The Effect of Admissions Test Preparation: Evidence from NELS:88

Derek C. Briggs

Introduction

www.photodisc.com

For students planning to apply to a four-year college, scores on standardized admissions tests — the SAT I or ACT — take on a great deal of importance. It may be the quality and quantity of an applicant's high school coursework that receives the closest scrutiny at the more prestigious institutions, but these are cumulative indicators of performance. Standardized admissions tests, by contrast, are more of a one-shot deal. Such tests are blind to a student's high school record — instead, they are intended as an independent, objective measure of college "readiness." For students with a strong high school record, admissions tests provide a way to confirm their standing. For students with a weaker high school record, admissions tests provide a way to raise their standing. A principal justification for the use of the SAT I and ACT in the admissions process is that such tests are designed to be insensitive to the high school curriculum and to short-term test preparation. If short-term preparatory activities prior to taking the SAT I or ACT can have the effect of significantly boosting the scores of students above those they would have received without the preparation, both the validity and reliability of the tests as indicators of college readiness might be called into question.

There is an emerging consensus that particular forms of test preparation have the effect of improving scores on sections of the SAT I for students who take the tests more than once. That such an effect exists is not under dispute. The

actual magnitude of this effect remains controversial. Some private tutors claim that their tutees improve their combined SAT I section scores on average by over 200 points. Commercial test preparation companies have in the past advertised combined SAT I score increases of over 100 points. There are two reasons to be critical of such claims. First, any estimate of a commercial program effect must be made relative to a control group of students who did *not* prepare for the test with a commercial program. If test preparation companies or private tutors advertise only the average score gains of the students who make use of their services, the "effect" of this preparation is misleading. A second related problem is that students are not assigned randomly to test preparation conditions but self-select themselves into two groups, those receiving the preparatory "treatment" and those receiving the preparatory "control." Because the two groups of students may differ along important characteristics related to admissions test performance, any comparison of average score gains that does not control for such differences will be biased.

When researchers have estimated the effect of commercial test preparation programs on the SAT while taking the preceding factors into account, the effect of commercial test preparation has appeared relatively small. A comprehensive 1999 study by Don Powers and Don Rock published in the *Journal of Educational Measurement* estimated a coaching effect on the math section somewhere between 13 and 18 points and an effect on the verbal section between 6 and 12 points. Powers and Rock concluded that the combined effect of coaching on the SAT I is between 21 and 34 points. Similarly, extensive meta-analyses conducted by Betsy Jane Becker in 1990 and by Rebecca DerSimonian and Nan Laird in 1983 found that the typical effect of commercial preparatory courses on the SAT was in the range of 9-25 points on the verbal section and 15-25 points on the math section.

One of the most remarkable aspects of this line of research has been the lack of impact it has had on the public consciousness. The proportion of test-takers signing on for commercial test preparation shows no signs of abating, and many companies are now expanding their efforts into online test preparation. Furthermore, the widespread perception remains that students participating in commercial test preparation will improve their test scores dramatically rather than marginally. One explanation for this phenomenon may be a certain degree of suspicion regarding the motivations of those who have found small effects for commercial test preparation. Most researchers with access to student scores from the SAT I and ACT are themselves affiliated with the companies designing the tests. Faced with conflicting messages about the effectiveness of test preparation, the public may choose to embrace the more optimistic one.

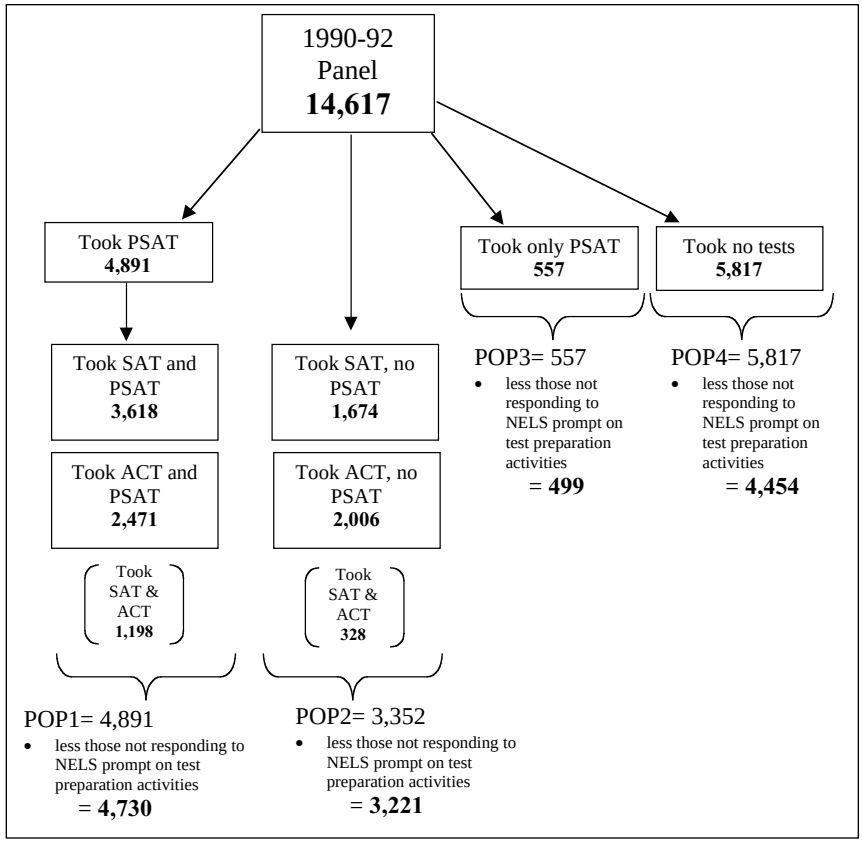

Figure 1. NELS:88 sample populations considered in analysis.

