SAT Coaching, Bias and Causal Inference

by

Derek Christian Briggs

B.A. Carleton College 1993
M.A. University of California, Berkeley, 1998

A dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Education

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Mark Wilson, Chair
Professor David Freedman
Professor Paul Holland
Professor David Stern

SAT Coaching, Bias and Causal Inference

by

Derek Christian Briggs

ABSTRACT

SAT Coaching, Bias and Causal Inference

by

Derek Christian Briggs

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

This study considers the extent to which unbiased causal inferences can be drawn about the effect of coaching on SAT performance. Following a review of the literature, I present the linear regression model and the Heckman Model as two statistical approaches that might be used control for bias in an estimated coaching effect. The assumptions necessary before an estimated effect can be given a causal interpretation are described in some detail. I estimate coaching effects for both sections of the SAT using data from the National Education Longitudinal Study of 1988 (NELS). There is some indication that the linear regression model successfully reduces bias due to omitted variables. It appears that commercial coaching programs have an effect of about 3 to 20 points on the verbal section of the SAT, and an effect of about 10 to 28 points on the math section of the SAT. These effects may be somewhat bigger or smaller if coaching is defined more broadly. There is some evidence that coaching is more effective for certain types of students. I demonstrate the sensitivity of the Heckman Model to the choice of variables included in the selection function. Small changes in the selection function are shown to have a big impact on estimated coaching effects.

ACKNOWLEDGMENTS

This dissertation benefited from the help of a very supportive committee. I thank the chair of my committee, Mark Wilson, for his mentoring throughout the course of my graduate training. Before I met Mark I didn't know an item from an instrument. David Stern has also been a valuable mentor for me; I've learned a lot about striving for clarity in working with him. Paul Holland first introduced me to the topic of causal inference. My discussions with Paul have done a great deal to influence my thinking about causation, educational research and good red wines. Special thanks to David Freedman for his patience in working with me to get chapter 2 on track.

Don Powers contributed useful comments on earlier versions of this study, and Don Rock helped me check my Heckman Model estimates produced in SPSS against those produced in STATA. I have had many engaging discussions about causal inference with my friend and colleague Ben Hansen. I am particularly grateful for Ben's encouragement while I was struggling with a moment of self-doubt in the spring of 2002.

I thank my father, Ray Briggs, for convincing me to take courses that developed my quantitative skills while at Carleton College. Little did I realize how important those skills would become for me! Though she may well think it is an inappropriate setting for the sentiment, I thank my mother, Sylvia Mauthner, for always prodding me to think critically and look at the world through a multicultural lens. I also acknowledge the support of Doris Shimabukuro, who was a second mother to me during my formative years.

I have saved the most important acknowledgment for last. This project has been the culmination of a marathon for someone who is by nature a sprinter. It has involved a degree of single-mindedness bordering on obsession, and it takes a special partner to put up with all that. I thank my wife and best friend, Whitney Pinion, for bearing with me. Without her companionship and support over the past five years, "I'd have become so unpleasant, I'd have gotten on my own nerves."

I dedicate this dissertation to Ralph Waniek, who once wrote to me "Keep pushing, you'll get there." Thanks Ralph, I did.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The SAT is a standardized test required for admission at almost all competitive four-year colleges in the United States.[1] The test has a math and verbal section, each scored on a scale that ranges from 200 to 800 with standard deviation of about 110 points. Each year about two million high school students take the SAT at a cost of about $25 each. Coaching for the SAT (and many other standardized tests) is a multimillion dollar industry. Companies such as Kaplan and The Princeton Review charge roughly $800 for 30-40 hours of instruction, and attribute to their programs average gains of 100-140 points on the combined math and verbal sections of the test. Private tutors, books, videos and computer software are also available, at a price, to help students prepare for the test. It has become widely accepted among the general public that coaching has a large effect on student scores. Yet most of the published research on the topic suggests that the combined coaching effect is fairly small, in the range of about 20 to 30 points.

Clear causal inferences about the effectiveness of coaching have proven elusive. Few studies of coaching have attempted randomized designs, and for the those that have the results remain equivocal for reasons I discuss in chapter 1. Instead, coaching effect estimates are generally based upon studies with observational designs. Imagine that one group of students takes part in a coaching program and another group does not. When

[1] As of 1994, the SAT became the SAT I. For the sake of consistency, the term SAT is used throughout generically to represent a multiple-choice test used for purposes of college admission. A historical description of the SAT in the context of its use in college admissions decisions is beyond the scope of this dissertation. For these details, cf Zwick, 2002; Lawrence et. al., 2001; Lemann, 1999.

average SAT scores for the two groups are compared, can differences between the groups be attributed to the program? This is the key question in making causal inferences. Treatment and control groups in observational studies are not randomly assigned. Thus, outcome differences between the groups may be explained by other characteristics on the two groups differ. Hence, an estimate of a causal effect usually suffers from bias[2], which can lead to incorrect inferences about the effectiveness of coaching.

A number of statistical methods have been used in observational settings to control for bias. There is a common thread running through all these approaches: it is the idea that an observational study can be considered as a randomized experiment, conditional on certain covariates. The approaches differ in the statistical assumptions they make and the methods they apply to the data. In this dissertation I will be analyzing two different methods of controlling for bias in the context of coaching for the SAT: the linear regression model and the Heckman Model.[3] I ask two fundamental research questions:

1. How can the linear regression model and the Heckman Model be used to make unbiased causal inferences in observational settings?
2. Using these models, what can be concluded about the effect of coaching on SAT performance?

[2] The term bias in this dissertation is defined in a statistical context (e.g. an estimated causal effect is biased), not an educational measurement context (e.g. the test items are biased against certain types of students).

[3] Three other popular approaches that are sometimes used in this context include the Propensity Matching Model (Rosenbaum & Rubin, 1983), two stage least squares (Greene 1993, 603-10), and structural equation modeling (Jöreskog & Sörbom, 1996).

The general idea presented here is that the effect of coaching can be modeled using two equations. One equation represents the process by which students decide whether or not they will be coached before taking the SAT. This is called the selection function. A key feature of the selection function is a latent random variable. Another equation represents the process by which exposure to coaching, conditional on certain covariates, has an effect on SAT performance. This is called the regression equation. A key feature of the regression equation is what is usually presented as the "error" term. The linear regression and Heckman Model approaches differ primarily in the assumptions they make about the relationship between the latent variable in the selection function, and the error term in the regression equation. If the latent variable and the error term are assumed to be independent, then the linear regression model may be used to estimate an unbiased effect for coaching from the regression equation. If the latent variable and the error term are allowed to be correlated, then the linear regression model will probably not generate an unbiased estimate of the coaching effect. Given certain strong distributional assumptions, the Heckman Model uses both the selection function and the regression equation in a two-step process to estimate an asymptotically consistent coaching effect.

In chapter 1, I review the literature on SAT coaching studies. My focus is on how each study went about estimating an effect for coaching. I describe a number of obstacles that have hindered the causal interpretation of estimated effects: unclear definitions of the term coaching, small and unrepresentative student samples, and numerous factors that have confounded comparisons between coached and uncoached students. I point out that randomized experiments, an ideal methodological design in scientific research, have not

had a successful history in the context of coaching studies. Observational designs are the more typical scenario. Bias typically clouds the interpretation of effects estimated from observational coaching studies.

In chapter 2, I discuss the two-equation behavioral model under which an estimated coaching effect has a causal interpretation. When certain assumptions are made, the linear regression or the Heckman Model can be used to estimate an unbiased effect of coaching. Each statistical approach makes different assumptions about the nature of bias, and the two-equation behavioral model is used to distinguish these assumptions. Under the linear regression model, bias is due to confounding from covariates omitted from the regression equation; under the Heckman Model, bias is due to confounding from omitted variables and self-selection. In the latter case, the idea is that subjects select themselves into coached and uncoached conditions as a function of at least one covariate that is latent. If selection bias exists in observational data, then the linear regression model is not equipped to deal with it. On the face of things, the Heckman Model is an attractive solution to the problem of bias in an estimated coaching effect. In theory, it accomplishes all the things we would want from the linear regression model, and more. Because it is a relatively new statistical approach, I describe in some detail the two-step process by which the Heckman Model—given its underlying assumptions—would estimate an asymptotically unbiased coaching effect. Criticism of the Heckman Model on theoretical grounds has often emphasized the potential sensitivity of the model to different covariate specifications. I discuss briefly why this might be a problem, setting the stage for an empirical analysis in chapter 4.

Chapter 3 introduces the data I use to estimate coaching effects with the linear regression and Heckman Model approaches. The National Education Longitudinal Study of 1988 (NELS), sponsored and maintained by the United States Department of Education, contains nationally representative information on a cohort of students who were high school seniors in 1991-92. Population weights and design effect corrections necessary for the NELS sampling design are introduced. I describe the NELS subsamples of interest, along with the data needed to evaluate the effectiveness of coaching: SAT scores, variables indicating how students prepared for the SAT, and variables associated with higher or lower SAT scores. These latter variables are analyzed relative to coaching status, which is defined as enrollment in a commercial test preparation course. I use the data to describe the characteristics of coached and uncoached students nationally.

In chapter 4, I estimate coaching effects for each section of the SAT using first the linear regression model, and then the Heckman Model. For the Heckman Model, I present five alternate specifications, each of which might prove compelling for a researcher hoping to estimate a causal effect for coaching. I show that the coaching effects estimated under the different specifications of the Heckman Model vary dramatically. Multicollinearity, reflected by large standard errors and unstable parameter estimates, is discussed as a potential problem. I compare the coaching effect estimates generated under the linear regression model and Heckman Model to effects estimated in previous observational coaching studies. According to the linear regression model, the

combined effect of coaching is about 30 points; according to the two most plausible specifications of the Heckman Model, the combined effect is about 60 points.

Both the linear regression and the Heckman Model approaches assume that it makes sense to estimate a single effect for coaching. In chapter 5, I consider the extent to which this constraint seems to hold up for the linear regression model. I test whether the estimated coaching effect is different for students coached in different ways, or for students with different demographic, academic and motivational characteristics. I also consider whether estimated coaching effects are consistent using a different subsample of students from the NELS data. The estimated coaching effect does depend on how coaching is defined, but not dramatically. There are some suggestive interactions between commercial coaching and certain student characteristics. Estimated commercial coaching effects are roughly consistent across different student subsamples in the NELS data.

In chapter 6, I summarize the important results of this dissertation. I return to the two research questions posed above, and offer some tentative answers. Finally, I conclude with some suggestions for further research on the topic of SAT coaching.

CHAPTER 1: REVIEW OF THE LITERATURE ON SAT COACHING

1.1 Coaching and the SAT

The SAT consists of two sections administered over two and a half hours, with items that measure, respectively, verbal and mathematical reasoning ability. The test is sponsored by the College Entrance Examination Board (CEEB), but developed and administered by the Educational Testing Service (ETS). Since its inception as a tool for college admissions in the late 1940s, the SAT has become one of the most widely known large-scale standardized tests in the United States. It has also become very controversial. Much of this controversy stems from the early association of the SAT with IQ testing. Initially, the SAT was devised as a test of aptitude, and its acronym—the Scholastic Aptitude Test—reflected this belief. Over time however, both the format of the test and the position of its developers has changed. Messick (1980) and Anastasi (1981) have suggested that standardized tests can be conceptualized as solely measuring either achievement or aptitude, and that the SAT falls somewhere in between these two poles. Messick wrote:

> The Scholastic Aptitude Test was developed as a measure of academic abilities, to be used toward the end of secondary school as a predictor of academic performance in college…The SAT was explicitly designed to differ from achievement tests in school subjects in the sense that its content is drawn from a wide variety of substantive areas, not tied to a particular course of study, curriculum or program. Moreover, it taps intellectual processes of comprehension and reasoning that may be influenced by experiences outside as well as inside the

> classroom…The specific item content on the SAT attempts to sample the sort of cognitive skills underlying college-level performance. (p. 7)

A key element in Messick's description of the SAT, and one which The College Board has maintained in subsequent descriptions of the test, is the notion that the SAT measures reasoning abilities that are developed gradually over the years of primary and secondary schooling that precede college. As such, SAT performance should be sensitive to long-term instruction (unlike an aptitude test), but insensitive to short-term instruction or cramming (unlike an achievement or aptitude test). The question of the effectiveness of SAT coaching is predicated on this latter aspect.

Multiple definitions of the term coaching are possible. In this review I define it broadly as short-term, systematic test preparation for a student or group of students that involves any or all of the following: content review, item drill and practice, and an emphasis on specific test-taking strategies and general testwiseness. Coaching varies in its setting, its mode of instruction and its duration. All forms of SAT coaching share one common characteristic: the presumption that students being coached will perform substantially better on the SAT than if they had not been coached. The debate over this presumption has been described as "the leading measurement dispute of our time" (Bond, 1989) because it gets to the heart of the fundamental psychometric concept of test validity.

Along these lines, a number of authors have decomposed the term coaching into different components (Pike, 1978; Anastasi, 1981; Cole, 1982; Messick, 1982; Bond,

1989). Messick describes three general ways that coaching could be expected to increase test scores: 1) improving the skills and abilities measured by the test, 2) reducing test-taking anxiety, and 3) teaching students to use test-taking strategies and tricks. Messick concluded that only coaching that emphasized this third approach posed a threat to test validity. Cole delineated similar components in greater detail. Under Cole's taxonomy, coaching which emphasizes testwiseness should not threaten test validity, as a well-constructed test is by definition insensitive to test-taking strategies and tricks. Cole suggested that the greatest threat to test validity is coaching that results in the instruction of test-specific content unrelated to the construct (i.e. reasoning) of interest. A critical distinction then, is whether coaching effects are transferable. Does a coaching effect of some amount of points on the SAT indicate a student with improved reasoning abilities ready for college study, or does it merely indicate a student who has “beaten” the test? To my knowledge, no coaching studies—including this one—have sought to answer this question.

