College Board Report No. 98-6

Effects of Coaching on SAT® I: Reasoning Scores

DONALD E. POWERS
and DONALD A. ROCK

The College Board
Educational Excellence for All Students

Donald E. Powers is principal research scientist at Educational Testing Service.

Donald A. Rock is an educational consultant.

Founded in 1900, the College Board is a not-for-profit educational association that supports academic preparation and transition to higher education for students around the world through the ongoing collaboration of its member schools, colleges, universities, educational systems and organizations.

In all of its activities, the Board promotes equity through universal access to high standards of teaching and learning and sufficient financial resources so that every student has the opportunity to succeed in college and work.

The College Board champions—by means of superior research; curricular development; assessment; guidance, placement, and admission information; professional development; forums; policy analysis; and public outreach—educational excellence for all students.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101-0886, (800) 323-7155. The price is $15. Please include $4 for postage and handling.

Acknowledgments

The authors would like to acknowledge the contributions of the following people to the study reported here:

Nancy Ervin for assistance in retrieving information from test score files;

Laura Jenkins for help in matching the PSAT/NMSQT and SAT® data bases;

Laura Jerry for processing and analyzing the data;

Ruth Yoder for help with several aspects of the study, including preparation of the final report;

Larry Stricker, Spence Swinton, and Stan Von Mayrhauser for helpful comments on an earlier draft; and

the members of the Joint Staff Research and Development Committee for their support.

Contents

Abstract.....1

Effects of Coaching on SAT® I: Reasoning Scores.....1

Method.....1

Sample.....1

Data.....2

Data Analyses.....2

Analytical Models.....3

Results.....6

Case Studies.....6

Score Changes.....7

Who Seeks Coaching?.....8

Estimates of Coaching Effects.....11

Differential Effects by Examinee Subgroups.....11

Discussion.....12

References.....14

Appendix A.....15

Appendix B.....17

Tables

1. "Case Histories" for 10 SAT I Takers.....6
2. Mean Pre-SAT I, Post-SAT I, and Gain Scores for All Coached and Uncoached Examinees.....7
3. Demographic and Background Characteristics of Coached and Uncoached Examinees.....8
4. Test Preparation and Test Reactions of Coached and Uncoached Examinees.....9
5. Effects of Coaching (All Programs) Based on Alternative Models of Analysis.....10
6. Effects of Coaching by Major Programs.....10

Tables: Appendix B

B.1. Effects of Coaching (All Programs) by Gender.....17

B.2. Effects of Coaching (All Programs) by Ethnicity.....17

B.3. Effects of Coaching (All Programs) by Initial Score Level.....17

This page is intentionally blank. Please continue on to the next page.

Abstract

A College Board-sponsored survey of a nationally representative sample of 1995-96 SAT® takers yielded a data base for more than 4,000 examinees, about 500 of whom had attended formal coaching programs outside their schools. Several alternative analytical methods were used to estimate the effects of coaching on SAT I: Reasoning scores. The various analyses produced somewhat different estimates. All of the estimates, however, suggested that the effects of coaching are far less than is claimed by major commercial test preparation companies. The revised SAT I does not appear to be any more coachable than its predecessor.

Effects of Coaching on SAT® I: Reasoning Scores

"New SAT proves more coachable than old," proclaimed the January 8, 1995, edition of the *Philadelphia Inquirer*. Undoubtedly, this announcement also appeared in some form in other major newspapers. The assertion is, however, altogether inconsistent with changes that have been made to the SAT I: a greater emphasis on critical reading, the elimination of antonym questions to measure vocabulary, a reduction in the number of analogy items, more generous time limits, and the use of some math questions that require examinees to construct (rather than choose) answers. For a variety of reasons, all of these modifications should, if anything, render the revised SAT *less* coachable[1] than its predecessor. Nonetheless, two of the major purveyors of commercial coaching currently boast (on their Web sites) of large average score increases (verbal and math combined):

- 120 points by the Kaplan Educational Centers
  Nov. 26, 1997
  www.kaplan.com/precoll/courses.html
- 140 points by the Princeton Review
  Nov. 25, 1997
  www.review.com/college/sat/satCourse.efm

1. "Coaching" can take many forms ranging from short-term cramming and practice aimed solely at honing test-taking skills to long-term instruction focused on the development of knowledge and abilities (Cole, 1982; Anastasi, 1981; Messick, 1982). We have sidestepped the definition of coaching here by simply considering it to entail any and all activities conducted in special preparation programs offered to students outside their schools.

The appeal to test takers is reinforced by Princeton Review's guarantee of 100-point increases and by Kaplan's claim that 28 percent of its students improve by at least 170 points upon retesting. Other companies make similar claims.

To date, however, coaching companies have, to our knowledge, documented their claims only by surveying previous customers to ascertain score changes after coaching. Although sometimes verified by prestigious accounting firms, these survey results do not constitute scientific studies. At a minimum, it is necessary to compare these score changes with those exhibited by uncoached test takers, who, for a variety of reasons (test practice, regression effects, and real growth in the abilities measured by the test, for instance) also show improvements upon retesting. As documented elsewhere (e.g., Powers, 1993), studies published in scholarly journals simply do not support current claims about the effectiveness of coaching for the SAT: in total, the average reported effect for the Princeton Review and for the Kaplan Educational Centers, for example, is about 25–40 points on the verbal and math portions of the SAT—less if only the best designed studies are used to gauge impact. All of this evidence, however, was collected for the pre-1994 version of the SAT, not the revision that was introduced in April, 1994. Therefore, as pointed out in a special report on the new SAT (College Board, 1994), until there is carefully controlled research, no one will know for certain whether or not the new test is more or less coachable than the old one. The primary aim of the study reported here was to estimate the effects of commercial coaching on the revised SAT, now known as the SAT I: Reasoning Test and hereafter referred to as the SAT I.

Method

Sample

The source of data for the study was a survey of 1995-96 SAT I test takers, undertaken to update prior estimates of students' involvement in SAT I test-preparation activities, including commercial coaching (Powers, *Preparing for the SAT I: Reasoning Test—An Update*). The survey involved a stratified random sample of some 6,700 SAT I registrants—1 in every 200 seniors who registered for the October, November, or December 1995 SAT I administrations, and 1 in every 200 juniors who registered for the May or June 1996 administrations. An analysis of replies from some 4,200 respondents, about

63 percent of the initial sample, revealed that nearly 12 percent attended coaching programs offered outside their schools. A total of 220 examinees reported that they had attended a program given by one of two major test-preparation firms, and 287 had attended some other formal programs conducted by other companies, other organizations, or colleges and universities.

Data

Besides information about students' test-preparation activities (e.g., which of a variety of activities they had engaged in, and, for coaching programs, the name of the offerer, the duration of the program, and the dates of enrollment), we also collected a variety of other information about survey respondents. This information included:

- test takers' evaluations of their earlier test scores as estimates of their abilities (pretty good, somewhat low, or much too low);
- test-takers' accounts of how nervous they were when taking the test (extremely, very, somewhat, slightly, not at all);
- test-takers' estimates of how important it was to attain good scores on the SAT I (extremely, very, somewhat, slightly, not at all); and
- test-takers' reports of the colleges that they regarded as their "first choices."

To make use of the information about test takers' first-choice colleges, we obtained information from SAT I test files about the SAT I scores of students applying to these schools. Specifically, for each of the approximately 1,800 different colleges and universities mentioned by study participants as their first choices, we retrieved the mean SAT I verbal and math scores for 1997 college-bound seniors who requested ETS to send their scores to these institutions.