Having no affiliation with either companies that test students or prepare students to be tested, I am throwing my hat into the ring with an analysis based on data taken from the National Education Longitudinal Survey of 1988 (NELS:88, hereafter referred to as "NELS"). NELS tracks a nationally representative sample of U.S. students from the 8th grade through high school and beyond. A panel of roughly 16,500 students completed a survey questionnaire in the first three waves of NELS — 1988, 1990, and 1992. For the purposes of this study, the relevant sources of information are specific student responses to survey items, high school transcript data, and standardized test scores collected during the first and second follow-ups of NELS. All of the NELS proxies for student performance used in this study, including variables for PSAT (essentially a pretest for the SAT), SAT, and ACT scores, derive from transcript data. Prior to 1993, the SAT I was known simply as the SAT. Because the data collected in NELS come from before 1993, I shall refer to the test as the SAT instead of the SAT I.

The NELS Data

Figure 1 presents a flow chart that details the sample of students used in this study. The target population in NELS is not those students taking the SAT or ACT in American high schools but rather all American high school students who

could have taken either the SAT or the ACT. Starting from the 14,617 students who both completed student questionnaires in 1990 and 1992 and for whom transcript data was collected, there are effectively four sample populations: The first consists of students who took the PSAT and also the SAT or ACT. The second consists of students who did not take the PSAT but did take the SAT or ACT. The third consists of students who took only the PSAT. The fourth sample population includes students who took none of the tests.

The focus in most past studies has been on those students in the first sample population for whom there is a test score *before* a subsequent test preparation treatment is introduced. It may be the case, however, that test preparation activities are actually most helpful for students in the second population who have not had the prior experience of taking the test. Finally, the third and fourth populations of students are of interest if there is reason to believe some or many of these students had college aspirations but self-selected themselves out of the other sample populations because they expected to do poorly on the SAT or ACT. In theory at least, if test preparation activities are effective in the short run, these are the students who might have had the most to gain from them.

The test preparation indicators used in this study were created from the following item in the NELS second follow-up questionnaire:

To prepare for the SAT and/or ACT, did you do any of the following?

A. Take a special course at your high school
B. Take a course offered by a commercial test preparation service
C. Receive private one-to-one tutoring
D. Study from test preparation books
E. Use a test preparation video tape
F. Use a test preparation computer program

What Are the Characteristics of Students Taking and Not Taking Admissions Tests?

It is reasonable to expect that students taking admissions tests are more academically able than those students choosing not to take admissions tests, given that the former group is planning to attend a four-year college. This is borne out by the NELS data. Academic ability is roughly monotonic as a function of sample population membership. On average, students who take admissions tests perform better on the external tests of academic achievement taken by students in the NELS sample. In addition, such students tend to take more math courses while in high school and get better grades in them than students taking fewer to no admissions tests.

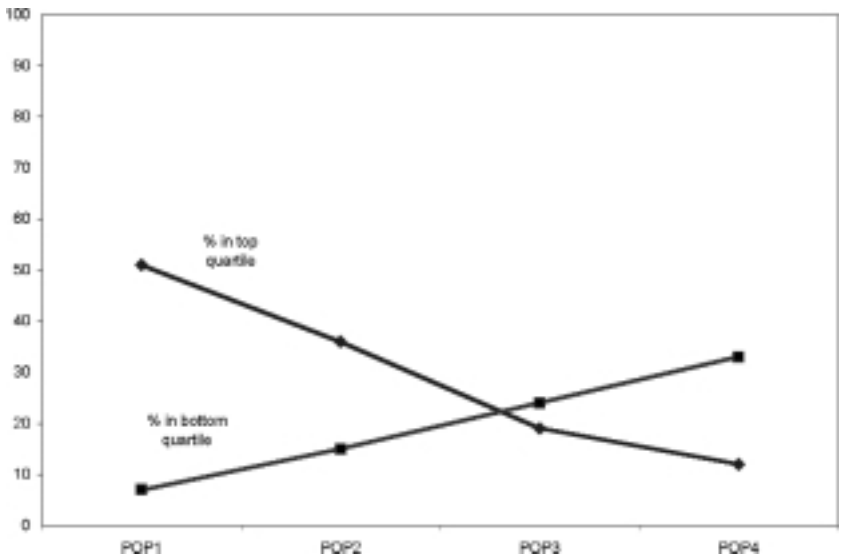

Figure 2. Proportion of students in top and bottom quartiles of SES index.

The demographic characteristics of students taking and not taking admissions tests are striking. In the two sample populations with students taking admissions tests, 13% and 17% of the test-takers are black or Hispanic. In the two sample populations in which students did not take admissions tests, the proportions of black and Hispanic students increase to 27% and 30%. Differences in socioeconomic status (SES) among the sample populations is also dramatic. The NELS SES variable combines information on household education, income, and occupational levels into a single index variable for each student. Generally, students with high SES index scores come from more educated, wealthier, and more successful households than students with low index scores. Figure 2 plots the percentages of students in the top and bottom quartiles of the SES index as a function of sample population membership. Students taking admissions tests are much more likely to be in the top quartiles of the SES index; students not taking admissions tests are much more likely to be in the bottom quartile.

Although over 6,000 students from the NELS sample did not take the SAT or ACT, many of these students nonetheless indicate that they engaged in test preparation activities. As Table 1 shows, the proportion of students engaging in test preparation activities is remarkably similar across the four sample populations. Among the students who took no admissions tests and responded to the NELS prompt regarding their test preparation activities, 8% indicated that they enrolled in a commercial preparation program, 7% indicated that they made use of a private tutor, and 40% claimed to have studied with books. This suggests that a significant number of students may consider taking the SAT or ACT while in high school but select themselves out of these sample populations because their test preparation activities are either discouraging or indicate that they will perform poorly on the exam. If this is true, then any study seeking to evaluate the effectiveness of test preparation activities using only the sample of students taking admissions tests is likely to be biased upward, depending on the number of students who opt out of such tests

Table 1 — Proportions of NELS:88 Sample Populations Engaged in Various Test Preparation Activities

