underscore the importance of not relying solely on tests of significance to support a conclusion that all plausible confounders are suitably balanced.

To compute effect estimates from our subclassified data we regressed the math or verbal portion of the SAT on a dummy variable indicating coaching status as well as a dummy variable indicating the propensity score subclass for the student. For math, the estimated effects were 12 and 22 points in POP1 and POP2 respectively. For verbal, they were 6 and 9 points. Only the POP2 effect for math was significant at the .05 level. In the optimally pair matched data, we began by regressing the sections of the test on a dummy variable for coaching status. This led to effects of 24 and 18 points in math for POP1 and POP2 and 14 and 0 points in verbal. However, due to the remaining imbalances on the PSAT tests shown in Figure 3, we ran subsequent regressions in which the variables PSATM and PSATV were included as controls for the POP1 sample, and the variables BYMATH and BYREAD were included as controls for the POP2 sample. In these regressions, the estimated math and verbal effects for POP1 fell from 24 and 14 points to 15 and 9 points. In contrast, because uncoached students had higher mean BYMATH and BYREAD scores than coached students after matching, the math and verbal effects increased from 18 and 0 points to 24 and 6 points after regression adjustment. These results are summarized in Table 4. Most of the coaching effect estimates for math are statistically significant at the .05 level in both POP1 and POP2 samples; in contrast, none of the verbal estimates meet this conventional threshold[7].