SAT coaching studies have instead focused on the issue of what constitutes a "substantial" effect from coaching that is "short-term" in nature. The SAT is reported on an arbitrary scale ranging from 200 to 800 points per test section. It could just as easily be reported on a scale from 20 to 80, and in fact this is the scale upon which the PSAT, essentially a pre-test for the SAT, is reported. A 10 point effect on the SAT scale has the same interpretation as a 1 point effect on the PSAT scale. Because the SAT scale has no absolute meaning, the best way to interpret the size of an estimated coaching effect is relative to the standard deviation (SD) and standard error of measurement (SEM) of each

test section. While these numbers vary from year to year, the SD tends to be roughly 100-120 points, while the SEM is about 30 points. Hence a 10 point coaching effect on one section of the SAT is about 1/10 of a SD and 1/3 of the SEM, a relatively small effect. A coaching effect of 60 points, on the other hand, is 3/5 of a SD and two full SEMs, a relatively large effect. There is no established frame of reference to distinguish between short- and long-term coaching. One useful approach is to note that a single high school course meeting 45 minutes a day involves somewhere between 40 and 50 hours of instructional student contact time over the length of a semester. Coaching programs that involve this amount of student contact time in preparation for a single SAT test section would seem more reasonably classified as traditional instruction. Previous researchers have used 40-hours as a threshold between short- and long-term preparation (Jackson, 1980), but clearly there is ample gray in this distinction.

In what follows I provide a historical review of SAT coaching studies through the lens of causal inference. Coaching studies differ, sometimes dramatically, in their choices of experimental sample, coaching treatment and methodological design. Because of this, it is difficult to report any one estimate for the effect of coaching on either the verbal (SAT-V) or math (SAT-M) sections of the test. The effect estimated in a given study will depend on the answers to three questions:

1. What types of and how many students were coached and not coached?
2. What was the nature of the coaching program?
3. What was done to make coached and uncoached students comparable?

The first question concerns the choice of sample and its generalizability to the full population of students taking the SAT. It is equally important to note the size of the experimental samples. Larger samples lend themselves to more precise coaching effect estimates that are less subject to chance error. The larger the sample size, the greater the power of a test of statistical significance to reject the null hypothesis that the effect of coaching is 0 when it is in fact false.

Answering the second question makes clear that the coaching treatment can differ in many ways. A key issue in summarizing the results of coaching studies is determining whether it makes sense to estimate a common coaching effect when more than one coaching program is being evaluated. As a starting point, coaching programs can be categorized into three types: school-based, commercial and computer-based. While there may be some overlap between these categories, they are used as a first grouping criterion by which coaching studies will be summarized in this review. School-based coaching encompasses programs that are incorporated into a high school's curriculum, usually as an elective during a school day, or as an extra-curricular program outside of the school day. These programs tend to be the longest in duration, and here the distinction between coaching and content-based instruction is often the blurriest. Commercial coaching is unique in three important ways. First, commercial programs charge students a fee,

sometimes a very large one, for their services. Second, to attract students to such programs, coaching companies have made public claims about the effectiveness of their services. Third, commercial coaching tends to be more short-term in duration compared to coaching in school settings; the longest program reviewed here devoted about 25 hours of student contact time per test section. Studies of computer-based coaching programs are considered separately because they involve a novel mode of coaching delivery. An advantage of computer-based programs is that coaching delivery is not dependent on the skills of the human coach. On the other hand, because the coaching often proceeds under the student's control, it becomes more difficult in the context of a study to ensure that students receive equal amounts of treatment exposure. Computer-based coaching programs tend to be shortest in duration, often only four or five hours per SAT test section.

The question of comparing coached and uncoached students gets to the heart of this dissertation. In short, we wish to know the strength of the methodological design used to arrive at coaching effect estimates. From the standpoint of causal inference, if treatment and control groups of students are not roughly equivalent along all other characteristics correlated with the outcome of interest, a treatment effect estimate may be biased. Three general design approaches have been taken in coaching studies: uncontrolled, observational (also known as quasi-experimental), and randomized experimental. These will serve as a second major grouping criterion for the coaching studies evaluated in this review. The crossing of the two grouping criteria—coaching

type by methodological design—for all published and unpublished SAT coaching studies since 1953 are presented below in Figure 1-1.

Figure 1-1. Studies by Coaching Type and Design

Coaching Type	Methodological Design		
	Randomized	Observational	No Control
School-based	Roberts & Openheim (1966) **Evans & Pike (1978)** **Alderman & Powers (1980)** Shaw (1992)	**Dyer (1953)** **French (1955)** **Dear (1958)** **Kintisch (1979)** Johnson [SF site] (1984)[1] Burke (1986) Harvey (1988) Schroeder (1992) Wrinkle (1996)	**Pallone (1961)** Marron (1965) Johnson [Atlanta, New York sites] (1984)[1]
Commercial		**Frankel (1960)** **Whitla (1962)** FTC Study & reanalyses BRO (1978) BCP (1979) Rock (1980) Stroud (1980) **Sesnowitz, Bernhardt & Knain (1982)** **Whitla (1988)** Zuman (1988)[1] **Snedecor (1989)** **Smyth (1989)** **Smyth (1990)** **Powers & Rock (1999)** **Briggs (2001)**	**Kaplan (2001)**
Computer-based	Hopmeier (1984) Laschewer (1985) Curran (1988) **Holmes & Keffer (1995)** McClain (1999)		Coffin (1988)[1]

Notes:
Bold type represents studies published in refereed academic journals.
[1] Design intent of these studies (randomized experimental) compromised by substantial sample attrition.

In this review the estimated size of coaching effects from study to study is of less interest than *how* the effect was estimated, and the extent to which it may be biased because of differences between coached and uncoached students. I purposely do not take a meta-analytic approach. Meta-analysis (Glass, McGaw et al., 1981; Hedges & Olkin, 1985) is technique for synthesizing quantitative results across studies evaluating the effect of a treatment with samples drawn from a common underlying population. It has been used by several researchers in previous reviews of coaching studies (DerSimonian & Laird, 1983; Kulik, Bangert-Drowns et al., 1984; Becker, 1990). It is not used here because I believe it adds a layer of restrictive assumptions to coaching effect estimates that have themselves frequently been made through the use of strong statistical assumptions (Berk & Freedman, 2001). In particular, the changing nature of the SAT test-taking population over time and the lack of independence among coaching studies threaten the usefulness of inferences drawn from meta-analysis. Finally, meta-analysis may obscure the nuances of coaching studies, particularly in their methodological designs. This review follows the approach taken by Pike (1978), Messick (1980), Bond (1989) and Powers (1993) in their reviews of coaching studies: studies are summarized individually and synthesized according to characteristics they have in common. At the same time, characteristics that make each study unique are highlighted, with special emphasis given to the methodological approaches taken to control for bias in coaching effect estimates.

Thirty-two distinct SAT coaching studies were located for review. These studies were found by consulting the reference sections of previous SAT coaching reviews (Pike,

1978; Messick, 1980; Slack & Porter, 1980; Anastasi, 1981; Messick & Jungeblut, 1981; Cole, 1982; Messick, 1982; DerSimonian & Laird, 1983; Kulik, Bangert-Drowns et al., 1984; Bond, 1989; Becker, 1990; Powers, 1993) and searching through the internet, academic journal indices and the ProQuest index of dissertation abstracts using combinations of the keywords "SAT", "Coaching" and "Test Preparation." To be included in this review, coaching studies were filtered through three criteria: 1) The study, whether published or unpublished, existed and was available in written form; 2) the study involved a program expected to increase student scores on the SAT; and 3) the sample of students participating in the study were not yet enrolled in college. Four studies referenced in previous reviews (Lass, 1961; Coffman & Parry, 1967; Fraker, 1987; Wing, 1987) and four doctoral dissertations (Keefauver, 1976; Winokur, 1983; Davis, 1985; Warch, 1996) did not meet all three criteria.

In the next three sections—1.2, 1.3 and 1.4—I present a review of all SAT coaching studies by methodological design (uncontrolled, observational and random experimental). Within each design category, studies are grouped and described in chronological order with respect to the setting under which the coaching took place (school-based, commercial-based, computer-based). In section 1.5 I summarize the results of the review and present salient patterns. Finally, in the last section, 1.6, I discuss the extent to which causal inferences can be justifiably draw about the effect of coaching from both past and future studies.

1.2 Uncontrolled Studies

Coaching in School Settings

Pallone (1961) investigated whether a developmental reading course could be used to improve SAT-V performance. This was the first published SAT coaching study to suggest on an empirical basis the possibility of a sizable average improvement in SAT performance due to a systematic program of instruction. Pallone reported average SAT-V score gains of 98 points for about 20 students who had participated in a daily, six-week long course emphasizing reading skills and the analysis of verbal analogy problems. For a larger group of about 100 students participating in a longer version of the course spanning six months, Pallone reported an average gain of 109 points.[4]

Marron (1965) hoped to determine if exposure to full-time instruction at a military preparatory school with course content directly related to both the SAT and College Board achievement tests would have an effect on subsequent test performance as well as grades in a military academy. Marron found weighted average gains in SAT-V and SAT-M scores across 10 schools for coached students of 58 and 79 points. Marron also discovered that coached students performed worse upon entering the military academy than would have been expected based on their SAT scores.

[4] Messick (1980) pointed out that this reported gain in Pallone's study was ambiguous. According to Pallone's Table 3, the average gain for the longer course was 84 points. Pallone's text however, reports an average gain of 109.

It is very difficult to evaluate the reported average SAT score gains by Pallone and Marron as coaching effects in the absence of comparable control groups. The Pallone and Marron samples consisted of boys at private college-preparatory schools, many of whom had already graduated from high school, and all of whom had been admitted to the prep schools on the basis of their performance on an IQ test. These samples were thus highly self-selected, with students that could be reasonably expected to improve their SAT scores substantially just by taking the test more than once. Pallone suggested that the average gains of his sample taking the long-term program might be best compared to the expected gains nationally (35 points) among all students taking the SAT-V twice. This notion of a post-hoc control group has also been suggested by others for studies that lack control groups (Messick, 1980; Slack & Porter, 1980; Messick & Jungeblut, 1981; Zuman, 1988; Becker, 1990; Kaplan, 2001). Such a strategy is almost guaranteed to be suboptimal as a means of estimating a treatment effect. Even if by luck the treatment and control groups were comparable, the data allowing a researcher to verify the equivalency of treatment and control group characteristics is seldom available for both groups, or has not been collected at all.

A different issue that applies to the Pallone and Marron studies is whether the programs should be classified as short- or long-term, or even as coaching. The short and long-term reading courses in Pallone's study involve 45 and 100 hours of student contact time. Pallone described his program as a developmental reading course, and sought to distinguish it from short-term coaching efforts. The full-time program in Marron's study constitutes somewhere in the neighborhood of 300 hours of contact time. Again, as in the

Pallone study, students were not being coached in specific test-taking strategies per se, but were instead exposed full-time to content related to the SAT. Whether such programs can be reasonably classified as coaching has been a topic of dispute in the coaching literature (Jackson, 1980; Slack & Porter, 1980).

Johnson (1984) reported on a study sponsored by the National Association for the Advancement of Colored People to improve the SAT scores of low-income minority students with school-based coaching programs. The study involved students at three different sites in the urban areas of Atlanta, New York City and San Francisco. Black students were given the opportunity to volunteer to be coached for the purpose of improving their SAT performance. The intent was to randomly assign these volunteers to treatment and control conditions, with students initially assigned to the control group receiving delayed exposure to the treatment. The coaching treatment involved about five hours of instruction per week over six weeks, split evenly on the math and verbal sections of the test. The focus of the instruction, developed by the National Association of Secondary School Principals, was content review and item practice. Coached students were to be tested with retired, shortened SAT forms before and after the treatment.

Because of administrative difficulties, random assignment in the Johnson study failed completely at the Atlanta and New York sites—students self-selected themselves into immediate and delayed treatment conditions. At these sites no coaching effect was estimated, but average SAT-V and SAT-M gains of 57 and 44 points were reported. At the San Francisco site, random assignment successfully placed a small number of

students into immediate and delayed treatment conditions. Those in the delayed treatment group were given an SAT pre-test at the same time that those in the immediate treatment group were given their SAT post-test. Sizable and statistically significant SAT-V and SAT-M coaching effects of 121 and 57 points were estimated as the difference between average scores of immediate and delayed treatment groups. The interpretability of these effects is threatened by substantial attrition in Johnson's San Francisco sample: only 23 out of the original 39 students assigned to the immediate treatment and only 12 out of the original 29 students assigned to delayed treatment remained in the study. Attrition of this magnitude suggests that while the Johnson study was intended as a randomized design, it is more appropriately evaluated as an observational study. As such, the study suffers from an extremely small sample size and probable bias of an unknown magnitude.