Test Preparation Activity*	POP1	POP2	POP3	POP4
Special high school course	21	17	15	14
Commercial course ("coaching")	14	12	8	8
Private tutoring	7	8	6	7
Study from books	63	60	48	39
Use of video tape	5	7	7	8
Use of computer program	13	12	10	9

POP1=Student took PSAT and SAT and/or PSAT and ACT
POP2=Student took SAT and/or ACT but not PSAT
POP3=Student took PSAT but not SAT or ACT
POP4=Student took neither PSAT, SAT, or ACT
*Proportions are of students in each sample population responding to NELS prompt on test preparation activities. Population sizes:
POP1 = 4,730 | POP2 = 3,221 | POP3 = 499 | POP4 = 4,454

Table 2 — Mean PSAT and SAT Scores for Students Taking Both Tests

PSAT	SAT	Gain
	Verbal	
439 (103)	466 (107)	27 (52)
	Math	
489 (109)	522 (114)	33 (58)

Standard deviations in parentheses, n = 3,494.

Table 3 — Mean PSAT and ACT Scores for Students Taking Both Tests

PSAT	ACT
Math	*Math*
475 (110)	22.2 (4.8)
Verbal	*English*
424 (98)	22.4 (5.0)
	Reading
	23.4 (6.0)

Standard deviations in parentheses, n=2,364.

after participating in preparatory activities.

Comparing Test Scores Without Controlling for Self-Selection

At this point, I restrict attention to the 4,730 students in the first sample population who have taken both the PSAT and SAT or ACT and responded to the survey question regarding their test preparation activities. It would be preferable to have data on students who have taken the SAT or ACT twice when considering score changes. Instead, PSAT scores are used as proxies for the SAT and ACT. This is reasonable since the PSAT — which is essentially a pretest for the SAT — is very similar in structure to the SAT, with multiple-choice verbal and math sections. The scores of students on each section of the PSAT have a very high correlation (almost .9) with their scores on the corresponding sections of the SAT. The ACT is different in structure than the PSAT; however, performance on the two tests is also highly correlated. The sections of the ACT most comparable to sections of the PSAT are the English, reading, and math sections. Student scores on the English and reading sections of the ACT have correlations of .8 with scores on the verbal section of the PSAT. The correlation of the PSAT and SAT verbal sections is only .08 higher. Similarly, student scores on the math section of the ACT have a correlation of .82 with scores on the math section of the PSAT, just .05 less than the PSAT-SAT math section correlation.

Previous studies have compared raw scores from the PSAT to SAT by multiplying PSAT scores by 10. The same tactic is taken here to illustrate an approach commonly taken in the analysis of test score changes. Tables 2 and 3 show the mean and standard deviation of student scores on the PSAT, SAT, and ACT. On average, students taking the test at least twice improved their scores on the SAT by about 33 points on the math section and about 27 points on the verbal section. Without knowing anything at all about student characteristics or test preparation activities, one might reasonably expect the combined SAT scores for any given student to increase by about 60 points, just by waiting a year and taking the test again. The question of interest here is whether students who prepare for the test in certain ways score significantly above this average. I consider a naïve and then, in the next section, a less naïve way to answer this question.

Table 4 compares the differences in mean PSAT-SAT section scores changes by splitting test-takers into dichotomous groupings as a function of their test preparation activities. A student is categorized as either making use or not making use of a particular preparation activity. Columns 3 and 4 show the "effects" of each of the six forms of test preparation — taking a course offered in high school, enrolling in a course offered by a commercial test preparation company, getting private tutoring, studying with a book, studying with a video, and studying with a computer. By far the largest effect sizes belong to those preparation activities involving either a commercial course or a private tutor, and the effects differ for each section of the SAT. On average, students with private tutors improve their math scores by 19 points more than those students without private tutors. The effect is less on the verbal section, where having a private tutor only improves scores on average by seven points. Taking a commercial course has a similarly large effect on

Table 4 — Raw "Effects" of Various Test Preparation Activities

SAT Preparation Activities	Number in Treatment Group	Change in SAT-M	Change in SAT-V
High school offered class to prepare for SAT	793	3 (2)	2 (2)
Took commercial class to prepare for SAT	573	17 (3)	13 (2)
Used private tutor to prepare for SAT	265	19 (4)	7 (3)
Used book to prepare for SAT	2,215	7 (2)	4 (2)
Used computer to prepare for SAT	473	0 (3)	0 (3)
Used video to prepare for SAT	173	0 (5)	-2 (4)

Note: Standard errors are in parentheses. Total N for each category = 3,492.

math scores, improving them on average by 17 points, and has the largest effect on verbal scores, improving them on average by 13 points. With the exception of studying with a book, no other activity analyzed in this manner has an effect on test score changes that is statistically different from 0 at a .05 significance level.

Depending on the relative characteristics of the students in the various test preparation categories, test score differences as presented above may be misleading. If the students who have prepared for an admissions test with a particular activity tend to be academically stronger or more motivated than the students not preparing with that activity, then one might expect the score increases of the 'test prep' group to be higher irrespective of the test preparation activity undertaken. If this is the case, then estimates of preparation effects based solely on test score comparisons are likely biased upward. If the converse is true — students engaging in test preparation activities are less motivated or academically inclined—then estimates of preparation effects are likely biased downward.

Most studies have focused on estimating the effect of one specific type of test preparation, known as "coaching." In this analysis, students have been coached if they have enrolled in a commercial preparation course not offered by their school but designed specifically for the SAT or ACT. The distinction made here is whether a test-taker has received systematic instruction over a short period of time. Preparation with books, videos, and computers is excluded from the coaching definition because, although the instruction may be systematic, it has no time constraint. Preparation with a tutor is excluded because, although it may have a time constraint, it is difficult to tell if the instruction has been systematic.