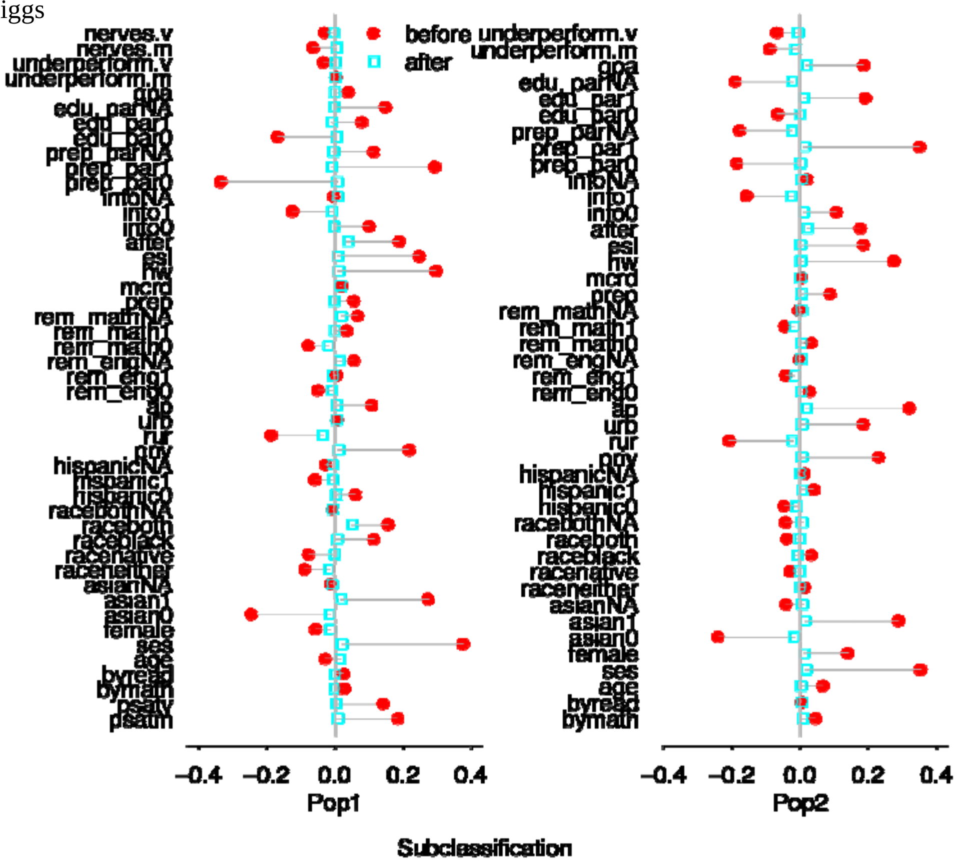

**Figure 2**. Improvement in Standardized Differences after Subclassification

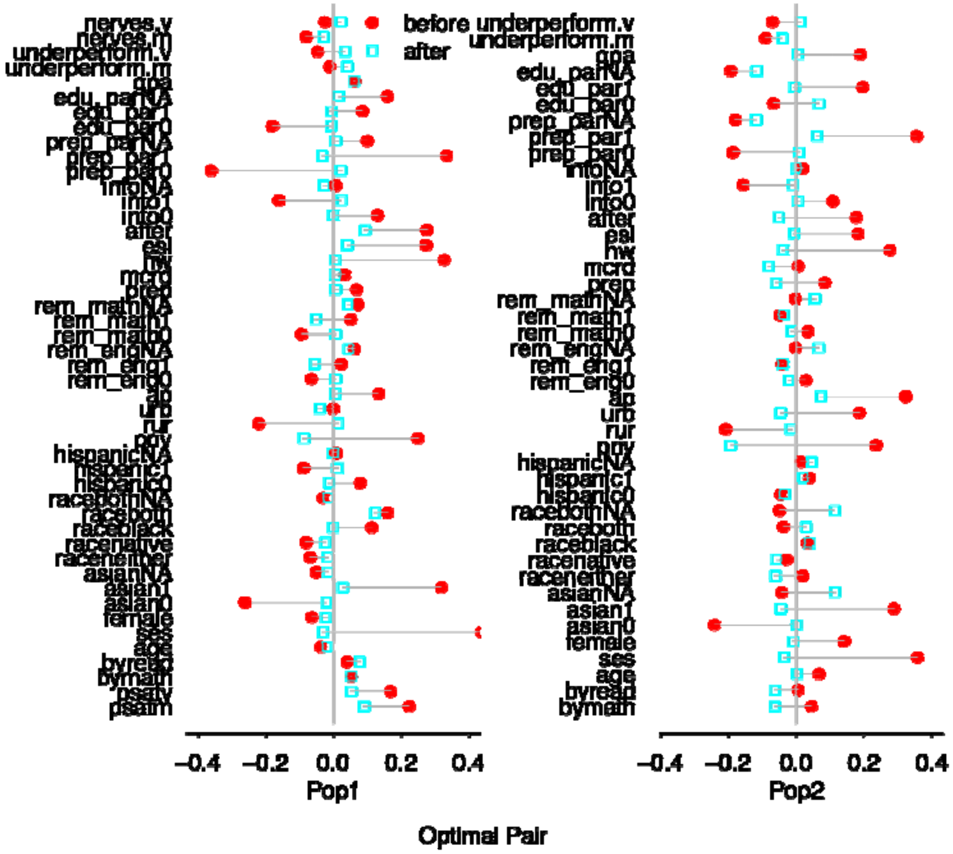

**Figure 3**. Improvement in Standardized Differences after Optimal Pair Match

**Table 4**. Effect Estimates from Matching with Huber-White Standard Errors in Parentheses

| | Math | | Verbal | |
|---|---|---|---|---|
| Matching Method | POP1 | POP2 | POP1 | POP2 |
| Subclass | 12.4 | 21.9 | 6.3 | 8.6 |
| | (8.4) | (6.5) | (8.5) | (6.6) |
| N | 1494 | 2385 | 1494 | 2385 |
| Optimal | 23.5 | 18.0 | 14.2 | 0.3 |
| | (10.3) | (9.1) | (10.5) | (8.9) |
| Optimal | 15.1 | 24.2 | 8.5 | 6.4 |
| with Smoothing | (4.7) | (4.5) | (4.4) | (5.7) |
| N | 706 | 894 | 706 | 894 |

One benefit of the subclassification approach is that it allows us to further probe the LR constraint that the effect of coaching can be best summarized with a single number. Table 5 shows the estimated coaching effect in each subclass as defined by ranges of the propensity score distribution. It also shows the ceiling for the estimated propensity scores in that subclass as well as the mean SES for the students in each subclass. Note that the effects vary quite dramatically with the highest effect estimates found in the higher subclasses. The higher subclasses contain those students most likely to be coached. Since the propensity score correlates strongly and positively with SES, one plausible explanation is that more affluent students are buying coaching that is both more expensive and more effective.