Commercial Coaching

Kaplan (2001) reported on coaching provided for two cohorts of nine students located in affluent New England suburbs. The cohorts had received coaching in the summers of 1999 and 2000 respectively from a program intended to help students improve their performance on the SAT-M. The program can be classified as commercial in the sense that Kaplan charged students a fee, but unlike most commercial programs it was provided on a very small scale and taught only by Kaplan himself. The elements of Kaplan's program included instruction, practice on test items, working in pairs, individual mentoring and homework for a total of about 20 hours of student contact time over one

month. For each cohort, Kaplan calculated average SAT-M score gains of 60 and 87 points from two official test administrations. In a comment on the study (Briggs, 2001) I noted the equivocal task of interpreting these gains as effects in light of the extremely small and self-selected sample used in Kaplan's study.

Computer-based Coaching

Coffin (1987) was unable to maintain a study that began with a randomized design. Coffin's research design had seemed straightforward: 40 student volunteers from an urban high school in Massachusetts were randomly assigned to either exposure or non-exposure to the Hayden SAT preparation software. An official version of the SAT was to serve as pre-test and post-test around the treatment. As motivation to participate in the study, all SAT fees were waived. Coffin and his research team were surprised to find substantial sample attrition—over half the students in both treatment and control groups. Student use of the Hayden software was difficult to monitor, and for many students, the software was not flexible enough to meet their needs, causing them to grow increasingly frustrated with their preparation. In a subsequent study involving no control group, Coffin reported score gains for students using Hayden, but could only speculate on the size of any coaching effects.

1.3 Observational Studies

Many coaching studies are designed as observational studies, and though these studies are a considerable methodological improvement over studies with no control groups, the coaching estimates they produce may be biased. Predominant approaches taken to account for this bias have included statistical matching and linear regression. When certain assumptions hold, these approaches allow the researcher to generate effect estimates equivalent to those that would have been reached in a randomized design. The assumptions behind the linear regression approach is given more detail in Chapter 2, but ceteris paribus, both statistical matching and linear regression will only succeed in reducing bias if an adequate set of covariates has been chosen to control for the differences between coached and uncoached group characteristics that are theoretically related to SAT performance.

Coaching in School Settings

The three earliest coaching studies (Dyer, 1953; French, 1955; Dear, 1958) were sponsored by the CEEB and conducted by researchers directly affiliated with ETS. All three studies had similar designs. Students from one or more schools participated in a series of coaching sessions embedded into their high school's curriculum. Subsequent SAT performance for these coached students was then compared to a different group of students not participating in the sessions, but taking the SAT in the same time period. The schools from which treatment and control students were taken were intended to be

the same in every relevant way with the exception of the coaching treatment. The nature of the coaching in the three studies was quite similar, involving drill, practice, feedback and test familiarization during 12 to 20 hours of contact time split evenly between the two sections of the SAT. In all three studies linear regression was used to estimate coaching effects as the difference in post-coaching SAT scores while holding constant covariates such as pre-coaching SAT scores, gender and academic course exposure.

The French study built upon the Dyer study, while the Dear study built upon both. Dyer's study, based upon a sample of 418 male students in two private schools, found statistically significant coaching effects of 5 and 15 points respectively for the SAT-V and SAT-M. Dyer's finding of an interaction effect of 29 points on the SAT-M for coached boys who were not taking math at the time of the test led him to conclude that 1) the SAT-M was more coachable than the SAT-V and 2) coaching might serve as a substitute for math instruction and thereby boost SAT-M performance. French's study was essentially a replication of the Dyer study with 319 male and female students from public rather than private schools. Dear's study involved 586 male and female students from both private and public schools, but now the coaching treatment was to be varied to involve both shorter and longer-term coaching sessions, and instruction would be individualized rather than at the group level.

The general magnitude of the coaching effects were consistent across the three ETS studies—the largest coaching effect found for any section of the test was about 20 points. Yet the patterns of the findings were inconsistent. The studies by Dyer and Dear

suggested that the SAT-M was more coachable than the SAT-V. In the French study there was some evidence to support the opposite conclusion. Dyer found a negative interaction with coaching for boys taking math courses at the time of the testings. French confirmed this finding for the boys in his study, but reported a *positive* interaction among girls who are coached and taking a math course. Positive interactions were also found for both boys and girls in the Dear study.

The coaching effects estimated in these early ETS investigations are likely biased because groups of coached and uncoached students were sampled from different schools. One mitigating factor is the relatively homogenous nature of students taking the SAT during the 1950s in terms of their demographic characteristics. Still, the differences in school experiences of students in treatment and control groups were unobserved and uncontrolled, and may serve to confound the estimates in the Dyer, French and Dear studies. In the French study there was evidence that the student groups from the three schools (A, B & C) in the study were not equally comparable. In all three schools a pre and post-coaching SAT score was available, regardless of a student's coaching status. Only school C students received coaching for the SAT-M, while school A and B students took the SAT-M but received no coaching. When average SAT-M score gains for school C were compared to average gains for school B, the math coaching effect was estimated as 18 points. Yet when school A was used as the control, the estimated effect for school C was significantly smaller—just 6 points. If students in schools A and B had been comparable to students in school C in every way other than exposure to coaching, one

would have expected to get roughly the same SAT-M coaching effect estimate regardless of which school—A or B—had been used as the control.

Another methodological weakness common to these studies is that the pre-coaching SAT score of students in both treatment and control groups was not based upon an official SAT administration, but upon a simulated administration using a retired test form. If coached and uncoached students were differentially motivated to perform well on the unofficial SAT, this would serve to bias estimates of coaching effectiveness. For example, coached students may know they will be getting instruction that will help them improve their SAT scores after the first testing. Since first testing scores will not be officially reported, coached students may save their best efforts for the second testing. Conversely, uncoached students may recognize that they will not be getting special instruction to improve their scores, and therefore give the first testing their best effort, because it is viewed as their best opportunity for practice before the real thing. Or, exposure to the SAT in the first testing may actually serve to motivate uncoached students to find other ways to prepare for the test before the second testing. All of this underscores the importance of controlling for student motivation and alternate methods of test preparation in coaching studies, as these are fundamental sources of confounding.

Kintisch (1979) attributed a small SAT-V effect to a high school reading course elective offered to 12th grade students in a Pennsylvania high school over a three year period. The course, which emphasized the efficient use of time, reading comprehension skills and the easing of test anticipation anxiety, met twice a week over 20 weeks for a

total of about 30 hours. All students in the study had taken an official SAT a first time as juniors and then took the test again as seniors. For half of these students, Kintisch's reading program was taken in between these testings. The other half of the sample served as the control group, chosen by matching students with similar 11th grade SAT-V scores. Kintisch reported a 14 point effect, but provided no information to support the statistical significance of this effect. Other than matching on 11th grade SAT-V scores, the possibility of bias in effect estimates was not addressed, so there is no way to validate Kintisch's assertion that her experimental groups were equivalent in all relevant ways other than exposure to the treatment.

A larger effect for a school-based reading program was presented in a doctoral dissertation by Burke (1986). Burke's sample consisted of 100 students from a large suburban high school in the South, half of whom were in the 11th grade, with the other half in the 12th grade. The students were described as generally upper-middle class with strong college ambitions. The samples of 11th and 12th grade students were analyzed separately. Each student cohort was enrolled in a semester-long course emphasizing vocabulary growth, reading comprehension and the analysis of literature. Students taking the course met five days a week over 12 weeks for a total of about 52 hours of classroom instruction, not including regular homework assignments also intended to strengthen reading skills. Students taking the reading course were matched to students in the same high school not enrolled in the course on the basis of prior PSAT-V or SAT-V scores, high school grade point average, and enrollment in either regular or advanced English classes at the time of the study. All 11th grade students had taken official administrations

of the PSAT before and after the reading course. All 12th grade students had taken official administrations of the SAT before and after the course. Burke surmised from the fact that all students in her sample had taken either the PSAT or SAT previously (multiple times for many of the 12th grade students), and from evidence gleaned from student interviews that both treatment and control groups were equally motivated to do well on the official PSAT or SAT administrations that served as the post-test in her study. Burke found a relatively large, statistically significant verbal coaching effect of 34 and 56 points respectively for her 11th and 12th grade students.

While the possibility of bias in estimated effects can always be raised as a methodological criticism given an observational design, Burke's study may be as close as one might expect to get to approximating a purely experimental setup. All students in the sample came from the same high school and had the same general academic background. Burke demonstrated empirically that her 11th and 12th grade samples were identical on average with respect to a number of covariates one would expect to influence PSAT or SAT test performance. The issue of differential motivation among treatment and control students was addressed by comparing scores only on official PSAT or SAT administrations and by interviewing treatment and control students about their college ambitions. Burke estimated coaching effects by comparing score gains for treatment and control students, and then testing for statistical significance. This is an example in which no post hoc statistical approach was taken to correct for potential bias. Instead, the author has carefully shown a priori that treatment and control groups are statistically equivalent along a host of observable variables related to SAT performance.

Harvey (1988) examined the effectiveness of three after-school workshops intended to help students prepare for the SAT-M. The sessions lasted a total of four hours and were offered to students at two high schools in a metropolitan area of Atlanta, Georgia. Coaching topics included an emphasis on test familiarization, strategies for reducing test anxiety, a review of math concepts and practice on sample test items. Harvey's control group was randomly selected from student volunteers interested in being coached. These students received delayed exposure to the coaching treatment, but were tested twice at the same times as the students receiving immediate exposure to the treatment. Retired versions of the SAT were used for all testings. Harvey estimated a statistically insignificant coaching effect of 21 points using ANCOVA with prior SAT-M score as a covariate. These findings did not change when considering the effect of coaching via a videotape of the live workshop.

Harvey made the case that while her treatment and control samples were not chosen randomly, they were effectively equivalent in all ways that might theoretically influence SAT performance. To this end Harvey presented a comparison of the characteristics of coached and uncoached students with respect to demographics, previous test-taking experience, academic achievement, other forms of test preparation and motivation. The experimental groups appeared roughly equivalent along these variables, unfortunately, this apparent equivalence was not demonstrated statistically. The use of delayed treatment and retired SAT forms as pre and post-test in Harvey's study also raises questions about differential motivation among the experimental groups.

Schroeder (1992) investigated the effect of small group coaching for the SAT-M. A small sample of students volunteering from New York metropolitan high schools were offered coaching in groups of 2-3 over eight weeks for a total of about 16 hours. The coaching focused on developing problem-solving skills, and beyond the scheduled instructional contact time, students were expected to follow a structured plan of study. Control students were randomly selected from area high schools. According to Schroeder, all students in the sample were of the same approximate age, came from upper middle class backgrounds, and had taken the same math courses in high school, though this was not demonstrated quantitatively. Both treatment and control group students had taken official administrations of the PSAT and SAT. Schroeder estimated a statistically significant SAT-M coaching effect of 46 points using ANCOVA with PSAT scores as a covariate.

Schroeder suggested his sample of students were generally equivalent in terms of their demographic and course-taking backgrounds, yet there was clear evidence that treatment and control students differed with respect to prior PSAT scores. Students being coached had average PSAT scores that were 30 points higher than students in the control group. Without correcting for this difference, the estimated coaching effect would have been 69 points. Correcting for this with ANCOVA reduced the estimated effect to 46 points. Schroeder did not collect, or at least did not report other relevant variables (e.g. gender, academic background) that might have been entered as additional covariates in

the ANCOVA model. It is unclear the extent to which other covariates would have further adjusted the estimated effect.

Wrinkle (1996) conducted a study of a school-based coaching program at a suburban high school in a metropolitan area in Texas. The program lasted nine weeks for a total of approximately 70 hours of instruction emphasizing the development of verbal reasoning skills. Wrinkle's sample included 10^{th}, 11^{th} and 12^{th} grade students characterized as upper middle class. Coached students in Wrinkle's sample were matched to uncoached students on the basis of their PSAT verbal score, high school grade point average, grade level and gender. Both coached and uncoached students took official administrations of the PSAT and SAT I. Using an ANCOVA model, Wrinkle estimated a statistically significant SAT-V coaching effect of 33 points.

A problem common to all of these observational studies of school-based coaching is that they lack generalizability. All involved sample sizes drawn from a small number of high schools, usually no more than one or two. Students that were sampled tended to be of high academic ability from upper-middle class families.

Coaching in Commercial Settings

The first studies investigating the effects of commercial coaching were published by Frankel (1960) and Whitla (1962). Both studies employed similar methodological designs: statistical matching of coached and uncoached students into pairs that had taken

official administrations of the SAT twice. In both studies the coaching treatment took place outside of the normal high school curriculum for a fee. The relative scores changes of coached and uncoached students were then compared statistically using a one-tailed t-test. Neither Frankel nor Whitla found coaching effects that were statistically significant.

Frankel's sample was drawn from a single high school of academically gifted students who had previously taken official administrations of the SAT. Every student reporting that he or she had received commercial coaching between two SAT testings over the span of six months was matched to an uncoached student from the same school on the basis of gender and first SAT score. Coached students were enrolled in a course involving 30 hours of contact time split evenly between the math and verbal sections of the test. If the high-ability students in this sample were equally motivated to improve their scores and only differed in their coaching status, gender and 1st SAT scores, then Frankel's coaching estimates would not suffer from bias due to confounding. On the other hand, if one assumes that both coached and uncoached students are equally motivated to perform well on their second SAT testing, then it may be the case that uncoached students prepared for the SAT in undocumented ways that resulted in the same gains as those of their coached peers. This criticism applies to any number of coaching studies with omitted variables describing the range of activities students use to prepare for the SAT.