Figures 3 and 4 plot students' SAT section scores relative to how they scored on the PSAT. Students who were coached are indicated by solid circles; uncoached students are indicated by empty circles. These scatterplots show that there is a great deal of variance in score changes for each group. The association between test performance is strong, yet many coached students performed significantly worse on the SAT than they did on the PSAT, and conversely many uncoached students performed significantly better than they did on the PSAT. On average, coached students do improve their SAT scores slightly more than uncoached students. The question that must be addressed is whether this difference in means is being confounded by corresponding differences in the characteristics of coached and uncoached students.

In fact, the characteristics of coached test-takers differ significantly relative to uncoached test-takers. Coached students are more likely to be Asian and in the top socioeconomic quartile than their uncoached counterparts. Coached students spend more hours studying outside of school, are more concerned about the reputations of the colleges to which they plan to apply, are more likely to have a private tutor helping them with their schoolwork, and are more likely to be encouraged by their parents to prepare for the SAT or ACT. Coached students are more likely to have higher scores on both sections of the PSAT. Interestingly, both groups are fairly similar along the range of other measures intended as proxies for academic ability.

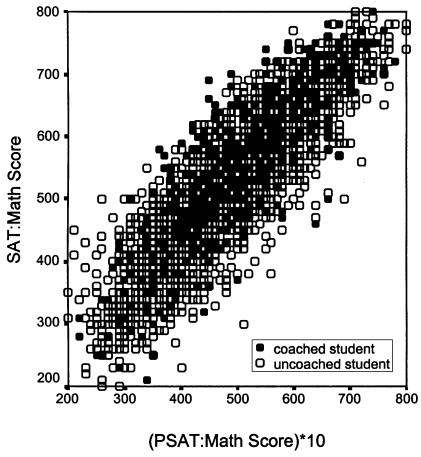

Figure 3. Scatterplot of students' scores on SAT math section and scale PSAT math scores.

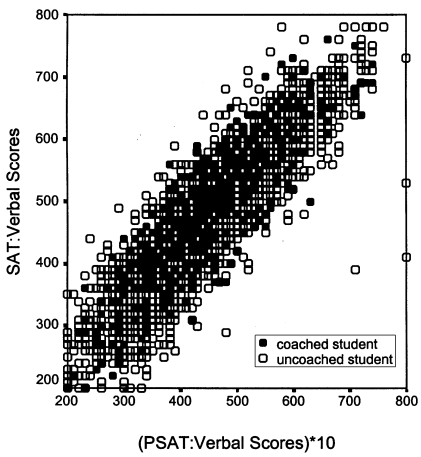

Figure 4. Same as Fig. 3 with "verbal" scores rather than "math" scores.

Table 5 — The Effect of Coaching on the SAT Under Linear Regression Models

	SAT-Math			SAT-Verbal		
	X1	X2	X3	X1	X2	X3
Coached/Total Students	573/3492	572/3468	379/2175	573/3492	572/3468	379/2175
% Coached	16%	17%	17%	16%	17%	17%
Adj R^2	.76	.79	.79	.78	.80	.81
Coaching Effect	19 (3)	14 (3)	15 (3)	14 (2)	8 (2)	6 (3)

X1: Baseline repeated measures model with no control variables other than previous test score.

X2: Additional control variables include demographic variables and indicators of student high school performance.

X3: Full model with all theoretically relevant NELS:88 control variables. Additional control variables include proxies for student motivation and dummy variables for other test preparation activities. Standard errors in parentheses.

In both groups over half the students scored in the top quartile of standardized tests in math and reading administered as part of NELS in the 10th grade. On average, both groups took the same number of math courses, and both groups got roughly the same grades in those courses. Finally, the two groups differ in their other test preparation activities. Coached students are more likely to make use of other test preparation resources, particularly private tutors, books, and computers.

The picture that emerges is that of a coached group of students who are wealthier, more motivated, and generally more prepared to take the SAT or ACT than uncoached students. It is not clear that the coached group is necessarily composed of academically "smarter" students. This pattern of differences suggests that an analysis restricted to test score changes will overestimate the effect of coaching. A less naïve estimate of the coaching effect involves the use of linear regression to control for group differences.

Controlling for Self-Selection Bias with Linear Regression

Using linear regression, the effect of coaching can be estimated from the model indicated by the equation

$$\text{Test Score} = b_0 + b_1 \text{ Coach} + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n + \text{error}$$

In this equation "Test Score" denotes score values on a particular section (e.g. math or verbal) of a standardized admission exam for a given sample of test-takers. The terms x_2 to x_n represent a set of variables thought to be related to performance on an admissions exam. They are included in the equation to hold constant quantifiable group differences between coached and uncoached students. I refer to these as control variables. "Coach" is the treatment of interest in this equation and equals 1 if a student has been coached on the test and 0 otherwise. Finally, "error" represents a random error term, assumed to average 0 across all students. Later I consider the significance of this assumption with respect to bias in the estimate of the coaching effect. For now I focus just on the results of linear regressions that model the effect of coaching on both the SAT and ACT. The effect of coaching, ($\hat{b}_1$), is estimated by fitting this regression model to the NELS data.

The Effect of Coaching on SAT Scores

The SAT has two sections that assess mathematical and verbal ability. The sections are timed and the questions are all multiple choice. Scale scores for each section of the test range from 200 to 800. Table 5 presents the results for linear regressions with three differing specifications of the control variables: X1, X2, and X3. In all specifications, the treatment of interest is the Coaching variable. The specifications differ in the degree to which they adjust for student differences. Under specification X1, a single control variable is included for a student's previous score on the PSAT section associated with Test Score. This simple repeated measures model is useful as a baseline for estimates of the coaching effect. The coaching effect estimated is the improvement for a student relative to a peer with the same PSAT score. The specification of X2 is an attempt to approximate the 1999 model developed by Powers and Rock using NELS variables to control for demographic background and academic ability. Here control variables include previous scores on both PSAT sections, dummy variables for student ethnicity, the SES index variable, and two proxies for student performance in high school, the number of math courses taken and the GPA from these courses. Now the estimated coaching effect is the improvement for a student relative to a peer with the same PSAT score, demographic background, and academic ability. We might expect that controlling for these other factors will lessen the effect. Finally, under specification X3 all NELS variables theoretically related to the improvement of SAT scores are included in the linear regression. Additional control variables include seven dummy variables that proxy for student motivation (e.g., time spent doing homework, aspirations, parental encouragement, etc.) and five dummy variables that reflect other test preparation activities besides coaching (e.g., private tutoring, use of books, etc.). Here the coaching effect estimate controls for anything that might be relevant, i.e., the