*Sensitivity Analysis*

The coaching estimates from our original regression analyses were predicated upon the constraint that there is a single mean causal effect that applies to all students. As the results from the subclassification analysis above indicate, this may be unreasonably restrictive. Through the inclusion of interaction terms in the regression model, we can readily evaluate the possibility of differential causal effects of coaching for selected subsamples of students. In what follows we report these results (summarized in Tables 6 & 7) only for student subgroups in the POP1 sample. We focus on the POP1 sample because we are more confident in these coaching estimates since they control for preexisting differences in PSAT performance.

One characteristic of coaching services that our treatment variable cannot operationalize is that these programs vary widely in cost. Extremely expensive small group work with college professors may be classified as coaching right alongside much less expensive coaching offered in community centers. The effects of these two types of coaching may be quite different. Since we do not have information on how much money students paid for coaching, we instead use a dummy variable indicating whether a student was in the top quartile of the distribution for the SES variable supplied in the ELS:02 data. Those students that were in the top quartile of the SES index are potentially paying for much more expensive (and possibly higher quality) coaching. When coached students in the top SES quartile are compared to uncoached students in the top SES quartile, we find a mean difference of 15 points on SAT math scores. In contrast, amongst students in lower SES quartiles, coaching has only a 5 point effect. Hence math coaching appears to be 10 points more effective for high SES students than it is for lower SES students. On the verbal section the coaching by SES interaction is weaker: the effect is 9 points for high SES students but just 2 points for lower SES students.

There appear to be no gender related differences in the effectiveness of coaching. On the other hand, there are some important differences as a function of race/ethnicity. Coaching is differentially effective for Asian students, for whom the coaching effect is 5 and 14 points higher on the math and verbal sections respectively. One explanation for this is that 70% of the coached Asian students are also ESL students. For Black students, the effect of coaching on the math portion of the test was 15 points higher than it was for non-Black students. Conversely, the effect on the verbal section of the SAT for Black students was 18 points lower than the coaching effect for non-Black students. The latter result is driven by our finding of a negative effect for Black students on the verbal section of the SAT. Finally, students with AP course experience seem to benefit considerably from coaching. On each section of the SAT, we found a coaching effect that was 12 points higher than similar coached students who had never taken an AP course.

**Table 5**. Effect Estimates across Each Subclasses

| | POP1 | | | | POP2 | | | |
|---|---|---|---|---|---|---|---|---|
| Subclass | Class Ceiling | SES | Math | Verbal | Class Ceiling | SES | Math | Verbal |
| 1 | 0.18 | 0.17 | -18 | -13 | 0.15 | 0.03 | 3 | -20 |
| 2 | 0.27 | 0.55 | 9 | -15 | 0.21 | 0.39 | 23 | 33 |
| 3 | 0.35 | 0.75 | 37 | 25 | 0.27 | 0.57 | 27 | 14 |
| 4 | 0.48 | 0.81 | 36 | 41 | 0.37 | 0.75 | 58 | 31 |
| 5 | 0.85 | 1.07 | 72 | 65 | 0.71 | 1.02 | 72 | 50 |