Whitla's study differed from Frankel's in three significant ways: 1) uncoached students were sampled from different schools than the one school attended by coached

students, 2) students pairs were matched only on the basis of scores on an unofficial SAT testing prior to the treatment, and 3) the length of the coaching treatment was substantially shorter, a total of 10 contact hours. Whitla acknowledged the difficulty of controlling for different levels of motivation among coached and uncoached students (1962, p. 33). As a strategy for motivating both groups equally to improve their scores in the second SAT testing, all students were pre-tested with an unofficial version of the SAT immediately before half the students were to receive coaching. Again, as was the case with the Frankel study, variables describing other ways students prepare for the test were unobserved. To the extent that such variables might be differentially correlated with SAT performance as a function of treatment status, coaching effect estimates would probably be biased.

The Federal Trade Commission (FTC) initiated a study in 1976 to investigate the advertising and marketing practices of commercial coaching companies. Enrollment data for students participating in the courses offered at three coaching companies in the metropolitan New York area between 1975 and 1977 were subpoenaed. PSAT and SAT scores for these students were also subpoenaed from the CEEB. A control group of SAT test-takers was selected at random from high schools in the same geographic areas. Demographic and academic background characteristics for coached an uncoached students were taken from the Student Descriptive Questionnaire (SDQ) filled out by students as part of taking the SAT. Coaching at Company A (Kaplan) consisted of a 10-week course with four hours of class per week split between preparation for the verbal and math sections of the test. At Company B (Test Preparation Center) coaching was

shorter in duration, spanning 24 hours of classroom instruction on both sections of the test. Coaching at Company C was not analyzed because of the small number of students involved.

The initial report on the FTC study was released as a staff memorandum by the FTC's Boston Regional Office (1978). While this memo reported SAT-V and SAT-M coaching effects at company A as large as 55 and 40 points, it was strongly criticized by the central administration of the FTC itself on the basis of flaws in the data analysis. Bias in effect estimates was at the heart of these flaws. Coaching effects had been estimated by comparing coached and uncoached students without controlling for obvious differences in observed variables such as socioeconomic status and academic background, as well as probable unobserved differences in motivation.

The data were subsequently reanalyzed by the FTC's Bureau of Consumer Protection (BCP)(Federal Trade Commission, 1979; Sesnowitz, Bernhardt et al., 1982). Coaching effects were re-estimated for Companies A and B using linear regression models. Covariates for academic background, demographic characteristics and test-taking experience were held constant and the relative SAT score gains of coached and uncoached students were compared as a difference between the intercepts of the regression surfaces. The estimated effects were mixed: coaching appeared to be have statistically significant SAT-V and SAT-M effects of as much as 30 points per section for Company A, but coaching effects were small and statistically insignificant for Company B.

The BRO analysis and BCP re-analysis of the FTC data were subsequently critiqued. The focus of the criticism was on what Messick (1980) described as the "bane of self-selection bias" in non-experimental designs. In the absence of a randomized assignment to experimental conditions, Messick cautioned, coaching effect estimates could only be interpreted "equivocally" as the *combined* effect of coaching and self-selection. As for statistical approaches to correct for selection bias, Messick wrote:

> "We hope to reduce—but cannot eliminate—this equivocality [in the interpretation of the coaching effect] by conducting multiple alternative statistical analyses."

This is very much a key point, because if a study suffers from selection bias, then linear regression, the statistical approach used in the FTC study, cannot produce unbiased effect estimates. As I discuss at length in Chapter 2, linear regression can only reduce bias due to confounding from omitted variables that have been measured.

Two alternative statistical analyses of the FTC data were undertaken by ETS researchers (Rock, 1980; Stroud. 1980). Stroud adapted an approach taken by Belson (1956) and Cochran (1969) as a way to estimate student-level coaching effects. The approach involved three basic steps.

1) A regression of SAT score on a specified set of covariates for all uncoached students.
2) Next, the coefficient estimates from this equation based on uncoached students would be used to calculate predicted SAT scores using only the

sample of coached students. These predicted values would reflect, in theory, the scores to be expected if coached and uncoached students came from the same underlying population.

3) Last, the predicted SAT scores for coached students were subtracted from observed scores. These residual values were estimates of the coaching effect at the student level.

In using this approach, Stroud's analysis differed from that taken by the BCP in his use of a fuller set of covariates, imputation techniques to adjust for missing data, and the estimation of interaction effects between coaching and other variables. Nonetheless, Stroud's basic findings supported those of the BCP re-analysis: a statistically significant effect for coaching of about 30 points on each section of the test for Company A, and small, statistically insignificant coaching effects for Company B.

A second ETS re-analysis by Rock focused on the BCP finding that Company A was equally effective at coaching verbal and math reasoning abilities. Rock was skeptical of this finding because previous studies (Dyer, 1953; Dear, 1958; Evans & Pike, 1973) had suggested the SAT-M was more susceptible to coaching than the SAT-V, and this seemed intuitively plausible to Rock because the math section had closer ties than the verbal to the high school curriculum. The BCP coaching estimates had been made under the assumption that after controlling for certain covariates, the only differences in rates of learning and growth between coached and uncoached students was a fixed constant equal to the coaching effect. In other words, the coaching effect could be estimated as the conditional difference in regression surfaces. Rock posed a question: what if coached

students, being more motivated and academically able, learn faster than uncoached students, even without exposure to the coaching treatment? This would result in an overestimate of the coaching effect. Rock's analysis considered students who took the PSAT once and the SAT twice, but who received coaching after taking the SAT for the first time. Rock found evidence that the SAT-V scores of these "to be coached" students were already growing at significantly faster rate than those of control students when conditional PSAT-V to SAT-V gains were calculated. A similar pattern was not found for SAT-M scores. Taking these rates of differential growth into account, Rock proposed an adjusted estimate for the Company A verbal coaching effect of 17 points.

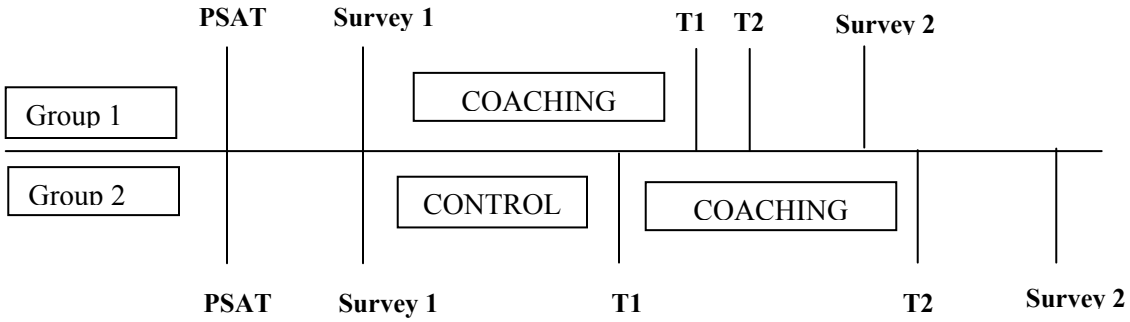
The FTC study was certainly not the first observational design to have its interpretation clouded by questions of bias. It was however, the first case in published record of researchers employing different statistical approaches with the intention of diagnosing and reducing bias in estimates of coaching effects. From a non-methodological standpoint, the FTC study also serves as a demarcation of sorts, because it helped bring the coaching issue to the forefront as something for both public and academic debate.

In a doctoral dissertation, Zuman (1988) investigated the effectiveness of commercial coaching for two different samples of high school juniors coached by The Princeton Review in 1986. The first sample of 55 students was taken from high schools in the New York City area, and consisted of primarily white males and females from relatively wealthy households. All students had signed up for the standard SAT

preparation course offered by The Princeton Review (~30 hours of contact time), but differed in the timing of their planned coaching sessions.

After controlling for a subset of background characteristics in a linear regression, Zuman found substantial coaching effects—48 and 56 points—on the verbal and math sections of the test for his observational sample. Just as with the FTC study, these findings can only be interpreted with much equivocation because of methodological design issues. Zuman's design is illustrated in Figure 1-2 below.

Figure 1-2. Zuman's Study Design



In figure 1-2, T1 refers to a special administration of a retired SAT form, while T2 refers to an officially administered version of the test. Ideally, one would wish to compare the SAT scores for two equivalent groups of students, one of which had received coaching from The Princeton Review. In Zuman's study, the students in each group were only roughly equivalent, and *both* groups received coaching before taking an official SAT administration. In order to estimate a coaching effect, Zuman compared official SAT scores (T2) for group 1 with the unofficial scores (T1) for group 2.

There were two sources of non-equivalency in Zuman's design: the characteristics of students in each group, and the characteristics of the SAT post-test measure. Zuman hoped to minimize the former source of non-equivalency by employing a regression model with covariates for PSAT scores, cumulative GPA and other background variables. The latter source of non-equivalency was due to the delayed treatment approach in the study's design—at the point in the study that group 1 had taken an official administration of the SAT with T2, group 2 had only taken T1. Zuman's solution to this problem may explain his large coaching effect estimates. Zuman correlated the scores of his group 1 students who had taken both T1 and T2 administrations. He found that while they were strongly correlated, the students' average T2 scores on the math and verbal sections were 6 and 20 points lower than their respective average T1 scores. Zuman extrapolated this finding to the students in group 2, subtracting 6 and 20 points from their T1 verbal and math scores with the intent of making them comparable to the average T2 scores from group 1. This adjustment essentially inflated Zuman's coaching effect estimates, and it is unclear that the adjustment made much sense. After adjusting the T1 scores for the control students in group 2, PSAT to T1 score changes indicate average *decreases* of 7 and 16 points on the verbal and math sections. There is also reason to suspect that student motivation in Zuman's sample had some interaction with group status. As Smyth (1990) pointed out, students in group 2, knowing that their promised coaching was imminent, and knowing that no stakes would be attached to their T1 performance, were almost surely differently motivated then group 1 students, who were taking the SAT in a high-stakes setting. Zuman himself noted that his results were at best suggestive, putting aside estimation bias issues, because his sample was very small.

Zuman also evaluated the effect of commercial coaching by The Princeton Review for a second experimental sample consisting of 48 low-income minority students also from the New York metropolitan area. These students were given scholarships to attend the coaching program, and were randomly assigned to experimental conditions using the same testing schedule as that used for the observational sample. Zuman found a statistically significant SAT-M coaching effect of 57 points, but no significant effect for the SAT-V. Again, the interpretation of these effects is clouded by the probable non-equivalency of Zuman's experimental groups due to substantial attrition rates (31%), difficulties in ensuring a standardized coaching experience, and the same motivational issues that applied to Zuman's observational sample.

Whitla (1988) and Snedecor (1989) published large-scale studies on the effectiveness of commercial coaching. Whitla surveyed incoming freshman at Harvard University, while Snedecor used the same survey questionnaire on a sample of seniors from 10 private high schools in Pennsylvania. Their results indicated that only small effects could be attributed to commercial programs. These studies, while suggestive, are of limited value in terms of their presented methodological approaches. Both samples involved highly self-selected samples, and each study's results were based upon self-reported student PSAT and SAT scores. Neither Whitla nor Snedecor provided any information on potential non-response bias and neither survey asked students about the timing of their coaching or involvement in other test preparation activities. Statistical

approaches taken to reduce bias were not presented. In fact, neither author reported or commented on the statistical significance of their relative score change comparisons.

Smyth's study (1990) of commercial coaching programs was a more rigorous version of a previous study (1989) and utilized a sample similar to that of Snedecor's—seniors from 14 private schools in the Mid-Atlantic region. Unlike Snedecor, Smyth presented a more carefully considered methodological design. Smyth's sample of senior students was surveyed about their SAT test preparation practices shortly before graduation in the spring of 1989. About 80% of the students completed the survey. All official PSAT and SAT scores of the participating students were provided to Smyth directly by school counselors. Smyth reported two types of estimates for pooled coaching effects: 1) "raw" effects—relative score changes from PSAT to best reported SAT of coached and uncoached students; 2) "adjusted" effects—relative score changes adjusting for group differences between coached and uncoached students. These latter adjusted estimates involved the use of an ANCOVA analysis with covariates for PSAT scores and number of times taking the SAT. Controlling for just these variables reduced the raw coaching effect estimates from a combined 33 points to a combined 24 points.

A study by Powers & Rock (1999) was the first to provide national estimates for the effect of commercial coaching programs. Powers & Rock surveyed a stratified random sample of students who had taken the SAT I nationally between the fall of 1995 and the spring of 1996. Students were asked to indicate by what method and how long they had prepared for the SAT I. Students reporting that they had participated in

coaching from a commercial company were prompted to specify the name of the organization providing the service. Students were also asked to answer questions about their motivation levels for performing well on the test. The responses to this survey were subsequently merged with official PSAT and SAT I score data, and with student responses to the Student Descriptive Questionnaire (SDQ). SDQ data include variables for student socioeconomic status, academic background and college aspirations.