Table 6 — The Effect of Coaching on the ACT Under Linear Regression

	ACT-Math		
	X1	X2	X3
Coached/Total Students	305/2390	305/2384	208/1544
% Coached	13%	13%	14%
Adj R^2	.68	.74	.73
Coaching Effect	.61 (.17)	.33 (.16)	.27 (.2)

	ACT-English		
	X1	X2	X3
Coached/Total Students	305/2396	305/2384	208/1544
% Coached	13%	13%	14%
Adj R^2	.58	.64	.65
Coaching Effect	.38 (.20)	.33 (.19)	.55 (.23)

	ACT-Reading		
	X1	X2	X3
Coached/Total Students	305/2396	305/2384	208/1544
% Coached	13%	13%	14%
Adj R^2	.61	.63	.63
Coaching Effect	-.66 (.23)	-.75 (.23)	-.66 (.29)

X1: Baseline repeated measures model with no control variables other than previous test score.
X2: Additional control variables include demographic variables and indicators of student high school performance.
X3: Full model with all theoretically relevant NELS:88 control variables. Additional control variables include proxies for student motivation and dummy variables for other test preparation activities.
Standard errors in parentheses.

"kitchen sink." Note that this lowers the available sample size.

The estimated effect of coaching on SAT scores is statistically significant at a .05 level for all three specifications of the control variables for each section of the test. In both the math and verbal sections of the SAT the estimated effect of coaching decreases from the baseline specification when control variables are added to adjust for group differences. From X1 to X2, the estimated coaching effect decreases by roughly 25% (19 to 14) in the math section and 40% (14 to 8) in the verbal section. From X1 to X3, the estimated coaching effect decreases by about 20% (19 to 15) in the math section and 60% (14 to 6) in the verbal section. When the control variables are limited to previous score on the related PSAT section, the coaching effect is estimated as a combined increase of 33 points (19 + 14) on the SAT math and verbal sections. When the equation is adjusted with control variables for student demographics and academic ability, the combined effect drops to 22 points. When the equation is also adjusted with control variables for student motivation and test preparation activities, the combined effect decreases to 21 points.

The linear regression model specified previously includes no interaction terms. It would be reasonable to suspect that the effect of coaching might be higher for certain types of students — for example, students who scored lower on the PSAT, students who also receive private tutoring, and so forth. To this end, I considered all possible two-way interactions with the coaching variable under the control-variable specification X3. The results suggest that coaching on the math section of the SAT is most effective for students with strong socioeconomic backgrounds, students who perform well in their high school math courses, and students who are actively involved in extracurricular activities. Conversely, coaching is least effective for students who previously scored high on the math section of the PSAT and for students who employ a private tutor to prepare for the exam. For the verbal section, only one interaction is statistically significant — SES, which is again positively related to the coaching effect.

These results are consistent with the hypothesis that the uncontrolled effect of coaching is overestimated because students who enroll in commercial programs tend to be more socioeconomically advantaged, more motivated to improve their scores, and better prepared to retake the test than their uncoached counterparts.

The Effect of Coaching on ACT Scores

The ACT has a different format and scale than the SAT. Although the students taking the SAT receive separate scores on two sections of the test, students taking the ACT receive separate scores on four multiple-choice sections of the test — math, English, reading, and science — along with one composite score summarizing overall performance. Scores on each section of the ACT are reported on a scale from about 5 to 36 points.

The effect of coaching and other test preparation activities can be modeled as before, using linear regression, where the dependent variable Test Score becomes the scores of students on either the math, English, or reading sections of the ACT. Table 6 parallels the form of Table 5.

Under the baseline repeated-measures specification, the estimated effect of coaching is statistically significant only for the ACT math and reading sections. The effect size of the coaching estimate is .6 and .4, respectively. Interestingly, the sign of the coaching effect for the reading section is negative, implying that coached students on average perform worse on ACT reading questions than their uncoached counterparts after controlling for prior performance on the verbal section of the PSAT. The following are a few trends worth noting in the X2 and X3 control-variable specifications for the three sections of the ACT:

For the math section, the estimated coaching effect size decreases rather dramatically as more control variables are added to the model. When control variables for socioeconomic background

and academic ability are included under specification X2, the coaching effect decreases to just .3 points. When control variables for student motivation and test preparation activities are added under X3, the estimated effect is no longer statistically significant.

For the English section, the estimated coaching effect is not statistically significant under control-variable specifications X1 and X2. When all possible control variables are included under specification X3, the estimated effect turns significant with an effect size of .6 points.

For the reading section, the estimated negative effect size of coaching increases in absolute value when socioeconomic and academic ability control variables are added to the model. When motivation and test preparation variables are added, the effect size of coaching returns to that of the baseline model.

Regardless of control-variable specification, when rounded to ones the estimated effect of coaching in absolute value is never more than a single point for any of the three ACT sections considered here.

Interactions with the coaching variable were tested for in the English and reading ACT sections. There were no significant interactions in the reading section. For the English section there were three significant interactions with the coaching variable, all with negative signs. These interactions suggest that if students are Asian or have scored well on the verbal section of the PSAT or have parents who encourage them to prepare for the test, then they are likely to benefit less from coaching.

Does Linear Regression Account for Self-Selection Bias?