Turning now to a sensitivity analysis of our PSM results, we consider the effects of some alternate specifications and assumptions. A potential mistake one may make in a PSM analysis is to misspecify the selection function. We established the selection function used in our analyses through careful consideration of what variables should theoretically influence coaching status and SAT results. Once this has been established, in our view there is no sound rationale for excluding such variables from the selection function on the basis of their statistical significance. Nonetheless, such exclusions are possible and even likely whenever the variables for a selection function are chosen by a stepwise algorithm[8]. To assess how sensitive our results are to the exclusion of certain variables from the selection function on the basis of their statistical significance, we estimated coaching effects after matching on the basis of a restricted selection function. In this approach only those covariates which were significant at the .05 level in the initial logistic regression were retained and then the logistic regression was estimated again to generate new propensity score predictions. After matching using the same subclassification and pair matching techniques described above, we computed new causal effect estimates (summarized in Table 8). Most notably, the estimated coached effect was 5 points higher for math (from 12.4 to 17.4) under subclassification (POP1 sample), and 6 points higher for verbal (from 6.4 to 12.1) under optimal matching (POP2 sample). In all other cases the coaching effects were about the same using either selection function. While a 5 to 6 point change to the effect may seem relatively small as a "worst case scenario", expressed as a percentage of the original effect (where the "original" effect is that deriving from the unrestricted selection function) they represent substantal increases of 40% and 89% respectively. Indeed, the differences that arise from this adjustment to the selection function are about as big as the largest differences found when shifting from an LR to a PSM approach to estimate coaching effects.

**Table 6.** Interaction Effects on SATM for Selected Variables (POP1 Sample)

| Variable | Effect Outside[1] | Effect Inside[2] | Difference |
|---|---|---|---|
| HI_SES | 4.7 | 15.1 | 10.4 |
| AP | 6.5 | 14.9 | 8.3 |
| FEMALE | 11.3 | 11.2 | -0.1 |
| ESL | 10.6 | 14.1 | 3.5 |
| ASIAN | 11.1 | 16.2 | 5.1 |
| BLACK | 11.4 | 26.4 | 15.0 |

**Table 7.** Interaction Effects on SATV for Selected Variables (POP1 Sample)

| Variable | Effect Outside[1] | Effect Inside[2] | Difference |
|---|---|---|---|
| HI_SES | 1.8 | 9.4 | 7.6 |
| AP | 3.7 | 7.5 | 3.8 |
| FEMALE | 5.7 | 6.0 | 0.3 |
| ESL | 4.8 | 10.4 | 5.7 |
| ASIAN | 2.6 | 17.1 | 14.5 |
| BLACK | 8.5 | -9.4 | -18.0 |

[1] These columns show the coaching effect considering only those students outside the group of interest.
[2] These columns show the coaching effect considering only those students inside the group of interest.

Another type of sensitivity analysis is to ask how our effect estimates would change in the presence of hidden bias

**Table 8**. Sensitivity of PSM Results to Selection Function

| Selection Function | SATM POP1 | SATM POP2 | SATV POP1 | SATV POP2 |
|---|---|---|---|---|
| | | Subclassification | | |
| Original | 12.4 | 21.9 | 6.3 | 8.6 |
| Significant | 17.4 | 23.9 | 7.1 | 10.2 |
| | | Optimal | | |
| Original | 15.1 | 24.2 | 8.5 | 6.4 |
| Significant | 14.0 | 24.5 | 7.1 | 12.1 |

(Rosenbaum, 2002). This method can be used to estimate the upper and lower bounds for how our effect estimate under the pair matching approach would change if we did not observe a variable which was predictive of treatment status. While the technical details of how such an analysis is conducted are outside the scope of this paper, we have found that our results are indeed sensitive to having fully observed all relevant confounders. In the presence of a moderate hidden bias[9], the effect of coaching in the POP1 sample may range anywhere between 5 to 45 points on the math section of the SAT and 0 to 25 points on the verbal section.