Test score histories for all survey respondents were obtained from test score files. Information included the dates of, and performance on, the PSAT/NMSQT, the SAT, and the new SAT I. Along with test scores we also retrieved examinees' responses to the Student Descriptive Questionnaire (SDQ), which all SAT I takers are asked to complete when they register to take the test. This questionnaire provided a variety of information about students' backgrounds and experiences. For our purposes, the most important elements were:

- years of study in each of several major areas (art and music, English, foreign languages, mathematics, and natural sciences)
- average grades for courses in each subject area
- overall cumulative grade average
- most recent rank in high school
- educational aspirations
- best language (English or other)
- parents' education and income
- ethnicity
- sex

Interest in these particular variables stemmed from their potential relationship both to SAT I scores and to participation in coaching programs. Without controlling for each of them (and thus the variables for which they may be proxies), it is not possible to obtain credible estimates of the impact of coaching. Even with such control, however, the results of our analyses would not be entirely unequivocal. Other (potentially numerous) important but uncontrolled differences between coached and uncoached test takers could make coaching appear to be more (or less) helpful than it really is. Nonetheless, the availability of several variables that relate strongly to SAT I scores should serve to reduce dramatically any effects of self-selection to coaching programs. Scores on the PSAT/NMSQT and on any prior SAT were anticipated to serve this function especially well.

Because much of our data was based on student self-reports, we recontacted a sample of 350 respondents approximately two weeks after receiving their responses to our test preparation survey. For this follow-up, we re-asked some of the initial survey questions in order to assess, if not the validity of student reports, at least their consistency. With respect to participation in coaching programs, for example, 96 percent of 139 follow-up respondents were consistent in their initial and subsequent responses regarding attendance at coaching courses.

Data Analyses

Initially, we identified a few carefully selected individual cases in order to make one critical point. Although sometimes thought to reflect the value of coaching, anecdotal accounts are, at best, insufficient and, most likely, very misleading estimates of the impact of coaching.

Next, we examined the simple raw gains (and losses) made by coached students and, subsequently, compared these changes with those exhibited by uncoached test

takers. Finally, we employed a variety of more sophisticated analytic methods to account for the fact that, as will be shown, coached and uncoached students differed systematically with respect to a variety of characteristics that are related to SAT I scores. These systematic differences are themselves of interest because they illuminate the bases on which students decide to enroll in coaching courses.

With regard to the more elaborate statistical analyses, two distinct approaches were followed. The first involved the use of analysis of covariance-related procedures; the second a variety of matching procedures. In some cases, both kinds of procedures were combined. The second strategy entailed the use, as comparison groups, of subsets of uncoached students who closely matched the backgrounds and other characteristics of coached test takers. These comparison groups were established with the aid of matching procedures such as those discussed by Rosenbaum (1995). Coached and (matched) uncoached groups were then compared with respect to their SAT I performances, with appropriate covariates included in order to maximize precision.

In all, six different formal computational models were used to derive estimates of the effects of coaching, both for all coached students and for those who attended major commercial coaching programs—those offered by each of two major test preparation companies and by all other programs combined. "Raw" changes in test scores were also compared for coached and uncoached examinees. (Except for lower precision, this analysis should agree, at the means at least, with the results of a repeated measures analysis.) We felt that the "real" coaching effects should lie within the range of estimates computed from the various methods. For all analyses we used only control variables that were antecedent—either logically or temporally—to decisions to attend coaching programs. Attending coaching programs did not therefore affect examinees' standing on them. Thus, their use as control variables could not increase any existing bias in the estimates of coaching effects. These variables included such demographic and background characteristics as sex, ethnicity, parental education, course-taking histories, high school grades, and earlier test scores.

To the extent that not all relevant variables have been observed, measured, and included in the various analyses, coaching effects will be underestimated, provided that unobserved variables relate in the same way, say positively, both to SAT I performance and to participation in coaching programs. Effects will be overestimated if relevant unmeasured control variables relate in opposite ways to SAT performance and to participation in coaching.

Analytical Models

Several different alternative procedures were used to analyze the data. We have not attempted to provide detailed descriptions of these procedures. Instead, we have provided references for the reader who may be interested in more information about the methods.

Model 1: Repeated measures. The first model employed was a simple repeated measures (RM) design. Only earlier SAT I scores (or PSAT/NMSQT scores) were used as the control variable. Thus, this design assumes that all "selection bias" is completely captured in the observed between-group differences in precoaching test score means. The model provides only a baseline comparison rather than a serious estimate of the effects of coaching. For this model, we included only test takers who had either previous SAT I scores or PSAT/NMSQT scores. (When both scores were available, we used SAT I scores.) Selection bias as used here refers to the fact that treatment status (i.e., membership in the coached or uncoached groups) is related to both measured and unmeasured characteristics (ability level and motivation, for example) that may themselves be related to treatment outcomes (Barnow, Cain, and Goldberger, 1980; Murnane, Newstead, and Olsen, 1985). Selection bias is of concern whenever the assignment to treatment and control groups is nonrandom (as is the case here) and/or is nonrandom conditional on the observable control variables that are used in the computational model.

Model 2: Analysis of covariance. The second computational technique reported here, the analysis of covariance (ANCOVA), makes the strong assumption that conditioning on covariates (i.e., adjusting for between-group differences) can render the coached and uncoached groups equivalent, as if there had indeed been random assignment to groups. The supposition here is that all relevant control variables have been included in the ANCOVA model and thus the analyses will yield unbiased estimates of the effects of coaching.

The assumption underlying the repeated measures model is that between-group differences in pretest scores completely capture any between-group differences on all other unmeasured variables that are related to self-selection. The ANCOVA model, on the other hand, does not make this assumption. Instead, it bases its adjustments on a total of nine covariates—two "dummy" variables indicating race/ethnicity (Asian, underrepresented minority, and white), earlier test scores in math and verbal, father's education, high school GPA, math grades (in the math equation), social science grades (in the verbal equation), and a variable reflecting the difference between an examinee's SAT I pretest

score (either verbal or math) and the mean SAT I score of applicants to the examinee's "first choice" college.

Model 3: Instrumental variable selection. The third method used here, the instrumental variable selection model (IVSM), was suggested by Barnow, Cain, and Goldberger (1980) for use in program evaluation. The IVSM attempts to correct for unmeasured selection effects by using information from two equations—the selection equation and the estimation equation. This analysis used the same variables as used for the ANCOVA model. Greene's (1992) Limited Dependent Variables computer program (LIMDEP) provided a way to implement the procedure. Although economic statisticians, particularly Heckman (1979), have devoted considerable attention to the development of methods to estimate and control for selection bias, there have been few applications of these procedures in educational research. Some notable exceptions are discussed by Murnane et al. (1985). Other economic statisticians, e.g., Greene (1981, 1997), Barnow et al. (1980), Olsen (1980), and Murnane et al. (1985), have contributed further to the early developmental work of Heckman. The variation of the original Heckman model that was used here, the IVSM approach, was suggested by both Barnow et al. (1980) and Murnane et al. (1985) to obtain not only plausible effect estimates in the presence of selection bias, but also unbiased estimates of the standard errors of these effects. Obtaining computationally efficient and unbiased estimates of the standard errors of treatment effects has proven to be difficult in much of the early work using the Heckman models. The latter accomplishment therefore represents a significant advance.

Model 4: Heckman model. The "pure" Heckman model, like the two-equation IVSM discussed above, assumes that the treatment model is incompletely specified because of two critical unmeasured variables. The first can be defined as the pretreatment "true ability" to achieve the outcome (Barnow et al.,1980); the second describes who was assigned to treatment and control groups. If an observed variable, say "t," were the only variable used to assign individuals to groups, then an unbiased estimate of the treatment effect could be obtained even without an accurate measure of "true ability." With random assignment, there would be no need to include a measure of "t" as a covariate in the outcome equation, since it would be uncorrelated with treatment status. Unfortunately, in naturalistic experiments such as this coaching study, precise measures of "t" are not available. Typically, therefore, a composite variable "t" is formed to predict who is treated and who is not. This model of the selection process, known as the selection equation, is the first of two equations of interest. Unfortunately, except when individuals are assigned to treatments either randomly or on the basis of some known observed variable, this equation is typically specified only incompletely. Usually, only proxies for "true ability," which we will call "w," are available. That is, typically neither "w" nor "t" is completely specified, as is the case for the situation we have in this coaching study. Thus, there is error in both "w" and "t," which in turn is likely to be correlated. It is necessary, therefore, to apply certain restrictions in order to purge the treatment effect, "z," in our second equation, i.e., the structural or outcome equation for the coaching effect, from any preexisting group differences in "w."