One critical assumption must hold if we are to believe that the linear regression estimate of the coaching effect is unbiased: We must assume that, conditional on the control variables included in the equation, the expected value of the error term across all students is 0. This is a strong assumption. In the context of coaching, we must believe that all the factors related to differences in the performance of coached and uncoached students on Test Score have been quantified in the equation as control variables.

Consider the scenario in which there is an omitted variable, some unobserved variable that predicts whether a student will perform well on the test in question. Consider further that this variable is positively correlated with a student's decision to seek coaching in the first place. In other words, students who are more "driven" are most likely to seek coaching, and driven students in turn are most likely the types of students that develop strong test-taking ability. Both "drive" and "test-taking ability" are unobservable yet related variables. In this scenario linear regression will not be a statistical model that produces unbiased estimates of the coaching effect.

Table 7 — Summary of Standardized Coaching Effects

Admissions Test	Coaching Effect	Standard Error
SAT-Math	14%	3
SAT-Verbal	5%	3
ACT-Math	6%	4
ACT-English	11%	5
ACT-Reading	-11%	5

Effect sizes and standard errors above are expressed as percentage of a standard deviation of the dependent variable. Estimates derived from the linear regression model specification with all control variables (X3).

Two statistical models popular in econometric research as a means for correcting the effects of selection bias are instrumental variables and the Heckman model, due to recent Nobel Prize winner James Heckman. The Heckman approach is a two-equation model that attempts to explicitly estimate and control for selection bias as an independent variable using either linear regression or generalized linear regression. (A more detailed description of this technique is outside the scope of this article.) When the Heckman model is applied to this analysis of coaching effects, the estimate of selection bias is not statistically significant for any section of the SAT or ACT, and the estimates for the coaching variable are virtually identical to those produced by linear regression.

What About Students Who Don't Take the PSAT?

Earlier the point was made that the effect of test preparation, and coaching in particular, might be the largest for students who do not take the PSAT first, precisely because test preparation activities might replace the experience of actually taking the SAT or ACT. This hypothesis can be tested by comparing the scores of students in the second sample population, controlling for their demographic characteristics, academic background, motivational proxies, and various test preparation activities with linear regression.

For students who do not take the PSAT first, the estimated effect of coaching is not statistically significant for any of the sections of the SAT or ACT. Coaching and other forms of test preparation do not seem to be particularly effective for students who have not had previous exposure to admissions tests in the form of the PSAT. In fact, the largest significant effect size for a test preparation variable is a negative one associated with the use of a preparatory video.

Conclusion

Does test preparation help improve student performance on the SAT and ACT? For students who have taken the test before and would like to boost their scores, coaching seems to help, but by a rather small amount. After controlling for group differences, the average coaching boost on the math section of the SAT is 14 to 15 points. The boost is smaller on the verbal section of the test,

just 6 to 8 points. The combined effect of coaching on the SAT for the NELS sample is about 20 points. The effect of coaching is similar on comparable sections of the ACT. The average score increase on the ACT math section probably lies within the range of 0 to .4 points, while the coaching effect on the English section is about .3 to .6 points. On the ACT reading section, coaching actually has a negative effect of about .6 to .7 points. Table 7 summarizes these empirical results, reporting coaching effect sizes in terms of standard deviations for both the SAT and ACT.

This analysis suggests unequivocally that the average effect of coaching is nowhere near the levels previously suggested by commercial test preparation companies. Private tutoring has a similarly small effect for students taking the math section of the SAT and no effect for students taking the math section of the ACT. Whether these benefits are worth the cost—commercial programs can charge anywhere from $700 up to $3,000, while private tutors often charge as much as $450 per hour—is unclear.

It is a potentially troubling finding in this study that there seem to be a significant number of students with aspirations for a college education who select themselves out of the sample of students taking college admissions tests. Students who engage in test preparation activities but choose not to take an admission test tend be less academically able and much less socioeconomically advantaged than their test-taking counterparts. These are not necessarily students who are unfit for college admission. Ideally, coaching should be most effective and at least readily available to these types of students, but in practice this does not seem to be the case.

A report in the *New York Times* (January 10, 1999) suggested that the benefits of coaching and private tutoring may extend beyond potential admission test score improvements by teaching students better study habits and imbuing them with greater discipline and self-confidence. This certainly might be the case. The data used in this analysis do not consider the potential side benefits of commercial test preparation. Furthermore, the data used here are from the early 1990s and may not reflect the state of the world 10 years later. It is possible that specific programs and tutors currently exist capable of producing higher than average score gains. The evidence for this, however, seems anecdotal at best. With respect to the NELS dataset, there is no evidence that commercial test preparation makes much of a difference in admissions test performance. Students and their parents should be careful before investing in test preparation with the expectation of dramatic improvements in SAT or ACT test scores.

References and Further Reading

Becker, B.J. (1990), "Coaching for the Scholastic Aptitude Test: Further Synthesis and Appraisal" *Review of Educational Research*, 60, 373–417.

DerSimonian, R., and Laird, N.M. (1983), "Evaluating the Effect of Coaching on SAT Scores: A Meta-analysis," *Harvard Educational Review*, 53, 1–15.

Heckman, J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

Powers, D.E. (1993), "Coaching for the SAT: A Summary of the Summaries and an Update," *Educational Measurement: Issues and Practice*, 12, 24–39.

Powers, D.E., and Rock, D.A. (1999), "Effects of Coaching on SAT I: Reasoning Test Scores," *Journal of Educational Measurement*, 36, 93–118.

Schwartz, T. (1999), *The Test Under Stress*, New York: The New York Times, Jan. 10, Section 6, Page 30, Column 1.

Smyth, F.L. (1989), "Commercial Coaching and SAT Scores," *The Journal of College Admissions*, 123, 2–9.

——— (1990), "SAT Coaching: What Really Happens and How We Are Led to Expect More," *The Journal of College Admissions*, 129, 7–17.