## Discussion

The substantive results from this study can be compared to the results of similar studies that have evaluated the effect of coaching. Earlier work suggests point estimates for an an overall coaching effect across both math and verbal sections of the SAT of roughly 25 points. In particular, we base our comparison on the work of Briggs (2001), Powers and Rock (1999), and Hansen (2004). The studies by Powers & Rock and Hansen used samples similar to our POP1 group, so comparisons should only be made directly to our POP1 findings. In contrast, the study by Briggs used data from NELS:88, which had the same structure and sampling design as ELS:02. So for this study coaching effect comparisons can be made for both POP1 and POP2 samples. The results from the individual studies are shown alongside our results in Table 9. In general, our estimates for the effect of coaching are similar to those of the earlier studies. Comparing effects based on regression-based estimates over time suggests that coaching has become slightly less effective for both sections of the SAT, at least on the math section of the exam. Comparing the different methodologies, Hansen's PSM analysis (which employed a "full matching" approach) produces estimates that are noticeably different for the math and verbal when compared to the others.

One interesting finding when comparing our results to those from the Briggs study using NELS:88 data is that the effect of coaching for students who have not taken the PSAT (POP2 sample) is about 10 points higher in math and 5 points higher in verbal. The difference in coaching effects in math for students in the POP1 and POP2 samples merits closer attention. On the one hand, this may indicate that coaching has a larger effect for those who have not previously taken the PSAT. On the other hand, this may be an artifact of uncontrolled confounding because we are missing information on differences in test-taking ability captured by PSAT scores. As a check on this, we re-ran the POP1 regressions after excluding the PSAT variables as controls. The resulting coaching effect in math increased from 11 to 16 points, much closer to the 22 point effect found for the POP2 sample. This leads us to believe that the higher effect estimates for the POP2 sample must be taken with a grain of salt.

In this study, the effects estimated on the basis of propensity score subclassification are more comparable to the regression results than they are with the optimal pair matching results. A key reason that coaching estimates deriving from the subclassication approach and LR approaches are so similar in this example is that the estimated propensity score distributions for coached and uncoached were not severely imbalanced at the outset. After matching through subclassification, relatively few students were excluded when estimating coaching effects. The linear regression and subclassification results use similar numbers of students, 1,552 and 1,494 for POP1 in the LR and subclassification respectively and 2,389 and 2,385 for POP2. The optimal match results are based upon far fewer students, only 706 and 894 for POP1 and POP2. This is one potential drawback to the pair matching approach[10].

Properly specifying a selection function can be a challenging part of a PSM analysis. In our view, the practice of removing variables from the selection function on the basis of their statistical significance is problematic. If theory dictates that a variable be included in the selection function, lack of statistical significance is an idiosyncratic reason for removal, one that encourages "fishing for significance." It is

**Table 9**. Coaching Effect Estimates from Various Studies

| Study | SATM | | SATV | |
|---|---|---|---|---|
| Powers & Rock (1999)-LR | 18 | | 6 | |
| Powers & Rock (1999)-PSM | 15 | | 6 | |
| Hansen (2004)-PSM | 23 | | 0 | |
| | POP1 | POP2 | POP1 | POP2 |
| Briggs (2001)-LR | 15 | 8[1] | 6 | 1[1] |
| Current Study | | | | |
| Regression | 11 | 22 | 6 | 8 |
| Subclassification | 12 | 22 | 6 | 9 |
| Optimal Pair (w/ smoothing) | 15 | 24 | 9 | 6 |

[1] These figures were not reported in the Briggs (2001) article but come from the same study.

also problematic because it can create a tautological case for balance. That is, balance is demonstrated only after the symptoms of imbalance—non-significant variables in the selection function—have been removed.