The Heckman model assumes that the errors from the two equations are normally distributed and derives a theoretical function (the so-called mills ratio) using, from the selection equation, the estimated probability of selection (i.e., being coached) and the information on who is actually coached. The validity of this derivation requires strong normality assumptions with respect to the two error distributions in the population. Assuming that the relatively strong assumption of normality of errors is correct, then the mills ratio can be entered into the outcome equation as an additional predictor. In theory, the use of this variable will purge the treatment estimate of the biasing effect of the correlated errors from the two equations. Although the parameter estimates in the Heckman model are believed to be consistent if the assumptions hold, the standard errors are not generally correct. The procedure employed here to estimate the Heckman model uses a maximum likelihood solution (Greene, 1997), which yields the correct standard errors. More technical details for the Heckman procedure are given in Appendix A.

Model 5: Propensity matching. The propensity matching model (PMM), suggested by Rosenbaum and Rubin (1983), uses the predicted probability from a probit equation to "match" students from the coached and uncoached groups on their relative probability of being coached. The objective here is to match (a) the distributions of the probability of being coached for students in the coached group with (b) a similar distribution of probabilities from the noncoached group. To the extent that this is possible, then the effects of self-selection bias can be reduced, if not eliminated. To work, this method typically requires a large reservoir of untreated (here, uncoached) students, from which a subsample can be identified that matches the probability distribution of treated (i.e., coached) students. However, even a large reservoir of uncoached students, which would facilitate good matches to coached examinees, does not guarantee that self-selection bias can be eliminated completely, as matching can be performed only on the basis of those variables that have been mea-

sured. As with other models, there may be numerous, relevant unmeasured variables related to self-selection into the coaching group that are not be considered in the analysis. This is clearly the case here, because most of the available background and ability variables (the same as those used for the ANCOVA and IVSM analyses) showed relatively modest relationships to decisions to attend coaching programs.

In the application used here, a nearest-neighbor matching procedure was implemented, where the nearest neighbors from above and below a given coached individual's probability were selected as matches for that individual. In a few cases only one nearest neighbor was chosen, if one of those was not sufficiently close. Before the matching was carried out, the average estimated probability of being coached (for those individuals who had *actually* been coached) was .21. These average probabilities are based on estimates from the probit selection equation. The average estimated probability of being in the coached sample for those who were actually in the *un*coached sample was .10. After matching was carried out, the average estimated probability of an uncoached student being in the coached sample was .18, suggesting a considerable reduction in potential bias following the matching. Our model of the selection process left much to be desired, however. Although the prediction of who would be coached was statistically significant, it was not, on the basis of the available background and ability variables, very accurate. Like the IVSM, the propensity model matches only on observed variables, but unlike the IVSM, it includes no mechanism for dealing with correlated error that results from a failure to include relevant variables in the selection equation and that may bias the coaching effects equation. Also, by its nature, the PMM significantly reduces sample sizes.

The third, fourth, and fifth methods used here, the IVSM, the Heckman, and the PMM, share one feature: They all entail modeling the selection process to control selection bias. Each method also requires the estimation of two equations. The first, a probit equation, attempts to predict from background, ability, and other measures, a student's membership in the coached versus the uncoached groups. Then, each method uses the estimated probability of membership, or a transformation of it, in a second step. The instrumental variable selection model (IVSM) uses this probability estimate as an instrumental variable in the second stage of a two-stage least squares procedure (Greene, 1981; Barnow et al., 1980). The PMM uses this estimate as a one-dimensional matching variable. The objective in the IVSM approach is to purge the explanatory variable of interest (i.e., the indicator of whether the student was coached or not) of possible selection bias due to correlated errors between the first (i.e., modeling) equation and the second (i.e., coaching effects) equation. If the residual (error) from the modeling equation is positively correlated with both SAT I performance on retesting and self-selection into the coaching group, ordinary least-squares, single-equation approaches to estimating coaching effects, e.g., analysis of covariance, will overestimate coaching effects (Greene, 1997).

Model 6: The Belson model. The sixth method used to estimate coaching effects was the so-called "Belson model." This method, which entails making adjustments on the basis of only the control-group regression equations (Belson, 1956), yields an estimate of coaching/self-selection when effect sizes are correlated with covariates. The approach has been recommended for situations in which the comparison group is much larger than the treatment group, as it was in our data set (Cochran, 1969). It also avoids some of the assumptions on which ANCOVA is based.

This procedure was used by Stroud (1980) in a reanalysis of data from a study of the effects of commercial coaching that was sponsored by the Federal Trade Commission (Sesnowitz, Bernhardt, and Knain, 1982), and later by Powers (1985), for a study of the effects of coaching for the GRE Aptitude Test. Here, this approach entailed the prediction of SAT I scores from the variety of background data, including previous test scores, that were available. These regression equations were established on the basis of only those examinees who did *not* attend coaching programs. Next, these baseline equations were used to predict the expected SAT I scores of each *coached* test taker as if he or she had *not* been coached, given his or her standing on each of the predictor variables. For each coached test taker, the effect of coaching (or, more precisely, the combined effect of coaching and self-selection to attend coaching) was defined as the difference between the actual (post-coaching) SAT I score and the score predicted from the regression equations based on uncoached students.

Initial input for all regression equations was a missing data correlation matrix based on all available variables for the uncoached sample. Only variables that correlated with both SAT I scores and with attendance at coaching programs were used in the analyses. Next, a total of 11 distinct patterns of data were identified for coached test takers. For instance, some coached examinees had only PSAT/NMSQT scores and no other background data, some had information for all variables, and so on. PSAT/NMSQT scores and/or previous SAT I scores—by far the best predictors of latter SAT I scores—were available for all but 60 coached test takers. Using data for uncoached students only, a regression equation was

TABLE 1

"Case Histories" for 10 SAT I Takers

Case	*Early Test Results*				*Intervening Test Preparation*	*Later SAT Results*				*"Effect"*	
	PSAT/NMSQT		*SAT*			*First*		*Second*			
	V	*M*	*V*	*M*	*Coached Test Takers*	*V*	*M*	*V*	*M*	*V*	*M*
1	47(0) 10/94	52(0) 10/94	380 1/95	460 1/95	Coached by major company 4/95-6/95	470 6/95	570 6/95			+90	+110
2	60(0) 10/94	57(0) 10/94	640 4/95	650 4/95	Coached by major company 8/95-9/95	620 10/95	620 10/95			–20	–30
3	57(0) 10/94	59(0) 10/94	550 4/95	710 4/95	Coached by major company 9/95-10/95	550 10/95	720 10/95			0	+10
4	51(0) 10/94	59(0) 10/94			Coached by major company 3/95-6/95	480 6/95	540 6/95	510 10/95	590 10/95	–30, 0	–50, 0
5	67(0) 10/94	77(0) 10/94			Coached by major company 3/95-5/95	650 5/95	700 5/95	670 10/95	710 10/95	–20, 0	–70, –60
					Uncoached Test Takers						
6			360 12/94	440 12/94	Used *Taking the SAT* and other test prep. materials	510 10/95	520 10/95			+150	+80
7	56(0) 10/94	55(0) 10/94	430 1/95	490 1/95	Nothing except take the PSAT/NMSQT and SAT as practice	590 10/95	540 10/95			+160	+50
8	66(0) 10/94	69(0) 10/94	570 5/95	670 5/95	Read *Taking the SAT* and reviewed math and English coursework on own	670 10/95	800 10/95			+100	+130
9			670 5/95	560 5/95	Nothing except take the SAT for practice	670 11/95	560 11/95			0	0
10			340 4/95	400 4/95	Used test prep. books and received special test prep. in class	280 10/95	350 10/95			–60	–50

Note: Multiple entries for "effects" indicate changes for first and second retests.

computed for each of these patterns to establish SAT I score expectations for each of the 11 different subsamples of coached test takers. In total, 469 coached students fit one of the 11 patterns. A total of 31 test takers who had neither precoaching test scores nor background data were excluded from the analyses.