Editor's Note: The Briggs study is noteworthy, and appears here, because the author is not affiliated with either a coaching service or a testing service. We thought it would be interesting to have comments from representatives of the two interested parties. Donald Powers of Educational Testing Service agreed and his comments follow. We had a committment from a researcher at a coaching service to provide comments, and sent a copy of the article to that researcher. We did not receive comments; phone calls and emails have gone unanswered for more than a month.

Forthcoming Articles in *Chance*

Measuring Bird Migration *by K. Cottrell, E. Bernstein, N. Altman, A. Dhondt, W. Hochackna, and R. Slothower*

Ancient Geometry in Image Processing *by Matt Reed and Vyvyan Howard*

Market Research Reveals Another Paradox of Means *by Randy Mason*

Surprises From Self-Examination *by Seth Roberts*

Alleged Racial Steering in an Apartment Complex *by Jason Connor and Joseph Kadane*

Three Sisters Give Birth *by Nathan Wetzel*

Comment:

Using National Education Longitudinal Study (NELS) Data to Evaluate the Effects of Commercial Test Preparation

Donald E. Powers

By providing an analysis of the National Education Longitudinal Study (NELS) database, Derek Briggs makes an important contribution to the literature on the effects of coaching for college admissions tests. First of all, unlike many of the "so-called" studies of coaching — typically, loosely designed surveys of previously coached students — the analysis offered by Briggs *does* qualify as a legitimate scientific inquiry, because he is clearly stressing one incontestable fact. As emphasized elsewhere, if estimates of the effects of commercial coaching are to have any scientific credibility whatsoever, they must, at a minimum, be made in relation to a comparison group that does not receive coaching. Thus, if there is a single message in Briggs's report that bears repeating, it is this:

> Test score *gains* made by coached students from one occasion to another cannot legitimately be regarded as equivalent to the *effects* of coaching. In short, GAINS ≠ EFFECTS!

The limitations of one-group, pre-post evaluations — the kind on which most coaching companies base their claims — have long been acknowledged. For coaching studies, the threats to inferences about program effects that are not controlled by this design are mainly three:

1. Growth/history (students improve on the knowledge and abilities measured by assessments like the ACT and SAT, regardless of whether they are coached)
2. Test practice (students get better at taking these tests just by taking these tests)
3. Measurement error (because test scores aren't perfectly reliable, they will change from one occasion to another, even in the short term, regardless of any intervening experiences)

Despite these uncontrolled factors, commercial coaching services virtually always tout the effectiveness of their programs in terms of pre-post test *gains* made by their customers. In effect, they take credit not only for that component of test-score gain that is properly attributable to their efforts but also for the parts that are due to other factors, such as those just listed. There is clear evidence, however, that, as an index of effectiveness, test-score gains seriously overestimate the real impact of coaching. The claims made by coaching companies (average "effects" of 120-140 points) appear to be inflated by as much as 100 points for the SAT (see Powers and Rock, 1999).The Briggs analyses are consistent with this conclusion.

The illogic underlying the claims for coaching becomes clearer when one considers *losses* as well as gains. If coaching schools are to use the test-score changes exhibited by their clients as the basis for evaluating their effectiveness, they must, it seems, take credit not only for any gains made by students who attend their programs but also for any losses. Because, as Briggs (and others) have shown, the test scores of some students do decrease after they are coached, this acknowledgement would in effect constitute an admission that coaching can be harmful. Although it may be unreasonable to think that coaching schools are liable for decreases in test scores, it is equally untenable to believe that they are responsible for all increases.

Yet much of the public seems prepared to accept evidence based on test-score gains as a legitimate indication of the effectiveness of commercial coaching. Why is this so? Perhaps, the threats to conclusions based on gains are less apparent for studies of mental abilities than they are for other traits, say physical abilities. For example, a physician whose treatment promised to stimulate the physical growth of adolescent students (to increase their height dramatically, for instance) would probably be called into question by discerning parents, who might be inclined to ask whether their offspring would grow despite the treatment. The physician could further ensure her success by selecting only the shortest students for treatment — those most likely, by virtue of an impending "growth spurt," to

"regress" to the mean of their peers. Similarly, by attracting students who are apparently in most need of coaching — that is, those with low initial test scores — a commercial coaching could increase its *apparent* effectiveness. Predictably, the scores of low-scoring students will, on average, increase even on immediate retesting.

Strengths of the Briggs Study

There are a number of specific aspects of Briggs's study that deserve special mention:

1. It is perhaps the largest coaching study ever conducted. The database — a nationally representative sample of some 14,000 secondary students, about 10% of whom were coached — is indeed impressive. That it was assembled by a disinterested party is also notable.

2. The study is noteworthy also in that it is one of very few that examine the effects of coaching on ACT test scores. It is, very likely, the only one that has included *both* ACT and SAT takers.

3. Besides including a comparison group of uncoached students, the investigator acknowledges that coached and uncoached students differ on a variety of potentially important variables. Because failing to take these differences into account could lead to biased estimates of the effects of coaching, the investigator includes a substantial number of covariates to control for relevant pre-existing between-group differences. Some of the covariates employed (e.g., time spent on homework) seem especially appropriate as proxies for student motivation, which may be related both to test performance and to enrollment in coaching programs. These variables do not seem to have been included in previous studies of coaching.

4. The database is large enough to enable separate analyses according to whether students had taken an earlier test or whether they were testing for the first time after receiving coaching. The significance here is that the effectiveness of coaching may depend on test takers' previous test-taking experiences. Moreover, the results of previous test taking may be a factor in decisions to attend coaching programs.

5. The basic data are displayed in a simple and compelling way. Scatterplots showing first versus second test scores for coached and uncoached test takers make clear how difficult it is to distinguish coached students from their uncoached counterparts. The variation from first to second testing clearly overwhelms the variation due to coaching, and one is hard pressed to detect many clear "outliers" — that is, coached students who did markedly better on retesting than would be expected on the basis of their pre-coaching scores.

The limitations of one-group, pre-post evaluations — the kind on which most coaching companies base their claims — have long been acknowledged.