Powers & Rock estimated a series of coaching effects with multiple statistical models. In their simplest "Raw Changes" model, coaching effects were calculated as the SAT I score changes in coached students relative to uncoached students. These effects amounted to 8 points on the verbal section and 18 points on the math section, both statistically significant. Powers & Rock demonstrated that their sample of coached and uncoached students differed along a number of characteristics related to SAT I performance. Given this fact, one might expect to find bias in a Raw Changes coaching estimate. Five statistical models—ANCOVA, Propensity Matching Model (PMM), Instrumental Variables (IV), Heckman Model, Belson Model—were applied to estimate alternative coaching effect estimates, with each model specification relying on differing combinations of covariates and underlying assumptions about the nature of bias. The ANCOVA model itself had been used extensively in previous SAT coaching studies, though the specification of the model varied with available covariates in a given study. The Belson model had previously been applied in Stroud's reanalysis of the FTC coaching study. The PMM approach (Rosenbaum & Rubin, 1983; Rosenbaum, 1995) is a more sophisticated form of statistical matching, also previously applied in numerous

coaching studies. The IV and Heckman Model approaches, long popular in econometric research (Heckman, 1979; Greene, 1993), were new in the context of SAT coaching, and involve the use of two related structural equations to estimate a coaching effect purged of bias due to both confounding and self-selection.

Interestingly, the use of the different models had little impact on the overall estimates of the coaching effect. Without correcting for bias using the Raw Changes model, the combined coaching effect on both sections of the SAT I was estimated as 26 points. Attempts to correct for bias with the five statistical models produced remarkably consistent combined effects ranging from 22 to 34 points. Hidden in this were some surprising results, as many of the models gave conflicting evidence for the direction of bias present in the Raw Changes estimates of the SAT-V and SAT-M effects. Different models suggested that the Raw Changes coaching effects might be underestimating or overestimating the true size of the coaching effect, though the magnitude of the bias appeared to be small in all the models.

In a precursor to this dissertation (Briggs, 2001) I evaluated the effect of coaching for a stratified random sample of students taking the PSAT and SAT nationally between 1989 and 1992. Unlike the Powers & Rock study, my study relied entirely on a pre-existing source of data: the National Education Longitudinal Survey (NELS). Though NELS was not designed with a coaching study in mind, the data included transcript information on academic performance (including official PSAT and SAT scores) and many covariates theoretically related to test performance. Students sampled in NELS

were asked explicitly about how they had prepared for the SAT, though no information was available as to the quality of the preparation.

I estimated coaching effects first with a Raw Changes model and then with three different specifications of a linear regression model. Each regression specification involved a wider set of covariates as a means of correcting for group differences between coached and uncoached students in the national sample. I found that under the most fully specified regression model—which included covariates for academic achievement, demographic characteristics and student motivation—the math and verbal coaching effects found under the Raw Changes model were reduced from 17 and 13 points respectively to 15 and 6 points. The coaching effect estimates in the this study were strikingly similar to those found in the Powers & Rock study.

A weakness in the national estimates for the effects of commercial coaching in both the Powers & Rock and Briggs studies is in the definition of the coaching treatment. Both studies assume a degree of treatment homogeneity that may be unwarranted. Perhaps low quality coaching drives down the large effects of high quality coaching? Powers & Rock addressed this issue by considering the subset of students in their sample who reported that they had been coached by either Kaplan or The Princeton Review, two of the most widely known commercial test preparation companies. At one company an estimated coaching effect on the SAT-M of 33 points was significantly larger than the pooled estimate of about 18 points, but other than this, the estimated coaching effects remained fairly consistent with those from the pooled estimates. In my study I was

unable to analyze subsamples in a similar manner because no information was available to differentiate commercial coaching companies in the NELS database.

Another criticism of the Powers & Rock study is that if suffers from non-response bias, since only 63% of the students surveyed at random chose to respond. The Powers & Rock coaching estimates could have been biased as a function of the characteristics of coached and uncoached students not responding to the survey. None of the statistical models used by Powers & Rock would have addressed this bias.

1.4 Randomized Studies

The classic solution to estimating an unbiased treatment effect in medical settings has been to randomly assign subjects to experimental groups. But even in medical settings, randomized designs can be difficult to establish and maintain. This difficulty is even greater in social science contexts. Ethical concerns are a prominent obstacle to randomized designs in coaching studies. If coaching really does have a substantial positive effect on SAT performance, then coached students in a given study would gain an unfair advantage over uncoached students, even if assignment of this advantage took place randomly. Randomized studies of SAT coaching studies often attempt to circumvent this problem by offering students selected to the uncoached group the opportunity for delayed exposure to coaching. Hence the study proceeds as idealized: coached and uncoached students take the SAT as a pre-test at the outset of the study, one group receives the treatment for a set period of time, and then both groups take the SAT

as a post-test. At this point the study is complete, but in the interests of fairness the uncoached students are now given the opportunity to be coached and re-take the SAT. To facilitate this delayed treatment design, randomized coaching studies tend to involve samples of high school juniors rather than seniors.

A popular randomization approach found in studies of computer-based coaching involves the use of what is known as the Solomon Four-Group Design, useful as a means for determining if score improvements can be attributed to coaching, the act of being pre-tested, or some combination of the two. In this design students are assigned to four conditions:

1) SAT pre-test, coaching treatment, SAT post-test
2) SAT pre-test, no coaching treatment, SAT post-test
3) no SAT pre-test, coaching treatment, SAT post-test
4) no SAT pre-test, no coaching treatment, SAT post-test

If the pre-testing is having an effect independent of the coaching, this will be revealed in a comparison of conditions 1 and 2 with conditions 3 and 4. If the pre-test is having an effect on SAT post-test scores that interacts with the coaching itself, this would be revealed in a comparison of condition 1 with condition 3. If pre-testing has no effect on post-test SAT performance, then—provided that the students in the experimental conditions are equivalent on average—the effect of coaching can be estimated by comparing the score gains of students in conditions 1 and 3 with those of students in conditions 2 and 4.

Coaching in School Settings

Roberts & Openheim (1966) conducted the first randomized SAT coaching study with an experimental sample consisting of disadvantaged male and female Black students from 14 high schools who had volunteered to "participate in a program designed to help high school students perform well on the multiple-choice tests that many colleges require for admission and financial assistance" (1966, p. 2-3). All students were pre-tested with a retired PSAT form and then assigned to coached and uncoached groups. Students in treatment groups at six schools received 7.5 hours of verbal coaching over a 4-6 week period. At eight other schools the treatment was 7.5 hours of math coaching over the same time period. Coaching followed a standardized curriculum emphasizing practice on specific SAT item formats and strategies for taking the test efficiently. Students in control groups were promised that they would also receive coaching once both groups had taken a post-test using a different retired PSAT form. Coaching effects were estimated as the average difference in gains between treatment and control groups, and the statistical significance of this difference in means was evaluated with a t-test. Roberts & Openheim reported a statistically significant coaching effect, pooled across schools, of 1.4 points (14 on the SAT-V scale) on the verbal section of the PSAT. There was great variability in verbal effect estimates from school to school, ranging from a statistically significant effect of 5.2 at one school to a statistically insignificant effect of –2 at another. For students coached in math, the reported pooled effect was not statistically significant, nor were school-level estimates, which ranged from –2.5 to 2.2 points (–25 to 22 on the SAT-M scale).

The Evans & Pike (1973) study was the first to systematically investigate the effect of coaching on specific SAT item formats. The study was initiated by the CEEB to determine if certain math item formats were more coachable than others. Eleventh grade student volunteers of average academic ability from 12 high schools were randomly assigned to one of three treatment conditions receiving 42 hours of combined instruction and homework specific to each of three possible item formats: 1) Regular Mathematics (RM) 2) Data Sufficiency (DS) and 3) Quantitative Comparisons (QC). These item formats ranged in their complexity with the RM format being most straightforward, and the DS and QC format more complex.[5] A different group of students was randomly assigned to the control condition and received no instruction on item formats. Students were pre and post-tested using retired SAT forms in a simulated test administration. Evans & Pike found that the QC and DS item formats were more susceptible to coaching than the less complex RM format, with QC items showing the greatest susceptibility to coaching. No estimate of a coaching effect for the full SAT-M section could be made because no student received coaching on all item formats. However, in a subsequent review of coaching studies Pike suggested that coaching tailored to the item formats in the SAT-M could be expected to produce an effect of 33 points (Pike 1978, p. 12).

Alderman & Powers (1980) investigated the effectiveness of school-based coaching programs on the verbal section of the SAT using mostly upper-middle class high school juniors sampled from eight private and public schools across seven

[5] At the time of the study the DS format was in the process of being replaced on the SAT-M by items with the QC format. Hence there was particular interest from the perspective of ETS to determine if the QC format was less susceptible to coaching than the DS format.

northeastern states. All schools participating in the study had agreed to allow students expressing an interest in special preparation for the SAT to be randomly assigned to treatment and control groups. Those students assigned to the treatment groups participated in verbal coaching sessions at their school ranging in length from 3 to 10 weeks for a total of 5 to 45 hours. Students assigned to the control group were given delayed exposure to the coaching programs after the study was completed. Both experimental groups participated in a simulated administration of the SAT using a retired form of the test after coaching took place for the treatment group. Each group had previously taken official administrations of the PSAT and the Test of Standard Written English (TSWE).

Alderman & Powers provided evidence to suggest that their randomized assignment had successfully established equivalent groups of students by comparing average PSAT-V and TSWE scores within high school samples. The means of treatment and control groups were virtually identical in every case. There were similar rates of attrition across treatment and control of about 90%. A linear regression model was used to estimate coaching effects to help correct for the bias that would have occurred if attrition happened differentially among coached and uncoached students (i.e. only the coached students scoring low on the PSAT dropped out of the study, while among the uncoached sample only students scoring high on the PSAT dropped out). Separate regressions, with PSAT and TSWE scores as covariates, were used to estimate a coaching effect for each of the eight schools in the study. The authors found substantial variability in the coaching effect from school to school, with effects ranging from a low of –2.75 to

a high of 28 points. Interestingly, the school with the largest effect was that with the shortest coaching program (5 hours), while the school with the smallest effect was more than twice as long (10.5 hours). Pooled estimates of the coaching effect were estimated using weighted averages and different regression model specifications.

Interpretation of the Roberts & Openheim and Alderman & Powers studies are hampered by the same problem: questionable motivation to perform for students in the control group. In both studies, as Messick first pointed out (Messick, 1980; Messick & Jungeblut, 1981), there was evidence to suggest that uncoached students lacked motivation to perform well on their post-test. In the Roberts & Openheim study, examining the scores of treatment and control students across testings revealed that the relative gains of coached students, particularly on the verbal section of the test, were driven by significant score *decreases* among uncoached students. In three of the eight schools considered by Alderman & Powers, the estimated coaching effect is driven by a decrease in average PSAT to SAT scores of the control group. This runs counter to expectations that students generally improve their scores over time upon retesting simply from the testing experience and normal maturation. At the very least, one would expect average scores for uncoached students to stay the same, within sampling error. In both studies students were given delayed treatment exposure and tested with retired PSAT or SAT forms. Because uncoached students were not taking an official version of the SAT, and knew that they would receive coaching in the future, these students may not have given an effort on the test comparable to that of their coached counterparts. This may explain the surprising score decreases among uncoached students.

Shaw (1992) conducted a randomized study with a small sample of ethnically mixed 12th grade students drawn from three high schools in southern California. All students selected for the study were planning to take an official administration of the SAT in the fall of 1988, had not previously taken the SAT, and had never taken an SAT coaching workshop. Half of these students were selected at random to participate in a one-day, 8 hour coaching workshop with an emphasis on general testwiseness. When the SAT scores of coached and uncoached students were compared, no statistically significant effect for coaching on either section of the test was found, nor were there any significant interaction effects for coaching with gender or ethnicity.

There are two methodological problems in Shaw's study. First, no information was collected about other forms of test preparation undertaken by students in his sample. There was a lag of up to a month between selection of the study sample and administration of the SAT. Students assigned to the control group may have sought out alternative means of preparing for the test, and this would not have been captured in Shaw's analysis. Second, Shaw provided no empirical evidence to verify the equivalency of treatment and control groups after random assignment.

Computer-based Coaching

With the advent of affordable personal computers in the 1980s, a new mode of coaching became available in the form of SAT preparation software. Five randomized

studies have been conducted to determine if coaching through a computerized medium can be considered effective.

Hopmeier (1984) examined a 7-hour computer-based coaching program for a sample of 9th, 10th and 11th grade students all enrolled in a geometry course in a Florida high school. Students were randomly assigned to individual or small-group training for the SAT using preparation software crafted by Harcourt-Brace, or no training at all. The coaching found in the software contained many of the same elements found under human instruction: math and verbal content review, diagnostic item practice and some emphasis on test-taking strategies. Coaching effects were estimated by comparing the average performances of the experimental groups on an unofficial administration of a retired SAT form following the treatment. Hopmeier found no statistically significant difference between the scores of students coached individually or in small groups. When the scores of all coached and uncoached students were compared, Hopmeier reported statistically significant effects of 57 and 37 points on the SAT-V and SAT-M.

In Hopmeier's study approximately 90 students were assigned to three experimental conditions. No empirical evidence was provided to verify that randomization had the desired, equalizing, effect on characteristics of the three student groups, hence bias cannot be ruled out as a factor inflating or deflating these coaching effect estimates. In addition, there was attrition in the sample across experimental conditions for those taking the verbal section of the SAT—which may explain or at least confound Hopmeier's finding of a larger coaching effect for the SAT-V than for the SAT-

M. Other problems were of a more logistical nature: students were tested with a retired SAT form in the last two weeks of the school year, and the test was administered as two separate sections (math and verbal) on different days. All of this was likely to affect student motivation in ways different from an official SAT administration, though it is not clear the extent to which this would have affected coached and uncoached students differentially.