Baseline regression equations provided very good predictions of latter SAT I scores. For SAT I verbal scores, multiple *R*s ranged from .88 to .93 for all patterns for which PSAT/NMSQT or earlier SAT I scores were available. For two patterns for which only background and questionnaire data were available, the multiple *R*s were .59 and .66. For SAT I math scores, the corresponding multiple *R*s ranged from .87 to .94 when earlier test scores were available, and .67 and .72 when they were not.

After individual effect estimates (i.e., differences between actual and predicted SAT I scores) were computed for each coached test taker, they were combined over all patterns of data. Mean effects were calculated for each of the three major categories of coaching courses, i.e., for each of two major coaching companies, and for all other kinds of courses combined. Finally, individual effect estimates were correlated with each background and questionnaire variable, and with earlier test scores, to ascertain if coaching may have been more effective for some kinds of test takers than others.

Results

Case Studies

Table 1 presents histories for 10 SAT I takers in our study. These cases were purposely selected to illustrate several points:

- Some test takers are remarkably consistent in their performance on the SAT I over time. For example, one SAT I taker shown in Table 1 (Case 9) obtained *exactly* the same SAT I verbal and math scores on two separate occasions five months apart. Most examinees, however, are much less consistent.
- There are indeed coached test takers who show dramatic score improvements after being coached. Case 1, for instance, shows a total increase of 200 points when June 1995 SAT I scores are compared to scores earned six months earlier, prior to

coaching. (The computed gains—a total of 50 points—are less impressive, however, when earlier PSAT/NMSQT scores are used as the precoaching baseline.)

- Not all coached students exhibit dramatic improvements. Some (Case 3, for example) show little or no improvement after participating in coaching courses. Still others (Cases 2, 4, and 5) seem to lose ground.
- Some *uncoached* students also display large increases upon retesting, even with little intervening preparation. Cases 6 and 8, for instance, each show a total increase of 230 points upon retesting (either 5 or 10 months later). These test takers reported that they prepared by reading the College Board's test familiarization, *Taking the SAT I: Reasoning Test*, and either reviewing math and English course work or using other test preparation materials.

In summary, the cases discussed above illustrate a variety of possible outcomes with respect to the effects of coaching (or the lack of it). It is probable, however, that the experiences of some examinees (Case 1, for instance) exert an undue influence over other test takers' perceptions of the usefulness of coaching. This conjecture is consistent with the tendency of decision makers to weigh confirmatory evidence more heavily than equally relevant disconfirmatory evidence (Wason and Johnson-Laird, 1972). Of all the cases we have selectively marshaled here, the one that may stand out is the single instance illustrating a substantial improvement.

Score Changes[2]

As noted above, the case histories we chose to make our point—that examining individual cases is at best insufficient and probably misleading—were by no means randomly selected, or even remotely representative of the study sample. Examining the test score changes of all test takers in our sample who took the SAT I and either the PSAT/NMSQT or the SAT I previously revealed that

- 12 percent of 427 coached examinees improved their SAT I verbal scores by 100 points or more upon retesting (mean gain = 29, *sd* = 59)
- 16 percent made equally large increases on the SAT I math exam (*m* = 40, *sd* = 58)

In contrast,

- 8 percent of 2,773 uncoached test takers improved by 100 or more points when retaking the verbal test (*m* = 21, *sd* = 52)
- 8 percent also improved by this much on the math test (*m* = 22, *sd* = 51)

But, for 36 percent of coached (and 38 percent of uncoached) examinees, SAT I verbal scores either decreased or remained exactly the same upon retesting. For the math portion, 28 percent of coached (and 37 percent of uncoached) examinees either made no improvement or else decreased upon retesting. With these data, the picture becomes somewhat clearer: Coached students were somewhat more likely than their uncoached counterparts to exhibit large score increases on both portions of the SAT I. However, examinees in both groups were much more likely to show *no increases* (or *decreases*) than they were to make large increases.[3]

TABLE 2

Mean Pre-SAT I, Post-SAT I, and Gain Scores for All Coached and Uncoached Examinees

	Test		
Group	***Pre***	***Post***	***Gain***
	Verbal		
Coached (*n* = 427)	500 (92)	529 (97)	29 (59)
Uncoached (*n* = 2733)	506 (101)	527 (101)	21 (52)
	Math		
Coached (*n* = 427)	521 (100)	561 (100)	40 (58)
Uncoached (*n* = 2733)	505 (101)	527 (101)	22 (50)

Note: Standard deviations are in parentheses.

In short, it is inadequate to estimate the effects of coaching on the basis of simple score changes for coached students only. However, though more defensible, even *comparisons* of score changes for coached and uncoached students, as shown in Table 2, are not entirely satisfactory either. As we will show next, coached and uncoached students differ with respect to a variety of characteristics, any one of which can contribute to biased estimates of the effects of coaching. To

2. For what we will refer to as raw score changes, we have used earlier SAT I scores, if available, as a baseline. If not, we have used PSAT/NMSQT scores, to which we have added a zero as the final digit in order to make them comparable to SAT I scores. This technique has also been used in several of the other analyses that employ precoaching test scores.

3. By comparison, the 419,000 students who tested as juniors in the spring of 1996 and again as seniors in the fall of 1996 improved their verbal scores by an average of about 12 points and their math scores by about 16 points. About 4 percent improved their verbal or math scores by 100 points or more (College Board, 1997).

TABLE 3

Demographic and Background Characteristics of Coached and Uncoached Examinees

Characteristic[1]	*Coached*	*Uncoached*	χ^2[2]
Female (%)	59	59	0.0
Ethnicity (%)			
American Indian	1	1	
Asian American	21	8	
African American	11	9	
Mexican American	3	4	89.7*
Puerto Rican	0	1	
Other Hispanic	3	3	
White	58	72	
Other	4	3	
Best Language (% English)	95	98	18.2*
Father's Education (%)			
High school or less	13	26	
Some college, A.A., B.A.	40	49	116.2*
Some graduate school or degree	47	25	
Mother's Education (%)			
High school or less	18	31	
Some college, A.A., B.A.	49	51	89.6*
Some graduate school or degree	34	18	
Parents' Income (%)			
Less than $40,000	23	39	
$40,000 - $80,000	34	43	156.0*
More than $80,000	43	18	
GPA (%)			
A– to A+	46	41	
B– to B+	45	48	25.2*
C+ or lower	9	11	
Degree Goal (%)			
Bachelor's or less	15	25	
Master's or higher	69	55	36.2*
Undecided	16	20	
Took PSAT/NMSQT (%) (as indicated on SDQ)	92	84	23.0*
Years Studied Various Subjects (Mean)			*F*
Arts and music	1.9	1.8	2.1
English	3.8	3.7	0.2
Foreign languages[2]	3.1	2.8	54.4*
Mathematics[3]	3.7	3.6	13.9*
Natural sciences[4]	3.4	3.3	7.6*
Social science/history[5]	3.4	3.3	3.2
Mean SAT Scores at First-Choice Colleges			
SAT-V	552	525	4.3*
SAT-M	566	533	3.2*
Mean PSAT/NMSQT Scores *(sds)*			
Verbal	51.0 (9.1)	50.7 (10.0)	0.6
Math	51.9 (10.1)	49.7 (10.0)	15.7*
Pre-SAT Score Means *(sds)*			
Verbal	487 (97)	512 (101)	11.5*
Math	522 (103)	518 (97)	0.2

1. *n*s range from 388 to 520 for coached students and from 2396 to 3556 for uncoached students. For pre-SAT scores, however, the *ns* were 198 and 1496 for coached and uncoached examinees, respectively.