Limitations of the Study

The limitations of the study are largely due to the use of an existing database:

1. The independent variable — how students prepared for the tests — is based on student reports to a single question. These days, with so many kinds of preparation available to test takers, it is difficult, as I have found, to classify students precisely into discrete test-preparation groups. One reason for this difficulty is the significant number of "hybrid" test preparations. For example, the "special course offered at high school" might very well be given by a commercial test-preparation company, and the test-preparation book that is mentioned could constitute the major resource employed in the course. So, it is not easy, especially with a single question, to determine exactly how best to classify test takers according to what test preparations they undertook.

2. It is not crystal clear from the study description that the exact time of coaching was determined — that is, whether the earlier and later test scores accurately bracketed the duration of coaching programs. To the extent that *both* testings may have come either before or after coaching, the effects of coaching could be either underestimated or overestimated. This limitation is, however, probably not a serious source of bias.

3. No distinction is made among the wide variety of commercial test-preparation services, which clearly differ with regard to cost, emphasis, and time required of students. The length of coaching programs in particular has been shown to relate to their effectiveness, although the returns due to extra time seem to diminish. So, it is likely that averaging the effects due to all kinds of programs may slightly underestimate the effects of the most effective ones. At least that is what the study methods will permit the coaching services to claim.

Noteworthy Findings

Regardless of its limitations, the study clearly provides useful information — some that corroborates earlier research, as well as some that has not been available heretofore. The following points seem worthy of mention:

1. With respect to the magnitude of coaching effects, the study results are remarkably consistent with the results of recent individual studies and with the conclusions of several major meta-analyses (some of these are listed under additional reading). They are also, for reasons discussed earlier, clearly at odds with the claims made by commercial coaching companies. With respect to the proportion of test takers who engage in commercial coaching and/or other kinds of test preparation, the results reported by Briggs are also quite consistent with other recent surveys of students' test-preparation activities.

2. The study identifies a potential source of bias that has heretofore been unconsidered in coaching studies. Briggs suggests that coaching's effectiveness may be overestimated when, after being coached, students self-select themselves out of the test-taking population. In fact, Briggs's data show that about 27% of the 1,445 NELS students who attended commercial coaching programs apparently did not bother to take

either the ACT or the SAT after they were coached. This statistic may simply mean that a significant portion of the sample had not *yet* tested when they completed the NELS questionnaire. If, however, as Briggs suggests, this dropout is the result of a discouraging test-preparation experience, it may mean that coaching companies are falling far short of their major objective — to help students gain entry to the "colleges of their choice."

3. The finding that coaching had a slight *negative* effect on ACT-Reading scores is intriguing also. The first reaction is that this is simply an artifact of failing to control for some important difference between coached and uncoached students. This interpretation seems plausible for ACT scores, because the primary covariate — performance on the Preliminary SAT (PSAT) — constitutes a somewhat less powerful control variable for the ACT than it does for its "big sibling," the SAT. Thus, a potential undercorrection for ACT effects seems credible. Nonetheless, the following notion seems worth entertaining. Some coaching companies may indeed impart faulty information that could serve to lower, not improve, students scores. For example, for the new GRE computer-based tests, some coaching companies have advised test takers to spend a disproportionate amount of time on the first few questions in the test — a strategy that has been shown to be fundamentally unsound. For reading comprehension questions in particular, some coaching firms have advised students that they can save time by proceeding directly to the questions, answering them immediately without first consulting the passages on which the questions are based. This strategy has been shown to be, at best, inefficient.

Conclusion

All in all, Briggs makes an important contribution to the literature on the effects of commercial coaching for college admissions tests. At the outset of his article, Briggs notes the remarkable lack of impact that the research on coaching has had on consumer behavior ("the public consciousness"). This has been discouraging to us also. It is hoped that consumers will find Briggs's study to be useful when assessing the potential value of commercial coaching. If so, it may have a greater impact than the previous research on coaching. Unfortunately, however, many prospective consumers of commercial coaching will undoubtedly continue to rely on anecdotal accounts rather than on the kind of compelling evidence that Briggs so ably provides in his article. The challenge is to make this information widely available, presenting it in ways that will enable test takers and their parents to make informed choices about whether to purchase the services offered by commercial coaching companies.

No distinction is made among the wide variety of commercial test-preparation services, which clearly differ with regard to cost, emphasis, and time required of students.

References and Further Reading

Becker, B. J. (1990), "Coaching for the Scholastic Aptitude Test: Further Synthesis and Appraisal," *Review of Educational Research*, 60, 373–417.

Campbell, D. T., and Stanley, J. C. (1966), *Experimental and Quasi-Experimental Designs for Research*, Chicago: Rand McNally.

DerSimonian, R., and Laird, N. M. (1983), "Evaluating the Effect of Coaching on SAT Scores: A Meta-Analysis," *Harvard Educational Review*, 53, 1–15.

Kulik, J. A., Bangert-Drowns, R. L., and Kulik, C. C. (1984), "Effectiveness of Coaching for Aptitude Tests," *Psychological Bulletin*, 95, 179–188.

Messick, S., and Jungeblut, A. (1981), "Time and Method in Coaching for the SAT," *Psychological Bulletin*, 89, 191–216.

Powers, D. E. (1993), "Coaching for the SAT: A Summary of the Summaries and an Update," *Educational Measurement: Issues and Practice*, 12, 24–39.

——— (1998), *Preparing for the SAT I: Reasoning Test—An Update* (College Board Report 98-5 and ETS Research Report RR-98-34), Princeton, NJ: Educational Testing Service.

Powers, D. E., and Leung, S. W. (1995), "Answering the New SAT Reading Comprehension Questions Without the Passages," *Journal of Educational Measurement*, 32, 105–129.

Powers, D. E., and Rock, D. A. (1999), "Effects of Coaching on SAT I: Reasoning Test Scores," *Journal of Educational Measurement*, 36, 93–118.