Going back to Hansen's question, will matching make regression obsolete as a statistical model for causal inference? We think not. Fundamentally, it does not seem any easier to model selection as opposed to outcome: both techniques depend on the quality of the available variables to capture sources of confounding. Linear regression is also less time-consuming to implement and makes checking for interactions much easier. However, there are a number of aspects of PSM which we find appealing. The emphasis that a PSM approach places on estimating causal effects on the basis of comparable units is important. When confronted with observational data in the context of a regression analysis, a fundamental lack of comparability between treatment and control units can be easily swept under the hood. In such a context Rubin has warned that "inferences for the causal effects of treatment on such a unit cannot be drawn without making relatively heroic modeling assumptions involving extrapolations. Usually, such a unit should be explicitly excluded from the analysis" (Rubin, 2001, p. 180). Furthermore, the sensitivity analysis approach suggested by Rosenbaum (2002) offers a powerful way of analyzing the robustness of one's results to hidden biases, and such an approach cannot be (or at least has not been) readily applied to linear regression. Conclusions based on the results of a PSM analysis can be judged relative to the robustness of the results in the presence of hidden biases. While more work is necessary to understand what does and does not constitute robustness using this method in educational research, such an approach is conceptually appealing.

Rubin (2006) has suggested that there is also an ethical advantage to the use of PSM to estimate causal effects in that all matching can done without reference to the outcomes. In principal this would seem to support the objectivity of causal inferences. In our view this is overselling the approach. Drawing causal inferences about student achievement in an observational setting fundamentally requires us to make an informed hypothesis about (a) why subjects choose to participate (i.e., self-select) in treatment and control groups, and (b) what characteristics of these subjects are associated with the outcome of interest. All statistical modeling that follows hinges upon the quality of this foundation. If the foundation has been well-established, then just as with PSM, the estimation of an aggregate causal effect using LR happens only once. This is precisely the approach that was taken in the present study. However, if the foundation has not been well-established—and we suspect this is the case whenever covariates are being chosen on the basis of statistical significance—then there is no magic that will pull the causal rabbit out of the observational hat.

## References

Abadie, A. & Imbens, GW. (2006). On the Failure of the Bootstrap for Matching Estimators. *National Bureau of Economic Research Technical Working Paper Series, 325.*

Bowers, J., Fredrickson, M., & Hansen, B. (2008). RItools: Randomization Inference Tools. R package version 0.1-3

Briggs, D. (2001). The effect of admissions test preparation: Evidence from NELS-88. *Chance, 14,* 10-18.

Briggs, D. (2002). SAT Coaching, Bias and Causal Inference. Unpublished doctoral dissertation, University of California, Berkeley.

Briggs, D. (2004). Evaluating SAT Coaching: Gains, Effects, and Self-Selection. In R. Zwick (Ed.), *Rethinking the SAT: The Future of Standardized Testing in University Admissions* (p. 217-234). New York: RoutledgeFarmer.

Caliendo, M. & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22* (1), 31-72.

Callahan, R., Wilkinson, L., Muller, C., & Frisco, M. (2009). ESL Placement and Schools: Effects on Immigrant Achievement. *Educational Policy, 23* (2), 355-384.

College Board. (2009). *SAT Reasoning Test*. Retrieved March 21, 2009 from http://www.collegeboard.com/student/testing/sat/about/SATI.html.

Frank, KA., Sykes, G., Anagnostopoulos, D., Cannata, M., Chard, L., Krause, A. (2008). Does NBPTS certification affect the number of colleagues a teacher helps with instructional matters? *Educational Evaluation And Policy Analysis, 30,* 3-30.

Gastwirth, J., Krieger, A., & Rosenbaum, P. (2000). Asymptotic seperability in sensitivity analysis. *Journal of the Royal Statistical Society: Series B, 62* (3), 545-555.

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

Hansen, B. & Bowers, J. (2008). Covariate Balance in Simple, Stratified and Clustered Comparative Studies. *Statistical Science, 23,* 219-236.

Ho, D., Imai, K., King, G., & Stuart, E. (2004). *Matchit: Matching as Nonparametric Preprocessing for Parametric Causal Inference*. Available from http://gking.harvard.edu/matchit

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15* (3), 199-236.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945-960.