2. Coached students were significantly more likely ($p < .05$ or higher) than uncoached students to have studied Chinese, Greek, Hebrew, Italian, Japanese, Korean, Latin, Russian, Spanish, but *not* French.

3. Coached students were significantly more likely ($p < .05$ or higher) to have studied trigonometry, pre-calculus, calculus, and computer math, but not algebra or geometry.

4. Coached students were significantly more likely ($p < .05$ or higher) to have studied chemistry, physics, and other science courses, but not biology or geology/earth/space science.

5. Coached students were significantly more likely ($p < .05$ or higher) to have studied U.S. history, world history, European history, ancient history, anthropology, economics, geography, psychology, sociology, and other social science/history, but *not* government/civics.

*$p < .01$

reiterate, if coached students differ from their uncoached counterparts on characteristics that relate to SAT I performance, then estimates will be inaccurate—either too high or too low.

Who Seeks Coaching?

Tables 3 and 4 reveal the many ways in which all coached and uncoached test takers in the study sample differed from one another. In short, according to Table 3, coached and uncoached students differed with respect to:

- ethnicity (coached students were more likely to be Asian American)
- best language (coached students' best language was slightly less likely to be English)
- parents' education (coached students' parents had more formal education)
- parents' income (coached students came from more affluent families)
- grades (coached students had slightly higher grades in high school)
- degree goals (coached students had slightly higher aspirations)
- previous test taking (coached students were more likely to have taken the PSAT/NMSQT)
- course-taking history (coached students had taken slightly more years of foreign language, mathematics, and science)
- previous test scores (coached students had slightly

TABLE 4

Test Preparation and Test Reactions of Coached and Uncoached Examinees

Characteristic	*Coached*	*Uncoached*	χ^2
Test Preparation (%)			
Read *Taking the SAT*	57	58	0.4
Tried sample test	51	51	0.1
Got College Board's *Real SATs*	21	9	81.2*
Got College Board's *Intro. the New SAT*	9	6	11.7*
Got College Board's video *Inside the SAT I*	2	2	0.3
Got other test prep books	62	28	239.9*
Got special SAT prep in class	39	32	9.3*
Attended special prep given by school	19	18	0.5
Tutored privately	15	5	75.0*
Used test prep software	26	18	18.8*
Used study aids	49	21	186.5*
Accessed test prep online	2	1	1.8
Used videos or related resources	2	2	0.0
Reviewed material from math courses on own	35	39	3.6
Reviewed material from English courses on own	30	33	1.9
Previously took the PSAT/NMSQT	88	80	18.3*
Previously took the SAT I	73	54	73.2*
Other	16	12	7.3*
Perception of Most Recent Previous SAT or PSAT/NMSQT Scores			
Pretty good estimates of my abilities	20	32	
Somewhat too low compared with my abilities	54	50	45.7*
Much too low compared with my abilities	26	18	
Nervousness Taking Most Recent SAT I			
Extremely nervous	13	8	
Very nervous	22	15	
Somewhat nervous	31	29	34.9*
Slightly nervous	20	26	
Not at all nervous	13	21	
Importance of Good SAT Scores			
Extremely important	52	40	
Very important	37	41	
Somewhat important	10	17	29.8*
Slightly important	1	2	
Not at all important	0	1	

Note: For coached examinees, *n*s ranged from 397 (for perception of most recent previous scores) to 534 (for test preparation). For uncoached examinees, comparable *n*s ranged from 2782 to 3733.

*$p < .01$

higher PSAT/NMSQT math scores and slightly lower precoaching SAT I verbal scores)

- college choices (coached students listed as their "first choices," colleges whose applicants had higher mean SAT I scores)

Also, as Table 4 reveals, in addition to attending coaching programs, coached students were more likely than uncoached students to have prepared for the SAT I in a variety of other ways. They were about twice as likely to have used study aids and to have obtained the College Board's book of practice tests, *Real SATs*, as well as other test preparation books. They were also slightly more likely to have tested previously and to have used a number of other test preparation resources.

Coached students also reported more often than uncoached students that they regarded their most recent previous PSAT/NMSQT or SAT I scores (i.e., those earned before the test they had just taken) as serious underestimates ("much too low") of their abilities. Uncoached students were more likely to report that their

TABLE 5

Effects of Coaching (All Programs) Based on Alternative Models of Analysis

Analysis Model	n *Coached*	*SAT I Verbal*		*SAT I Math*	
		Mean Effect	*Standard Error*	*Mean Effect*	*Standard Error*
Propensity matching model (PMM)	233	6	5	15**	4
Instrumental variable selection model (IVSM)	235	6	4	16**	3
Analysis of covariance (ANCOVA)	235	6	4	18**	3
Comparison of raw changes (RC)	427	8**	3	18**	3
Repeated measures (RM)	427	8**	3	18**	3
Belson model	469	8	9	26**	9
Heckman model†	237	12**	4	13*	6

Note: The *ns* for the uncoached comparison groups were 309 for the PMM, 1659 for IVSM, ANCOVA, and Heckman, 2733 for RC and RM, and 3494 for Belson. The total group analyzed here included 22 to 34 examinees (for the various models) who attended college- or university-based coaching programs. Because of the small number, these examinees were not included in subsequent analyses of major programs.

*$p < .05$, **$p < .01$

† The Heckman model tends to be quite sensitive to what variables are included in both the selection and the structural equations. Part of this sensitivity is due to the high colinearities between the estimates of the selection effect (lambda) and the dummy variable z. The estimates reported in the table were based on the *same* selection and structural equation as used in the instrumental variable approach. Other Heckman models characterized by different specifications yielded somewhat different estimates. The estimates of the coaching effects for mathematics varied from 13 to 18 while the corresponding estimates for verbal were 5 to 12.

TABLE 6

Effects of Coaching by Major Programs

Program/Analyses Model	n *Coached*	*SAT I Verbal*		*SAT I Math*	
		Mean Effect	*Standard Error*	*Mean Effect*	*Standard Error*
Company A					
ANCOVA	42	12	8	13	8
IVSM	42	13	8	11	8
Raw change	76	14*	7	5	8
Repeated measures	76	14*	6	5	6
PMM	41	14	9	12	8
Belson	82	15	17	17	21
Heckman	42	19*	8	9	8
Company B					
PMM	55	5	8	32**	7
IVSM	56	7	7	33**	7
ANCOVA	56	7	7	35**	7
Raw change	106	8	6	33**	6
Repeated measures	106	8	5	33**	5
Belson	113	9	17	38	25
Heckman	56	14*	7	31**	8
Other programs					
IVSM	115	–1	5	11*	5
PMM	115	0	6	9	5
ANCOVA	115	0	5	13**	5
Raw change	213	4	4	15**	4
Repeated measures	213	4	4	15**	4
Belson	240	4	16	23	13
Heckman	115	5	5	8	8

Note: The *n*s for the uncoached comparison groups were 309 for the PMM; 1659 for IVSM, ANCOVA, and Heckman; 2733 for RC and RM; and 3494 for Belson.