Laschewer (1986) explored the effectiveness of computer-assisted instruction software for the SAT that he had designed. Laschewer randomly assigned a small sample of 11th grade students from a New York Catholic high school to one of four different treatment conditions using the Solomon Four Group Design. While Laschewer's design was a good one, his study was compromised by sample attrition, the use of retired SAT forms and anecdotal evidence of low student motivation during the post-test SAT administration. None of Laschewer's estimated coaching effects were statistically significant.

Curran's study (1988) of computerized coaching was unique in that students were randomly assigned to five different modes of preparing for the PSAT. Four of the modes involved the use of SAT preparation software individually, in a group, and individually with an accompanying book or with the book alone. The control condition involved a review of incorrect responses from a PSAT pre-test administered to students in all experimental conditions. Hence, Curran's control condition was different from the usual control defined in other studies simply as the lack of exposure to the coaching program.

Instead, the control condition was defined as an alternative, less regimented way of preparing for the PSAT.

Curran's sample was drawn from student volunteers in four Catholic Schools in two New England states in the fall of 1987. Half the schools were all-male, the other half all female. According to Curran, many of the students had taken an official administration of the PSAT 11 months prior to the study in the fall of 1986. All students were administered a retired PSAT form as a pre-test immediately before assignment to experimental conditions. After preparing for the test using one of the five modes, these students were expected to take an official PSAT administration as a post-test. Students were also administered extensive background questionnaires and short instruments to gauge their anxiety levels while taking the PSAT. Unfortunately, numerical discrepancies, missing tables and misspecified statistical models in Curran's dissertation text cloud the findings from this ambitious study.[6]

Using a sample of students from a college-prep level English classes at one rural high school in northeast Georgia, Holmes & Keffer (1995) considered the effectiveness of a computer program for improving scores on just the SAT-V. This program did not involve the typical characteristics normally associated with coaching—content review, item practice and an emphasis on general testwiseness. Instead, the program's emphasis was on a specific strategy for solving antonym and analogy item formats by drilling

[6] The key results from Curran's study are in his Tables 4.7a and 4.7b (104). These tables suggest sample attrition and little to no effect for using computerized software and/or books relative to reviewing a practice test. Yet Curran does not comment on or test the statistical significance of these results. Instead, he mixes randomly assigned and self-selected experimental conditions into a multifactor ANOVA model (107-113).

students in Latin and Greek root words and the use of these words to decipher English terms. Using a Solomon Four Group Design, Holmes & Keffer randomly assigned 115 student volunteers to four conditions involving combinations of a pre-test with a retired SAT form, eight hours of exposure to the computer software, and a post-test with a different retired SAT-V form. The effect of the computer program was a statistically significant 40 points.

Beyond the small sample size and lack of generalizability, a problem with the Holmes & Keffer study is the use of only the verbal section of a retired SAT as the pre- and post-test. Part of the difficulty of performing well on the SAT comes in having to complete both sections of the test in one intensive sitting. This aspect of the testing experience was eliminated from the study. A bigger issue is the use of retired SAT forms *before* the test changed in format to become the SAT I. One of the biggest changes to the test was the elimination of the antonym format from the verbal section. These were replaced with more reading comprehension passages. Hence there is reason to suspect that that knowledge of Greek and Latin root words would be less useful for the SAT I then it might have been for the more vocabulary-dependent SAT used as the outcome measure in Holmes and Keffer's study.

McClain (1999) evaluated two commercially produced software programs intended to prepare students for the SAT I. The study's sample was drawn from 12th grade student volunteers at a suburban high school in Maryland. McClain describes the academic ability of students from the school as lower than that of other comparable

Maryland schools in the same district. Students were randomly assigned to one of three conditions: use of Davidson test prep software, use of Stanford test prep software, and use of no software. Students in the treatment groups were expected to use the software for approximately 36 hours. All students in the study took an official SAT I as a pre- and post-test. McClain reported the results of an ANCOVA analysis indicating that the effect of using the Davidson software was a combined 54 points on both sections of the test, while the respective combined effect of the Stanford software was 84 points.

Omitted information in the write-up of McClain's study makes it difficult to attach much weight to these findings. McClain reports only a combined verbal and math coaching effect for the Davidson and Stanford treatments—coaching effects specific to the verbal and math sections are not provided. This is a glaring omission, and may imply that only combined effects were reported to mask statistically insignificant findings for the individual SAT I test sections. There are also discrepancies in McClain's narrative and appendix that suggest his research design changed during the course of the study.

1.5 Summary of Coaching Studies

Thirty-two studies[7] of SAT coaching were found in the 48 year period between 1953 and 2001. Eighteen of these studies were published in peer reviewed academic journals. Four studies were produced as institutional research reports, and 10 were published doctoral dissertations. The patterns of findings across studies are best analyzed

[7] This number does not include the re-analyses of the FTC data. Also, some studies included evaluations of more than one coaching program or site.

with respect to the different experimental samples, coaching treatments and methodological designs, and these are described in Tables 1-1 through 1-5. Table 1-5 includes SAT-V and SAT-M coaching effect estimates for each study.

Most SAT coaching studies—21 out of 32—have employed samples of 11th and 12th grade students drawn from public and private high schools in the northeastern United States, particularly in the New York metropolitan area. Another seven studies were based in southern states. Only one study was found with a sample from a western state (Shaw, 1992). Generalizing these results to the full test-taking population of the time requires at the very least an assumption of geographic homogeneity among the schooling experiences of students taking the SAT. Three studies involved national samples (Whitla, 1988; Powers & Rock, 1999; Briggs, 2001), the latter two of which were stratified and drawn at random. The nature of the population of students taking the SAT has changed over time and this fact is reflected to some extent by coaching study samples. As Evans & Pike noted in their 1973 study, the population of test-takers in the 1950s tended to be a relatively homogenous group of high ability white students from upper-middle class households. By the 1970s the socioeconomic characteristics of test-takers nationally had become more mixed. Many of the studies between 1970 and 2001 involved a mix of students with a wider range of demographic and academic background characteristics. Only three coaching studies have restricted their analysis to samples of low SES students. All three studies employed a randomized design, but the designs were compromised in each case by high student attrition possibly due to low student motivation. Based upon the study samples reviewed here, very little can be said about

the effect of coaching for low-income students from poorly educated households with respect to school-based, commercial or computer-based programs.

Table 1-2 reports coaching study sample sizes as a function of study characteristics. In general, there is great variability in the sample size of SAT coaching studies. The sample sizes of SAT coaching studies range from as small as 18 to as large as 2,554. Within this variability there are some noteworthy patterns. It should come as little surprise that the sample sizes in published coaching studies tend to be substantially larger than those of unpublished studies. For published studies the median sample size, including both coached and uncoached students, is 487 for evaluations of both SAT-V and SAT-M effects. For unpublished studies, the respective SAT-V and SAT-M sample size is about 100. Other trends worth noting are that observational studies tend to involve larger sample sizes than randomized or uncontrolled studies, and within observational studies, those evaluating commercial coaching tend to have substantially larger samples that those evaluating school-based coaching. It is also clear that randomized studies of computer-based coaching have all been conducted on a small scale.

Table 1-1. SAT Coaching Studies: Sample Characteristics

Study	Sample Size[1] (Coached/Total) SAT-V	SAT-M	Grade Level	School Type	Location	Year(s) Tested	SES of Sample[2]
UNCONTROLLED STUDIES							
School-based Coaching							
Pallone (1960)	100	NA	Pre-college	1 private (all male)	D.C.	1959	High
Marron (1965)	714	715	11th, 12th	10 private (all male)	D.C.	1962	High
Johnson [Atlanta & NYC sites] (1984)	117	116	11th	multiple public (all Black, urban)	NY, GA	1983-94	Low
Commercial Coaching							
Kaplan (2001)	NA	18	12th	multiple public & private	CT	1999-2000	High
Computer-based Coaching							
Coffin (1987)	18	18	11th, 12th	1 public (urban)	MA	1986-87	Low
OBSERVATIONAL STUDIES							
School-based coaching							
Dyer (1953)	225/418	225/418	12th	2 private (all male)	NR	1951-52	High
French (1955)	161/319	161/319	12th	3 public	MI, MA	1954	High
Dear (1958)	60/586	60/586	12th	multiple public & private	NJ, NY, PA	1956-57	High
Kintisch (1978)	38/76	NA	12th	1 public (suburban)	PA	1976-78	NR
Burke (1986)	50/100	50/100	11th, 12th	1 public (suburban)	GA	1984-85	Mixed
Harvey (1988)	NA	21/54	11th	2 public (urban)	GA	1987	Mixed
Schroeder (1992)	NA	59/95	NR	1 public (urban)	NY	1991-92	High
Wrinkle (1996)	18/36	NA	9th, 10th, 11th	1 public (suburban)	TX	NR	High
Commercial Coaching							
Frankel (1960)	45/90	45/90	12th	1 public (urban)	NY	1958	High
Whitla (1962)	52/104	50/100	11th	multiple public & private	MA	1959	High
FTC: BRO/BCP (1978)	556/2122	556/2122	11th, 12th	multiple public & private (urban)	NY	1974-77	Mixed
Whitla (1988)	341/1558	341/1558	12th	multiple public & private	USA	1986-7	High
Zuman [high-SES sample] (1988)	21/55	21/55	11th	multiple public (urban)	NY	1985-86	High
Smyth (1989)	200/438	200/438	12th	8 private (suburban)	MD, D.C.	1987-88	High
Snedecor (1989)	264/535	264/535	12th	10 public & private	PA	1988-89	High
Smyth (1990)	631/1132	631/1132	12th	14 private (suburban)	MD,NJ	1989	High
Powers & Rock (1999)	427/2086	427/2086	11th, 12th	multiple public & private	USA	1995-96	Mixed
Briggs (2001)	379/2554	379/2554*	11th, 12th	multiple public & private	USA	1991-92	Mixed
RANDOMIZED STUDIES							
School-based Coaching							
Roberts & Oppenheim (1966)	154/265	188/310	12th	18 public (all Black, urban & rural)	TN	1965	Low
Evans & Pike (1972)	NA	288/417	11th	12 public (urban & suburban)	NJ, OH, PA	1970-71	Mixed
Alderman & Powers (1980)	239/559	NA	11th	8 public & private	7 NE states	1977-78	Mixed
Johnson [San Francisco site] (1984)	23/35	23/35	11th	multiple public (all Black, urban)	CA	1983-94	Low
Shaw (1992)	61/122	61/122	12th	3 public (suburban)	CA	1988	Mixed
Commercial Coaching							
Zuman [low-SES sample] (1988)	16/33	16/33	11th	multiple public (urban)	NY	1985-86	Low
Computer-based Coaching							
Hopmeier (1982)	42/71*	61/93*	9th, 10th, 11th	1 public (suburban)	FA	NR	Mixed
Laschewer (1985)	13/27	13/27	11th	1 private (suburban Catholic)	NY	NR	Mixed
Curran (1988)	204/408	204/408	11th	4 private (Catholic)	MA	1986-87	Mixed
Holmes & Keffer (1995)	28/58	NA	12th	1 public (rural)	GA	1990	Mixed
McClain (1999)	40/60	40/60	12th	public (suburban)	MD	1998	Low

NOTES:
1 Samples presented here are summed across all coached and uncoached subsamples considered in given study unless otherwise noted
2 Approximate socioeconomic status (parental income, education, occupation) of sample on average according to author
NA = not applicable; NR = not reported

Table 1-2. Summary of Coaching Sample Sizes by Study Characteristics

Verbal SAT Coaching Studies	Study Sample Sizes (Coached + Uncoached Students)					
	Number of Studies	Min Size	Median Size	Max Size	Mean	SD
ALL	30	18	194	2554	533	722
Source of Study						
Published	16	58	487	2554	806	851
Unpublished	12	18	94	714	169	204
Uncontrolled Studies						
School-based Coaching	3	100	117	714	310	350
Computer-based Coaching	1	---	---	---	---	---
Observational Studies						
School-based Coaching	6	36	209	586	256	221
Commercial Coaching	10	88	834	2554	1087	971
Randomized Studies						
School-based Coaching	4	35	194	559	245	230
Commercial Coaching	1	---	---	---	---	---
Computer-based Coaching	5	27	60	408	125	159
Math SAT Coaching Studies	**Study Sample Sizes (Coached + Uncoached Students)**					
	Number of Studies	**Min Size**	**Median Size**	**Max Size**	**Mean**	**SD**
ALL	29	18	204	2554	538	733
Source of Study						
Published	14	18	487	2554	895	871
Unpublished	13	18	95	715	154	185
Uncontrolled Studies						
School-based Coaching	2	116	416	715	416	---
Commercial Coaching	1	---	---	---	---	---
Computer-based Coaching	1	---	---	---	---	---
Observational Studies						
School-based Coaching	6	54	210	586	262	214
Commercial Coaching	10	88	834	2554	1087	972
Randomized Studies						
School-based Coaching	4	35	216	417	221	174
Commercial Coaching	1	---	---	---	---	---
Computer-based Coaching	4	27	77	204	96	77