Hong, G. & Raudenbush, S. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis, 27* (3), 205-224.

Hong, G., & Raudenbush, S. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101* (475), 901-910.

Keele, L. (2008). rbounds: Perform rosenbaum bounds sensitivity tests for matched data. R package version 0.1

Lawrence, I., Rigol, G., Essen, TV., & Jackson, C. (2004). A historical perspective on the content of the SAT. In R. Zwick (Ed.), *Rethinking the SAT: The Future of Standardized Testing in University Admissions* (p. 57-74). New York: RoutledgeFarmer.

Morgan, S. (2001). Counterfactuals, causal effect heterogeneity, and the catholic school effect on learning. *Sociology of Education, 744,* 341–374.

Powers, D. & Rock, D. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36* (2), 93-118.

R Development Core Team. (2008). R: A language and environment for statistical computing. Available from http://www.R-project.org

Rosenbaum, P. (2002). *Observational studies* (2nd ed.). New York: Springer.

Rosenbaum, P. & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70* (1), 41-55.

Rubin, D. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D. (2001). Using propensity scores to help design observational studies: Applications to the tobacco litigation. *Health Services & Outcomes Research Methodology, 2*, 169-188.

Rubin, D. (2006). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26* (1), 20-36.

Ruhm, K. & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review, 26,* 33-51.

## Appendix: Variable Index

1. *SATM*: Score on the SATM. Based on ELS:02 [txsatm].
2. *SATV*: Score on the SATV. Based on ELS:02 [txsatv].
3. *PSATM*: Score on the PSATM. Based on ELS:02 [f1rpsatm].
4. *PSATV*: Score on the PSATV. ELS:02 [f1rpsatv].
5. *AGE*: Approximate age in months at time of first follow-up survey. Modification of ELS:02 [bydob_p].
6. *SES*: The SES index created for ELS. Based on ELS:02 [byses2].
7. *FEMALE*: Dummy variable indicating respondent is female. Modification of ELS:02 [bysex].
8. *RACE*: Composite variable indicating whether respondent is black or a Native American (including natives of Alaska and Hawaii). Based on ELS:02 variables byrace_2, byrace_4, and byrace_5.
9. *ASIAN*: Dummy variable indicating respondent is Asian. Based on ELS:02 [byrace_3].
10. *HISPANIC*: Dummy variable indicating respondent is Hispanic. Based on ELS:02 [bys15].
11. *PRIVATE*: Dummy variable indicating whether respondent attends private school. Modification of ELS:02 [bysctrl].
12. *RURAL*: Dummy variable indicating whether respondent attended a rural school. Modification of ELS:02 [byurban].
13. *URBAN*: Dummy variable indicating whether respondent attended an urban school. Modification of ELS:02 [byurban].
14. *AP*: Dummy variable indicating whether respondent took an AP or IB class. Modification of ELS:02 [f1rapib].
15. *REM_ENG*: Dummy variable indicating whether respondent took a remedial English class. Based on ELS:02 [bys33d].
16. *REM_MATH*: Dummy variable indicating whether respondent took a remedial Math class. Based on ELS:02 [bys33e].
17. *COLL_PREP*: Dummy variable indicating whether respondent's high school program was college preparatory. Modification of ELS:02 [byschprg].
18. *BYREAD*: Student's score on the ELS:02 administered Base Year reading test. Based on ELS:02 [bytxrstd]
19. *BYMATH*: Student's score on the ELS:02 administered Base Year math test. Based on ELS:02 [bytxmstd].
20. *MATH_CRD*: Number of math credits from the student's transcript around the time of the first follow-up survey. Based on ELS:02 [f1rhma_c].
21. *HW*: Dummy variable indicating whether respondent spent more than 10 hours/week on homework. Modification of ELS:02 [bys34b].
22. *ESL*: Dummy variable indicating that respondent speaks English as a second language. Modification of ELS:02 [bys67].
23. *EDU_AFTER_HS*: Dummy variable indicating whether respondent planned to continue education immediately after high school. Modification of ELS:02 [bys57].
24. *COLLEGE_INFO*: Dummy variable indicating that a respondent did not seek college information from any of the listed sources (parents, counselors, etc). Based on ELS:02 [bys59k].
25. *PRNTS_DISC_PREP*: Dummy variable indicating whether respondent discussed SAT/ACT preparation with parents often. Modification of ELS:02 [bys86f].
26. *PRNTS_DISC_SCH*: Dummy variable indicating whether respondent discussed school courses with parents often. Modification of ELS:02 [bys86a].
27. *COACH*: Dummy variable indicating whether respondent received coaching for the SAT/ACT. Modification of ELS:02 [f1s22b].
28. *UNDERPERFORM.M*: Dummy variable indicating that a student may have underperformed on the Base Year Math test relative to their GPA. Defined further in the data section.
29. *UNDERPERFORM.V*: Dummy variable indicating that a student may have underperformed on the Base Year Reading test relative to their GPA. Defined further in the data section.
30. *NERVES.M*: Dummy variable indicating that a student's PSATM score was much lower than anticipated based on the Base Year Math score, possibly due to nervousness. Defined further in the data section.
31. *NERVES.V*: Dummy variable indicating that a student's PSATV score was much lower than anticipated based on the Base Year Reading score, possibly due to nervousness. Defined further in the data section.
32. *HI_SES*: Dummy variable indicating that a student's value on the ELS ses variable was in the top quarter for all students in the survey. Modification of ELS:02 [byses2].