*$p < .05$, **$p < .01$

earlier scores were "pretty good" estimates. Finally, coached students reported being somewhat more nervous about taking the SAT, and they tended to place more importance on getting good scores than did their uncoached classmates.

Although all of the differences mentioned above are statistically significant, they are, with few exceptions, "small" (or "small" to "medium") in Cohen's (1988) terms. The only difference that can be considered as "medium" to "large" concerns the greater likelihood of coached than of uncoached examinees to obtain other books on test preparation.

Estimates of Coaching Effects

Table 5 provides estimates, for each analytic model, of "coaching effects" for all coaching programs combined. As is clear, the estimates do vary somewhat according to method of analysis, both with respect to their size and their precision. Some of this variation results, undoubtedly, not only from the different methods of analysis, but also from the use of slightly different samples and sets of variables for each method. Nonetheless, the effect estimates for SAT I verbal scores are remarkably consistent, ranging from 6 to 12 points; the range of effects for math scores is 13 to 26 points.

As pointed out earlier, the various models were expected to yield somewhat different results. For the repeated measures model, estimates are adjusted for between-group differences in initial status on only pretest scores. These estimates, therefore, may contain more bias than other estimates. The assumption underlying the repeated measures model is that between-group differences in pretest scores completely capture any between-group differences on all other unmeasured variables that are related to self-selection.

For the ANCOVA model, the most important covariates in the estimation model (for SAT I math) were earlier SAT I math scores, math grades, GPA, and the difference between the earlier SAT I math score and the mean SAT I math score at the examinee's first choice college. The important covariates in the verbal equation were earlier SAT I verbal scores, the difference between this score and the mean score at the first choice college, and grades in social science courses. The inclusion of additional control variables in the ANCOVA model, as compared to the repeated measures model, led to only a slight reduction in the estimated coaching effects for SAT I verbal scores, but no change for math scores.

For SAT I verbal scores, the IVSM yielded overall estimates that were virtually identical to those obtained for the ANCOVA model, suggesting that there was little correlated error between the selection model equation used in the first stage and the coaching effects equation used in the second. This in turn suggests that the potential omission of relevant unmeasured variables may not have been a major problem here. Also, the IVSM math estimates were only slightly lower than the ANCOVA estimates. This suggests the possibility of a slight correlated error across the selection equation and the coaching effects equation, which was adjusted by the IVSM procedure but not by the single-equation ANCOVA model. The reduction, however, was relatively trivial, as the two methods produced essentially the same results.

The propensity matching model yielded overall SAT I verbal effect estimates that are virtually the same as those computed from the ANCOVA and IVSM analyses. The estimates for SAT I math scores are only slightly lower than the ANCOVA and IVSM estimates. Using a much larger portion of the cases, the Belson model yielded verbal score effect estimates that are very similar to the ANCOVA and IVSM results; the estimates for math score effects, however, are the largest of any that were computed.

In summary, if two outlier estimates are discounted (i.e., the Belson model estimate for math scores and the Heckman estimate for verbal scores), then, on average, coaching seems to affect SAT I verbal scores by about 6–8 points, and SAT I math scores by about 13–18 points (or about twice as much as for verbal scores). By commonly used standards (Cohen, 1988), these effects can be regarded as small.

Table 6 shows a breakdown of results by major coaching programs.[4] This analysis serves mainly to illustrate the consistencies and differences among methods of estimation. For example, for each offerer, the Belson estimates tend to be larger than others, especially for math scores.

Differential Effects by Examinee Subgroups

As stated above, one of the advantages of the Belson model is that it enabled us to obtain estimates of coaching effects according to examinee background

4. Because the effects of coaching on SAT I performance are small for both of the major programs included in the study, and because sample sizes did not permit separate estimates for the variety of the other smaller companies, we have not identified the companies in our analyses. The hope is that this will discourage the misuse of our findings in ways that might provide a business advantage to any one firm over the variety of other test preparation programs and services, whose effects we were unable to estimate.

characteristics. Correlations of the individual effect sizes (residuals) with examinee characteristics revealed statistically significant, though slight, relations with several variables. Effects for SAT I math correlated negatively with PSAT/NMSQT math scores (r = .12, $p < .05$), suggesting that initially lower-scoring examinees may have benefited slightly more from math coaching than did their higher-scoring counterparts. Math coaching effects also correlated positively with high school grades in both English (r = .14, $p < .01$) and math (r = .12, $p < .05$) and with mean SAT I scores of other test takers who applied to the student's first-choice college (r = .13, $p < .05$). Effects for SAT I verbal correlated with number of years of English taken (r = .13, $p < .01$), with grades in English (r = .16, $p < .01$), with English best language (r = .11, $p < .05$), and with both mother's and father's education (r = .12 for each, $p < .05$). Thus, there is some indication in our data that some test takers may have benefited slightly more than others from coaching. For instance, students who had obtained good grades in their high school courses seemed to benefit slightly more than students who received lower grades. It seems plausible that the same traits that enabled students to get good grades also served them well in coaching classes. In any case, however, the size of any differential effects appears to be quite small.

Appendix B gives, for selected methods of analysis, estimates of coaching effects by gender, ethnicity, and initial score level. As is clear, effects do not vary substantially according to these classifications. The only exception is that, as mentioned above, effects on math scores appear to be greater for examinees who scored low initially.

Discussion

The major limitation of the study described here is its observational (or nonexperimental) nature: There was no random assignment to treatments, the most effective procedure for controlling differences between treatment and control groups. In addition, although our initial survey sample was representative of all SAT I takers, examinees on whom our analyses are based are less so, as more than a third of the initial sample chose not to respond. The major (known) effect of this nonresponse was to render our study sample somewhat more able on average (in terms of SAT I scores) than the population of SAT I takers. Thus, there is slightly more uncertainty in our estimates of coaching effects than our analyses may suggest. Despite these shortcomings, we believe that our controls were relatively effective in accounting for most of the relevant preexisting differences between coached and uncoached test takers, and that, at least for the responding sample, our estimates are reasonably defensible.

The major strength of the study, we believe, is its use of "real" SAT scores, i.e., scores from actual operational administrations of the PSAT/NMSQT and SAT I. To generate precoaching baseline data, other coaching studies have sometimes administered special editions ("retired" or publicly disclosed forms) of the SAT I under nonoperational conditions; still other studies have administered "SAT-like" tests that, although yielding scores corresponding to the SAT scale, were not equivalent to the SAT I. The motivation of test takers for these precoaching examinations has been called into question, and in at least some studies, examinees appear not to have been as motivated as if the test had actually counted (Messick, 1980).

From the outset, we fully expected that, because they were based on different sets of variables, on different samples of examinees, and on different analytical models, the alternative analyses described above would lead to somewhat disparate estimates of the effects of coaching. A willingness to entertain multiple estimates is consistent, we believe, with Tukey's (1996) advice to be open to more than a single answer, depending on the assumptions underlying the analyses, and with Rosenbaum's (1995) strategy of employing multiple control groups. Our expectation, however, was that *all* of the estimates generated in this study would be more defensible (and probably substantially lower) than the raw-gain-score "effects" based on only coached students as reported by coaching companies.

Our expectations were confirmed. The various analytical models did yield estimates that were not entirely consistent, though estimates were generally far less variable than we had anticipated. As noted, there were several likely reasons for the variation: the models rested on different assumptions, employed alternative treatments of missing data, used slightly different samples, and made adjustments on the basis of different sets of control variables. Nonetheless, despite the slight variation among the alternative estimates, two findings are clear:

1. There is an effect of coaching, and as with the previous version of the SAT, the effect is larger for the math than for the verbal part of the exam.
2. The variability among the estimates computed here—21 to 34 points over all programs for verbal and math scores combined—is far less than the discrepancy between these estimates and the claims made by the Kaplan and Princeton Review programs, for example.