Table 1-3. SAT Coaching Studies: Treatment Characteristics

Study	Coaching Type	Coaching Duration (Hours)	
		SAT-V	SAT-M
UNCONTROLLED STUDIES			
Pallone (1960)	School-based		
Short verbal coaching		45	---
Long verbal coaching		100	---
Marron (1965)	School-based	300	300
Johnson [Atlanta, New York] (1984)	School-based	17.5	17.5
Kaplan (2001)	Commercial	---	20
Coffin (1987)	Computer-based	NR	NR
OBSERVATIONAL STUDIES			
Dyer (1953)	School-based	10	8.3
French (1955)	School-based		
Verbal & Math coaching		8.3	8.3
Vocabulary coaching		4.5	---
Dear (1958)	School-based		
Short math & verbal coaching		6	6
Long math coaching		---	12
Kintisch (1978)	School-based	30	---
Burke (1986)	School-based	52	---
Harvey (1988)	School-based	---	4
Schroeder (1992)	School-based	---	16
Wrinkle (1996)	School-based	68	---
Frankel (1960)	Commercial	15	15
Whitla (1962)	Commercial	5	5
FTC: BRO/BCP (1978)	Commercial		
Company A		20	20
Company B		12	12
Whitla (1988)	Commercial	NR	NR
Zuman [High-SES sample] (1988)	Commercial	13.5	13.5
Smyth (1989)	Commercial	NR	NR
Snedecor (1989)	Commercial	NR	NR
Smyth (1990)	Commercial	NR	NR
Powers & Rock (1999)	Commercial	15	15
Briggs (2001)	Commercial	NR	NR
RANDOMIZED STUDIES			
Roberts & Oppenheim (1966)	School-based	7.5	7.5
Evans & Pike (1972)	School-based	---	21
Alderman & Powers (1980)	School-based		
School A		7	---
School B		10	---
School C		10.5	---
School D		10	---
School E		6	---
School F		5	---
School G		11	---
School H		45	---
Shaw (1992)	School-based	4	4
Johnson [San Francisco] (1984)	School-based	17.5	17.5
Zuman [Low-SES sample] (1988)	Commercial	12	12
Hopmeier (1982)	Computer-based	3.5	3.5
Laschewer (1985)	Computer-based	4.5	4.5
Curran (1986)	Computer-based	5	5
Holmes & Keffer (1995)	Computer-based	8	---
McClain (1999)	Computer-based	18	18

NR = not reported

The coaching treatment has varied from study to study in terms of its duration and instructional characteristics. As Table 1-3 indicates, for all studies the median duration of the coaching treatment per test section was 10 hours for the SAT-V, and 12 hours for the SAT-M. Coaching duration per section ranged from as short as 3.5 hours to as long as 100, but the median length was relatively stable as a function of study source, methodological design and type of coaching, ranging from about 5 to 15 hours. Published studies of SAT-V coaching tended to involve slightly shorter amounts of instruction than unpublished studies; for SAT-M coaching, published studies featured instruction that was more than twice as long as that found in unpublished studies. With the exception of uncontrolled studies of school-based coaching and randomized studies of computer-based coaching—which respectively involved evaluations of programs with long and short amounts of student contact time—median program duration has been fairly consistent across methodological designs and coaching settings, ranging from 8 to 14 hours for either section of the test. The longest median program duration has been associated with observational studies of commercial coaching.

For the most part, the coaching techniques in the studies evaluated here involved a number of common instructional characteristics: content review, test familiarization, item practice and review, and general tips on test-taking strategies. Four school-based coaching curricula (Pallone, 1961; Kintisch, 1979; Burke, 1986; Holmes & Keffer, 1995) differed from this norm in giving a greater emphasis to the development of skills specific to the SAT-V, and a de-emphasis of item practice, review and general test-taking

strategies. Such an approach was mirrored by Schroeder (1992) in his coaching curriculum for SAT-M. The coaching curriculum developed by Evans & Pike (1973) was unique in its focus on student mastery of a single item format on the SAT-M.

Most SAT coaching studies have involved randomized experimental or observational designs as a framework for estimating coaching effects. A small number of studies have been conducted with no control group. For these studies no coaching effect can be readily estimated, but the large average SAT gains reported are suggestive. Apart from one small-scale study of commercial coaching conducted by Zuman, studies with randomized experimental designs were limited to evaluations of school and computer-based coaching. There were a number of examples where randomization may have failed to create equivalent experimental groups either because of attrition, small sample size, and/or unequal motivation levels related to the timing and type of SAT administered in the study. The Solomon Four Group design was introduced as a means of elucidating this last factor.

Observational designs have been most common in coaching studies. All studies of commercial coaching have involved observational designs. Coaching estimates from such designs will suffer to varying extent from bias because the decision to seek coaching is made by the student, not by the researcher. Attempts to control for bias due to confounding have primarily included the use of statistical matching, linear regression, or both approaches used together. In the larger observational studies, a greater number of covariates were gathered, including variables for academic grades and course-taking

patterns, socioeconomic status and even proxies for levels of student motivation. The use of these covariates with the linear regression model resulted in smaller coaching effect estimates. Only one study has estimated coaching effects using Instrumental Variables and the Heckman Model, two statistical approaches specifically intended to control for selection bias. There were a total of eight studies with observational designs that evaluated school-based coaching programs. Seven SAT-V coaching effects were estimated ranging from -2 to 56 points. The effects estimated by the early CEEB-sponsored studies tended to be small and involved coaching of 10 hours or less with fairly large student samples. The effects estimated by Kintisch, Burke and Wrinkle tended to be larger, and involved coaching programs of substantially longer duration (30, 52 and 68 hours) with small student samples. For SAT-M coaching, the five estimated effects ranged from 6 to 46 points. Smaller effects (13, 6 and 21 points) were found in studies with about 8 to 12 hours of math coaching with an emphasis on content review and item practice. The largest effect was estimated by Schroeder for 59 students enrolled in a 16-hour long course emphasizing the development of problem-solving skills.

Ten studies with observational designs estimated effects for commercial coaching programs. The resulting ten SAT-V and SAT-M coaching effects ranged from 0 to 52 and -5 to 58 points respectively. In seven of the 10 studies the combined verbal and math effects were less than 30 points. Two of the studies with the smallest estimates were the most methodologically flawed (Whitla, 1988; Snedecor, 1989).

Table 1-4. Summary of Coaching Treatment Duration by Study Characteristics

Verbal SAT Coaching Studies	Studies	Student Contact Hours with Coaching Treatment					
		Coaching Treatments	Min	Median	Max	Mean	SD
ALL	25	32	3.5	10.2	100	18.3	21.6
Source of Study							
Published	11	21	4.5	10	100	18.3	22.1
Unpublished	11	11	3.5	12	68	18.3	21.6
Uncontrolled Studies							
School-based Coaching	3	4	17.5	72.5	300	116	128
Computer-based Coaching	1	1	---	---	---	---	---
Observational Studies							
School-based Coaching	6	7	4.5	10	68	25.5	25.4
Commercial Coaching	5	6	5	14.3	20	13.4	4.9
Randomized Studies							
School-based Coaching	4	11	3.8	10	45	11.8	11.7
Commercial Coaching	1	1	---	---	---	---	---
Computer-based Coaching	5	5	3.5	5	18	7.8	6
Math SAT Coaching Studies	**Studies**	**Student Contact Hours with Coaching Treatment**					
		Coaching Treatments	**Min**	**Median**	**Max**	**Mean**	**SD**
ALL	20	22	3.5	12	21	11	6.1
Source of Study							
Published	9	11	5	12	21	13	5.7
Unpublished	11	11	3.5	5	18	9.3	6.1
Uncontrolled Studies							
School-based Coaching	2	2	17.5	159	300	159	---
Commercial Coaching	1	1	---	---	---	---	---
Computer-based Coaching	1	1	---	---	---	---	---
Observational Studies							
School-based Coaching	5	6	4	8.3	16	9.1	4.3
Commercial Coaching	5	6	5	14.2	20	13.4	4.9
Randomized Studies							
School-based Coaching	4	4	3.8	10.8	21	11.6	9
Commercial Coaching	1	1	---	---	---	---	---
Computer-based Coaching	4	4	3.5	4.75	18	7.75	6.9

NOTES:
For five observational studies of commercial coaching and one uncontrolled study of computer-based coaching, program duration was not reported.

Six of the 10 studies with observational designs estimated single effects for multiple commercial coaching programs. In only one of these studies (Powers & Rock, 1999) was

the mean and median duration of the coaching program reported.[8] The study with the largest effect estimate (Zuman, 1988) also involved the smallest sample, and must be interpreted with caution because of differences in the tests administered to treatment and control groups.

All studies with randomized designs evaluated coaching programs that were either school- or computer-based. In five studies of school-based coaching, students were supposed to be randomly assigned to treatment and control conditions. In one case (Johnson, 1984) the randomization did not hold due to substantial sample attrition. Of the remaining four studies, three produced estimates of an SAT-V effect ranging from 8 to 21 points, and three produced effect estimates for the SAT-M ranging from 6 to 17 points. None of these studies involved the use of two official SAT administrations as pre and post-tests, and there was some evidence to suggest that treatment and control groups may have been differentially motivated to give their best efforts.

Six studies with randomized experimental design intents evaluated the effectiveness of computer-based coaching. Methodological problems make these studies difficult to interpret and may explain why only one of the six was published in an academic journal. All six studies involved very small sample sizes, and only in one study were students given official administrations of the SAT. In three studies section specific

[8] According to the results of a national student survey conducted both in 1986-87 and 1995-96, (Powers, 1988) reported that 20 hours was the median duration for the commercial coaching received by students in the Powers & Rock national sample. Among commercially coached students, about half received instruction from either The Princeton Review or Kaplan. So for about half the coached students in the Powers & Rock and Briggs samples, one can reasonably infer that the nature of the treatment is a well-known and standardized commodity. For the other half of the coached samples, however, the quality of the treatment is less clear and possibly quite heterogeneous.

effects could not be estimated because of problems or inconsistencies with the study's methodological design and/or reporting of results (Coffin, 1987; Curran, 1988; McClain, 1999). In two studies (Hopmeier, 1984; Laschewer, 1986) the equivalency of experimental conditions was threatened by attrition in the control groups. Three studies (Hopmeier, 1984; Holmes & Keffer, 1995; McClain, 1999) suggest fairly large effects for computerized coaching of as much as 40 points per test section, but are based upon small, non-generalizable samples.

Revisiting the Messick & Jungeblut Analysis

An issue that merits special attention is the relationship between coaching duration and effect. It seems intuitively plausible that irrespective of coaching type or even methodological design, studies evaluating longer coaching programs should find larger coaching effects. In a thorough review of coaching studies conducted between 1953 and 1980, Messick and Jungeblut (Messick & Jungeblut, 1981) analyzed this relationship by calculating the rank correlation between program hours and coaching effect by test section. In estimating Spearman rank correlations rather than Pearson product moment correlations, less weight is placed upon the issue of the specific point estimates of coaching effects that generally suffer from bias to varying extent. Under the rank correlation, studies with large and small effects have less influence, as the set of study effects are compared only in an ordinal sense. Messick and Jungeblut found strong correlations of .77 and .71 between program duration and effect for 19 and 14 SAT-V and SAT-M coaching studies respectively.

Table 1-5. SAT Coaching Studies: Effect Estimates

Study	Design Intent	Coaching Type	Estimation Approach	Pre-test	Post-test	SAT-V		SAT-M	
						Effect	Stat Sig	Effect	Stat Sig
Dyer (1953)	Observational	School-based	Regression	Retired SAT	Official SAT	5	<.05	13	<.01
French (1955)	Observational	School-based	Regression	Retired SAT	Official SAT	18	<.01	6	<.01
Dear (1958)	Observational	School-based	Regression	Retired SAT	Official SAT	-2	NS	21	<.01
Kintisch (1972)	Observational w/ matching	School-based	NR	Official SAT	Official SAT	14	NR	No coaching	
Burke (1986)	Observational w/ matching	School-based	ANOVA	Official PSAT	Official SAT	45	<.01	No coaching	
Harvey (1988)	Observational	School-based	Regression	Retired SAT	Retired SAT	No coaching		21	NS
Schroeder (1992)	Observational	School-based	Regression	Official PSAT	Official SAT	No coaching		46	<.05
Wrinkle (1996)	Observational w/ matching	School-based	Regression	Official PSAT or SAT I	Official SATI	31	<.01	No coaching	
Frankel (1960)	Observational w/ matching	Commercial	t-test.	Official SAT	Official SAT	8	NS	9	NS
Whitla (1962)	Observational w/ matching	Commercial	ANOVA	Official SAT	Official SAT	11	NS	-5	NS
FTC: BRO/BCP (1978) [Company A]	Observational	Commercial	Regression	Official PSAT or SAT	Official SAT	28	<.01	24	<.01
FTC: BRO/BCP (1978) [Company B]	Observational	Commercial	Regression	Official PSAT or SAT	Official SAT	2	NS	4	NS
Whitla (1988)	Observational	Commercial	NR	Self-reported PSAT or SAT	Self-reported SAT	11	NR	16	NR
Zuman [high-SES] (1988)	Observational	Commercial	Regression	Official PSAT	Official SAT for treatment/ retired SAT for control	52	<.001	58	<.001
Smyth (1989)	Observational	Commercial	ANOVA	Official PSAT or SAT	Official SAT	6	NS	32	<.01
Snedecor (1989)	Observational	Commercial	NR	Self-reported PSAT or SAT	Self-reported SAT	0	NS	15	NC
Smyth (1990)	Observational	Commercial	Regression	Official PSAT or SAT	Official SAT	9	<.01	18	<.01
Powers & Rock (1999)	Observational	Commercial	Regression, PMM, IVSM, HM, Belson	Official PSAT or SAT I	Official SAT I	6	NS	18	<.01
Briggs (2001)	Observational	Commercial	Regression	Official PSAT	Official SAT	15	<.05	6	<.05