## Notes

1. In particular, his methodology reduced pre-matching imbalances between treatment and control subjects substantially in 27 covariates and still allowed more of the data to be used than would have been possible in a regression analysis (Hansen, 2004, p. 617).
2. The PSAT is essentially a pre-test for the SAT taken by grade 11.
3. "Substantially worse" here is defined as having a PSAT score less than two RMSEs below their predicted score based on the ELS:02 Base Year test.
4. The variables with the bulk of the missing data were ASIAN, BLACK, NATIVE, HISPANIC, REM_ENG, REM_MATH, COLLEGE_INFO, PRNTS_DISC_PREP, and PRNTS_DISC_SCH
5. For those variables with this extra level, the variable name shown in later tables was modified to include a postscript of "1" or "NA". For example, "REM_ENG" becomes either "REM_ENG1" or "REM_ENGNA".
6. It is worth noting that this choice often appears rather fluid in empirical applications of PSM. Because there are few guiding principles that inform the number of subclasses, some researchers essentially use this as a variable that can be manipulated to demonstrate that sufficient balance has been obtained.
7. We have taken a parametric approach to estimate standard errors for the purpose of conducting tests of significance. These standard errors are presented in Table 4. One potential alternative to this parametric approach would be the bootstrap, but this was shown to be a poor choice for matched data (Abadie and Imbens, 2006). Ho et al. (2007) take the position that since pretreatment variables are typically assumed to be fixed and exogenous, standard parametric adjustments are appropriate. Further discussion of this issue is outside the scope of the present study.
8. For example, see , as in Hong & Raudenbush (2006). Though it is not made explicit in the paper that a stepwise selection method was used, this was indicated to us in a personal correspondence (G. Hong, Personal Communication, September 8, 2008).
9. In the terminology of Rosenbaum (2002), these ranges correspond to a $\Gamma$ of 1.3.
10. An alternative to pair matching that still preserves the notion of finding specific matches for each treatment unit would be fixed ratio matching or variable matching.

Send correspondence to: Ben Domingue
University of Colorado
Email: benjamin.domingue@colorado.edu

This research was supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Sciences Foundation and the National Center for Education Statistics of the Institute of Education Sciences (U.S. Department of Education) under NSF Grant #DRL-0634035. Opinions reflect those of the authors and do not necessarily reflect those of the granting agencies.