What accounts for the disparity between our computations and the assertions of the coaching companies? Undoubtedly, the single most significant factor is the way in which "effects" are calculated by coaching companies. Score *changes* made by coached students—the most often reported information—are simply not effects at all. As asserted earlier, at the very least, some comparison is needed with the scores of uncoached test takers, who also often improve upon retesting. According to Cook and Campbell (1979), the simple one-group pre/post design, the basis for most claims about coaching, is likely to be adequate only in very rare circumstances. Studies of coaching's effect on SAT I scores is not one of these circumstances: Few if any of several plausible threats to the interpretation of study results (history, maturation, selection, and statistical regression, for example) can be discounted by this design.

There must also, however, be other reasons for the discrepancies between the claims of coaching schools and the results of more rigorous analyses. Even our analysis of simple change scores for coached students does not agree with claims made by coaching schools: We found far fewer large gains by coached students than has been suggested in the advertisements of coaching schools. We can only speculate on the reasons. One plausible explanation is that coached examinees who register large score increases are more likely to respond to surveys by coaching schools than are examinees who do not exhibit such large increases. This selective reporting of outcomes is likely to be a significant factor in the higher effect estimates given by coaching companies. Based on survey data, our results are also subject to nonresponse bias. However, we have no strong reason to believe that nonresponse bias skewed our estimates in the opposite direction.

The confidence we place in our estimates stems in part from the more rigorous methods of data collection and analysis that we have used. It also results from the correspondence of our estimates with the results of several previous meta-analyses of the effects of coaching for an earlier version of the SAT (Messick and Jungeblut, 1981; DerSimonian and Laird, 1983; Kulik, Bangert-Drowns, and Kulik, 1984; Becker, 1990). Although the test has changed, the revision is not, most observers would probably agree, radically different from its predecessor and, as suggested earlier, the putative result of the modifications is a less coachable test. Thus, the earlier research still provides a basis for comparison that is at least somewhat relevant.

In particular, our results are extremely similar to those obtained by Becker, whose analysis is arguably the most comprehensive of the several available quantitative summaries of coaching studies. Combining the results from a total of 48 separate studies, Becker concluded that, on average, coaching increased SAT verbal scores by about 9 points and SAT math scores by about 19—less if only the best designed studies are considered. Our estimates (medians for all analyses and over all programs) were 8 points for verbal scores and 18 for math. In addition, the estimates computed here for the two major commercial test preparation companies are generally consistent with the results of studies that have provided estimates for these companies (Powers, 1993, Table 2).

To put the benefits of coaching into perspective, a potential buyer might consider the following. The largest effect that we noted in any of our analyses was about 33 points for SAT I math for one of the coaching programs. This effect is equivalent to about three or four additional questions correct on the 60-question math portion of the SAT I. Assuming that a coached student attending a major program spends nearly 40 hours in classroom instruction and perhaps another 10–20 hours completing homework assignments (and that approximately half of this time is devoted to the math portion of the test), the benefit is approximately one additional question correct for every eight or so hours of effort. Making the same assumptions for the verbal portion of the test and basing our calculations on a 10-point effect, the result is about one additional question correct on the 78-question verbal part of the test for every 25–30 hours of effort.

In conclusion, given (a) DerSimonian and Laird's observation (i.e., that studies employing rigorous designs have yielded estimates that are only 20 to 25 percent as large as those obtained with less defensible methods), (b) the current assertions of major coaching companies, (c) the results of earlier studies of the effects of coaching for the SAT, and (d) the results of the analyses reported here, our conclusions are that:

- Coaching companies' current estimates of the effects of coaching for the SAT I are much too high; and
- the revised SAT is no more coachable than its predecessor.

References

Anastasi, A. (1981). "Coaching, test sophistication, and developed abilities."*American Psychologist,* 36: 1086–93.

Barnow, B., G. Cain, and A. Goldberger. (1980). "Issues in the analysis of selection bias." In *Evaluation Studies* (Vol. 5), E. W. Stromsdorfer and G. Farkus (eds.). Beverly Hills, Calif.: Sage Publications.

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

Belson, W. A. (1956). "A technique for studying the effects of a television broadcast." *Applied Statistics* 5: 195–202.

Cochran, W. G. (1969). "The use of covariance in observational studies." *Applied Statistics* 18: 270–75.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Second edition). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cole, N. (1982). "The implications of coaching for ability testing." In *Ability testing: Uses, consequences, and controversies part II: Documentation sections,* A. Wigdor and W. R. Garner (eds.). Washington, D.C.: National Academy Press.

College Board. (Sept. 1994). *Coaching and the new SAT: Common sense about how to prepare* (Special Report No. 6). New York: College Entrance Examination Board.

College Board. (1997). *Handbook for the SAT® Program 1997-98.* New York: College Entrance Examination Board.

Cook, T. D., and D. T. Campbell (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

DerSimonian, R., and N. M. Laird. (1983). "Evaluating the effect of coaching on SAT scores: A meta-analysis." *Harvard Educational Review* 53: 1–5.

Greene, W. H. (1997). *Econometric Analysis.* Upper Saddle River, N.J.: Prentice Hall.

Greene, W. H. (1992). *LIMDEP version 6.0: User's manual and reference guide.* Bellport, NY: Econometric Software.

Greene, W. H. (1981). "Sample selection as a specification error: Comment." *Econometrica* 49: 795–98.

Heckman, J. (1979). "Sample bias as a specification error." *Econometrica* 47: 153–61.

Kulik, J. A., R. L. Bangert-Drowns, and C. C. Kulik. (1984). "Effectiveness of coaching for aptitude tests." *Psychological Bulletin* 95: 179–88.

Messick, S. (1980). *The effectiveness of coaching for the SAT: Review and reanalysis of research from the fifties to the FTC.* Princeton, N.J.: Educational Testing Service.

Messick, S. (1982). "Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice." *Educational Psychologist* 17: 67–91.

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

Murnane, J., S. Newstead, and R. J. Olsen. (1985). "Comparing public and private schools: The puzzling role of selectivity bias." *Journal of Business & Economic Statistics* 3: 23–35.

Olsen, R. J. (1980). "A least squares correction for selectivity bias." *Econometrica* 48: 1815–20.

Powers, D. E. (1993). "Coaching for the SAT: A summary of the summaries and an update." *Educational Measurement: Issues and Practice* 12 (2): 24–39.

Powers, D. E. (1985). "Effects of coaching on GRE aptitude test scores." *Journal of Educational Measurement* 22 (2): 121–36.

Powers, D. E. (1998). *Preparing for the SAT I: Reasoning Test—An Update* (College Board Report No. 98-5). New York: College Entrance Examination Board.

Rosenbaum, P. R. (1995). *Observational studies.* New York: Springer-Verlag.

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

Sesnowitz, M., K. L. Bernhardt, and D. M. Knain. (1982). "An analysis of the impact of commercial test preparation courses on SAT scores." *American Educational Research Journal,* 19: 429–11.

Stroud, T. W. F. (1980). *Reanalysis of the Federal Trade Commission study of commercial coaching for the SAT* (ETS RR-80-10). Princeton, N.J.: Educational Testing Service.

Tukey, J. (1996, October 9). "How should we frame our questions?" Presentation at Educational Testing Service, Princeton, N.J.

Wason, P. C., and P. C. Johnson-Laird. (1972). *Psychology of reasoning: Structure and content.* Cambridge, Mass.: Harvard University Press.

Appendix A

Heckman Selection Model

The Heckman selection model is a two-equation model. The first equation attempts to model the selection process, that is, here we are modeling coaching school participation as

$$z_i = r' w_i + e_{1i} \qquad (1)$$

where (1) is probit equation with w an unobserved and/or incompletely specified vector representing ability and motivation and z is an estimate of the probability of being coached.

The second or structural equation is

$$y_i = B' x_i + \delta z_i + e_{2i} \qquad (2)$$

where y = posttest score, and where x is a vector of incompletely measured and typically incompletely specified covariates, many or all of which are shared with w. z_i is a "dummy" variable indicating whether an individual is in the treatment or control group. The problem here is that e_1 and e_2 are likely to be correlated since the variables in w in (1) are either incompletely specified or measured with error or both. If the variables in w only partially measure initial true ability and/or motivation and what is left out correlates positively with performance on y (say coaching) gains then we can expect e_1 and e_2 to be correlated. If students were assigned entirely on the basis of the observed variable w and w was included in x, then δ would yield an unbiased estimate of the coaching effect. Similarly, random assignment to "z" status eliminates any bias in the estimate of the coaching effect since $r = 0$ in (1) and the expected value of the covariance of z and w will be *0*. The other unlikely situation that yields an unbiased estimate of the coaching effect is if the conditional covariance of zw is *0* conditioning on x, i.e., $(zw \mid x) = 0$.

One way to look at the Heckman selection model is as a special case of a truncated distribution. If those students who self-select to go to coaching school are of initially higher ability and/or motivation, then one can think of a shift of their subpopulation mean to the right of the mean of the total sample, i.e., coached and uncoached. Because of the symmetry of the assumed normal distribution of ability and motivation, the same argument applies if one thinks of the uncoached sample as a subsample from the total population who have suffered "creaming" and thus have a truncation in their upper tail.

To the extent that the cov $(zw \mid x) \neq 0$ and "creaming" takes place, the initial differences in ability and motivation will be attributed to the coaching effect. Greene's (1997) derivation of the treatment/control variation of the Heckman selection model is based on the truncated normal and is sketched out below.

If y and z have a bivariate normal distribution with correlation p_{yz} and we are interested in the marginal distribution of y given that z exceeds a particular value assuming that $p_{yz} > 0$, then the truncation of z (creaming) should push the distribution of y to the right. That is, the joint density of y and z is

$$p(y, z \mid z > a) = \frac{f(y,z)}{prob(z > a)} \qquad (3)$$

integrating z out of (3) Greene (1997) shows that

$$E[y \mid z > a] = u_y + p\sigma_y \lambda(a_z) \qquad (4)$$

where $a_z = (a - u_z)/\sigma_z$ and $\lambda = \phi(a_z)/[1 - \Phi(a_z)]$ where ϕ = normal density and Φ = the normal cumulative distribution function. If the truncation is $z < a$, then $\lambda(a_z) = -\phi(a_z)/\Phi(a_z)$. Lambda ($\lambda$) is the so-called inverse mills ratio.

In the Heckman variation known as the treatment selection model, the probit equation is used to estimate z_i^* not z in equation (1). That is $z_i^* = r' w_i + e_{1i}$ where z^* is an imperfectly estimated probability of going to coaching in the so-called selection equation. Then $z = 1$ if $z_i^* > 0$, otherwise $z = 0$ (no coaching).

$$\begin{aligned} E(y_i \mid z = 1) &= B' x_i + \delta + E(e_2 \mid z = 1) \\ &= B' x_i + \delta + p\sigma_{e2}\left[\frac{\phi(\gamma' w_i)}{\Phi(\gamma' w_i)}\right] \\ &= B' x_i + \delta + p\sigma_{e2}\lambda(\gamma' w) \end{aligned} \qquad (5)$$

where $p = correlation\ (e_1\ e_2)$, and

$$E(y_i \mid z = 0) = B' x_i + p\sigma_{e2}\left[\frac{-\phi(\gamma' w_i)}{1 - \Phi(\gamma' w_i)}\right] \qquad (6)$$

The expected difference in performance on y for the coached and uncoached groups is

$$E(y_i \mid z = 1) - E(y_i \mid z = 0) = \delta + p\sigma_{e2}\left[\frac{\phi_i}{\Phi_i(1 - \Phi)_i}\right] \qquad (7)$$

The quantity to the right in equation (7) will be inappropriately assigned to δ, the coaching effect, if λ is not explicitly included in equation (2). If $p>0$ in equation (7) then one can assume some "creaming" took place. Conversely, if $p=0$, then the Heckman "treatment" model should yield similar results to the standard ANCOVA model shown in equation (2). If $p>0$ then the errors are positively correlated and the coaching effect is overestimated. Theoretically, the solution then requires two steps. First, estimate equation (1) using probit regression and obtain maximum likelihood estimates of γ and then compute λ for the coached $\frac{\phi(\gamma' w_i)}{\Phi(\gamma' w_i)}$ and for the uncoached $\lambda = \frac{-\phi(\gamma' w_i)}{1+\Phi(\gamma' w_i)}$ and insert λ in equation (2) and perform ordinary least squared (ols) regression. Unfortunately, due to colinearities among the regressors, the ols estimates are not efficient. Computer programs such as LIMDEP (1997) use ols as starting values for a final maximum likelihood solution. The regression weight δ, associated with the "dummy" z in the following equation, will give an unbiased estimate of the treatment effect

$$y_i = B' x_i + \delta z_i + m\lambda_i + e_3 \qquad (8)$$

and

$$m = \hat{\rho}_{e_1 e_2} \hat{\sigma}_{e_3}$$

and

$$\hat{\rho}^2_{e_1 e_2} = \frac{\hat{m}^2}{\hat{\sigma}^2_{e_3}}$$

Instrumental Variable Approach

A much simpler approach but not necessarily yielding the same estimates is the 2SLS approach using $\Phi(\gamma' w_i)$ as one of the instrumental variables in a 2SLS solution. This will purge e_2 of its correlation with e_1 from the probit equation. The regression coefficient (δ) associated with z_i in equation (2) will then give an unbiased estimate of the treatment (coaching effect).

Propensity Analysis

Propensity analysis simply uses equation (1) to estimate z_i and then uncoached students and coached students are matched based on the similarity of their "z" scores. The assumption here is that the sample of uncoached students is considerably larger than that of the coached students. Propensity matching also assumes that self-selection can be accurately modeled by the observed variables.

Appendix B

Estimates of Effects of Coaching by Subgroups

TABLE B.1

Effects of Coaching (All Programs) by Gender

			SAT I Verbal		SAT I Math	
Analysis Model	*Gender*	n *Coached*	*Mean Effect*	*Standard Error*	*Mean Effect*	*Standard Error*
Comparison of raw changes	F	239	9	4	16	4
	M	167	7	5	20	4
Belson	F	300	5	15	26	12
	M	200	14	16	24	13
IVSM	F	144	1	5	18	4
	M	98	13	6	12	5

Table B.2

Effects of Coaching (All Programs) by Ethnicity

			SAT I Verbal		SAT I Math	
Analysis Model	*Ethnicity*	n *Coached*	*Mean Effect*	*Standard Error*	*Mean Effect*	*Standard Error*
Comparison of raw charges	Asian	31	9	12	16	11
	White	275	7	4	19	4
	Other	81	5	6	14	7
Belson	Asian	96	6	21	35	23
	White	269	7	12	22	11
	Other	91	10	17	19	20

Note: "Other" designates all underrepresented minority examinees (American Indian, African American, Mexican, Puerto Rican, and other Hispanic). Sample sizes were deemed too small to provide estimates using other models.

TABLE B.3

Effects of Coaching (All Programs) by Initial Score Level

			SAT I Verbal			SAT I Math	
Analysis Model	*Initial Score Level*	n	*Mean Effect*	*Standard Error*	n	*Mean Effect*	*Standard Error*
Comparison of raw changes	<400	49	13	10	39	23	11
	400-600	315	6	3	285	20	4
	>600	34	20	10	54	6	7
Belson	<400	49	28	26	39	41	29
	400-600	319	10	14	299	30	12
	>600	59	17	25	89	20	21