(continued next page)

Table 1-5. SAT Coaching Studies: Effect Estimates (continued)

Study	Design Intent	Coaching Type	Estimation Approach	Pre-test	Post-test	SAT-V		SAT-M	
						Effect	Stat Sig	Effect	Stat Sig
Roberts & Oppenheim (1966)	Randomized	School-based	t-test	Retired PSAT	Retired PSAT	14	NS	8	NS
Evans & Pike (1972)	Randomized	School-based	MANOVA	Retired SAT	Retired SAT	No coaching		17[a]	<.05
Alderman & Powers (1980)	Randomized	School-based	ANCOVA	Official PSAT	Retired SAT	8[b]	<.05	No coaching	
Johnson (1984)	Randomized	School-based	t-test	Shortened, retired SAT	Shortened, retired SAT	121[c]	<.05	57[c]	<.05
Shaw (1992)	Randomized	School-based	ANOVA	None	Official SAT	21	NS	6	NS
Zuman [low-SES] (1988)	Randomized	Commercial	Regression	Official PSAT	Official SAT for treatment/ retired SAT for control	-1	NS	57	<.001
Hopmeier (1982)	Randomized	Computer-based	ANOVA	None	Retired SAT (over two days)	57	<.05	37	<.05
Laschewer (1985)	Randomized Solomon 4 Group	Computer-based	MANOVA, Regression	Retired SAT	Retired SAT	-1	NS	12	NS
Holmes & Keffer (1995)	Randomized Solomon Four Group	Computer-based	F-test	Retired SAT-V	Retired SAT-V	39	<.03	No coaching	
McClain (1999)	Randomized	Computer-based	ANCOVA	Official SAT I	Official SAT I	d		d	

Coaching Effect Estimates only reported for studies involving control groups.
NR = Not Reported in Study NS = Not Significant
a Average effect across three item format treatments; these are hypothetical effects if all items on the SAT-M were the same format.
b Average effect across 8 schools; evidence of poor student motivation among control groups.
c. Interpretation of effects threatened by severe sample attrition; no attempt made to control for group differences statistically.
d. Effect not reported for each test section, only for combined sections.

Table 1-6. Coaching Duration by SAT Coaching Effect Estimate

	Verbal SAT		
	Messick & Jungeblut[1]	Messick & Jungeblut Updated[2]	Messick & Jungeblut Updated & Revised[3]
Number of Estimates	19	30	24
Rank Correlation	.712*	.459*	.312
	Math SAT		
	Messick & Jungeblut[1]	Messick & Jungeblut Updated[2]	Messick & Jungeblut Updated and Revised[3]
Number of Estimates	14	25	17
Rank Correlation	.711*	.481*	.408

* P-value of correlation under two-tailed t-test is less than .05

Basis for SAT-V Correlations:

[1]Dyer (1953), French (1955) [two program estimates], Dear (1958), Frankel (1960), Whitla (1962), Alderman & Powers (1980) [five program estimates], Pallone (1961) [two program estimates], Marron (1963) [four program estimates], FTC (1979) [two program estimates]
[2]All studies and program estimates in Messick & Jungeblut plus Kintisch (1979), Hopmeier (1982), Johnson (1984), Laschewer (1985), Burke (1986), Zuman (1988) [two program estimates], Shaw (1992), Holmes & Keffer (1995), Wrinkle (1996), Powers & Rock (1999)
[3]Excludes all program estimates from uncontrolled studies: Pallone (1961), Marron (1963)

Basis for SAT-M Correlations:

[1]Dyer (1953), French (1955), Dear (1958) [two program estimates], Frankel (1960), Whitla (1962), Evans & Pike (1973) [three program estimates], Marron (1963) [three program estimates], FTC (1979) [two program estimates]
[2]All studies and program estimates in 1 plus Hopmeier (1982), Johnson (1984), Laschewer (1985), Schroeder (1988), Schroeder (1992), Zuman (1988) [two program estimates], Shaw (1992), Powers & Rock (1999), Kaplan (2001) [two program estimates]
[3]Excludes all program estimates from uncontrolled studies: Marron (1963), Kaplan (2001)

More than twenty years have passed since the Messick and Jungeblut analysis, and we may reasonably ask whether the collection of new SAT coaching studies produced during this period conforms to the same correlational pattern. In Table 1-6 the Messick and Jungeblut analysis is replicated with three different collections of coaching studies. The first collection is identical to those used by Messick and Jungeblut in their review. The results using these studies should be identical to those found by Messick and Jungeblut. The second collection of studies adds to the first all studies conducted since

the Messick & Jungeblut review that reported coaching duration in hours and SAT effects in points per test section. The third collection of studies considers the same set as the second collection but excludes those studies that lacked a control group.

The results here do not support the association between program duration and coaching effects found by Messick and Jungeblut. For both sections of the test the rank correlation drops by about 30-35% when new studies not reviewed by Messick and Jungeblut are included in the calculation. When the new studies are included and uncontrolled studies are excluded from the calculation, the rank correlations are small and no longer statistically significant. It may still be intuitively plausible to believe that longer coaching programs will produce larger coaching effects, but there is little empirical support for this belief.

1.6 Discussion

It is tempting to present point estimates of mean and median coaching effects grouped by study characteristics as was done to summarize study samples and treatment duration in Tables 1-2 and 1-4. This is a temptation I resist. Unlike sample size and coaching duration, variables that are fixed quantities in each study, a coaching effect is a parameter that must be estimated. If multiple coaching studies were to be conducted with identical methodological designs and coaching treatments using samples drawn from the same population of students, then summarizing resulting coaching effects with a weighted average across studies would be informative. Instead, this review suggests that

the battery of coaching studies conducted since 1953 is a heterogeneous collection of methodological approaches, coaching treatments and samples.

Consider, for example, the eight observational studies on school-based coaching programs listed in Table 1-5. Can these studies be combined to produce a single point estimate for the SAT-V and SAT-M coaching effect? If a simple average was taken across studies, it would suggest a SAT-V coaching effect of 22 points and a SAT-M coaching effect of 21 points. This is very similar to the respective median effects of 18 and 21 points. Yet if these averages were weighted by sample size, these effects would change to 9 and 17 points. And none of these numbers takes into account that two of the SAT-V and one of the SAT-M effect estimates were not statistically significant.

A popular methodological approach for combining effect estimates across studies is meta-analysis. Two fundamental assumptions of meta-analysis are 1) studies being combined are independent and 2) the samples and treatments are drawn from the same underlying populations (usually with convenient normal distributions). How well would these assumptions hold for the seven SAT-V and five SAT-M coaching effects estimated from the eight observational studies in school settings? Not very well. The studies by Dyer, French and Dear are very much dependent. Each was sponsored by the CEEB, each made use of the same materials to develop its coaching curriculum, and each sought to examine questions raised by the prior study. Nor is it tenable to assume that the eight study samples and treatments reflect the same underlying population. The samples of students in the early coaching evaluations by Dyer, French and Dear are likely to

represent very different underlying student populations relative to the later studies by Burke, Schroeder and Wrinkle. In addition, the instructional emphasis of coaching programs evaluated by Kintisch, Burke and Schroeder were of a different nature than those evaluated by Dyer, French, Dear, Schroeder and Wrinkle.

The approach taken here has been to consider the continuum of coaching effect estimates suggested by each collection of studies grouped by methodological design and coaching category. This process reveals that coaching effects range as a function of study characteristics. A clear problem in the existing literature is that the studies suggesting the largest coaching effects are those with the smallest and least generalizable samples. It is unfortunate that these studies have not been replicated with bigger samples on a larger scale. On the other hand, studies suggesting the smallest coaching effects, particularly for commercial coaching, often involve a treatment condition that is rather crudely specified. There is a clear tradeoff in having on the one hand, a carefully delineated treatment, and on the other, a nationally generalizable sample. The larger the sample, the more difficult it is to gather enough information to adequately describe the full range and variability of the coaching programs under consideration.

Irrespective of the size of the sample or the definition of the treatment, when SAT coaching studies are conducted using observational designs, the causal inferences that can be drawn from such designs are threatened by bias. An optimal solution to this problem would seemingly be a reliance on randomized experimental designs. Yet these designs

have met with limited success in the realm of SAT coaching, primarily for ethical and logistical reasons.

The ethical dilemma comes in denying one group of students access to a coaching program that could improve their performance on the SAT. In the three large-scale randomized studies reviewed here (Roberts & Openheim, 1966; Evans & Pike, 1973; Alderman & Powers, 1980) three steps were taken to ameliorate ethical concerns. First, experimental samples were selected from students in the 11th grade or earlier, well before the college applications process that starts in the 12th grade. Second, treatment and control conditions were both exposed to the same coaching program, but for the treatment group the exposure was immediate while for the control group exposure was delayed. Third, both experimental groups were tested with retired SAT forms. While this approach adequately addresses ethical concerns, there is reason to suspect that it has a significant interaction with student motivation: 11th grade students are probably less motivated to do well on the SAT than 12th grade students, and students taking an unofficial SAT before receiving the coaching treatment (delayed treatment condition) are less motivated do well than students taking the test after having received the coaching treatment (immediate treatment condition). The three small-scale randomized studies that were reviewed (Shaw, 1992; Holmes & Keffer, 1995; McClain, 1999) involved samples of 12th grade students taking official SAT administrations. None of these studies included an explicit plan for delayed treatment exposure across experimental conditions, and neither author explained how they were able to circumvent the ethical problems associated with denying exposure to coaching to students in their control group.

A second class of problems endemic to randomized studies is more logistical in nature. Randomization requires a degree of control over student behavior that may not be feasible in high school settings. A researcher can do little to force treatment students to attend all scheduled coaching sessions. Worse yet, maintaining blind conditions among students is virtually impossible, particularly when the design calls for no delayed coaching for the control group. If the control condition is "doing nothing," such students may well seek out other confounding forms of test preparation during the course of a study. Such problems become exacerbated as the size of the experimental sample increases. This may explain why many randomized coaching studies have tended to be small in nature.

Randomized studies with small treatment and control samples of about 20 to 50 students pose an additional methodological difficulty. Implicit in coaching studies by Hopmeier and Shaw is the assumption that so long as students have been assigned at random to coached or uncoached conditions, any differences in their subsequent SAT scores can be attributed to the coaching treatment. Yet random assignment is more properly viewed as a means to an end: creating experimental groups equivalent in all ways except in exposure to the treatment. That this end has been attained after random assignment must be verified empirically, especially when a small sample is being assigned. The smaller the sample, the greater the probability that the experimental groups will *not* be equivalent on average, just by chance.

I do not mean to suggest that good randomized studies of SAT coaching are not possible, only that the track record of previous attempts has not been encouraging. One solution to the ethical dilemma of random assignment would be different specifications of the control condition from "doing nothing" to preparing for the SAT in a variety of ways. It is an empirical issue whether coaching per se has an effect above and beyond systematic test practice, and this is an issue that might be properly explored in a randomized design.

In the absence of randomized coaching studies, those with observational designs are clearly the next best thing. How close do coaching estimates from observational studies come to approximating those that would have been estimated under an idealized experimental design? SAT coaching studies have invoked a number of different statistical approaches attempting to estimate coaching effects as if coached and uncoached students had been randomly assigned. Among these approaches specifications of linear regression models have been applied most frequently. Unfortunately, linear regression only addresses bias due to observed variables omitted from the model, not bias caused by student self-selection. The Heckman Model has been applied as a two equation approach to purging effect estimates of bias due to both confounding and self-selection bias. The use of these approaches to estimate SAT coaching effects merits closer scrutiny. In the next section I develop formal definitions for different types of estimation bias and describe the assumptions under which the linear regression model and the Heckman Model can be used to reduce or eliminate this bias in observational settings.

CHAPTER 2: BIAS IN COACHING EFFECT ESTIMATES

In this chapter, I describe bias more explicitly in the context of observational studies of SAT coaching. Bias in estimated coaching effects may occur because of observed variables that confound the relationship between coaching and SAT performance, or because students systematically self-select themselves into coached and uncoached conditions for unobserved reasons. The focus of this chapter is on a description of two statistical approaches for reducing bias in treatment effects estimated from observational studies—the linear regression model and the Heckman Model. I present each approach in some detail, discussing key similarities and differences in model assumptions. This sets the stage for an empirical analysis of the two approaches.

2.1 Causal Inference in Randomized and Observational Settings

To make the issues clearer, I start with a model for estimating a coaching effect in a randomized experiment. Assuming that the experiment has been well-designed, the effect estimate will be unbiased. Let *y* represent a score, ranging from 200 to 800 points on the math section of the SAT. Let *COACH* be a dichotomous variable that takes a value of 1 if a student is coached, and 0 if not coached. Students are indexed by the subscript *i*. A response schedule is defined for student *i* by the pair

$$(y_{ti}, y_{ci}) \tag{1}$$

The subscript *t* stands for treatment, *c* for control. For student *i*, there are two *potential* responses: y_{ti} represents the SAT score that student *i* would obtain if coached, while y_{ci} represents the SAT score that student *i* would obtain without coaching. The notion of potentially observable responses is a fundamental part of what has been called the "Neyman-Rubin model for causal inference" (Holland, 2001). If it were possible to observe both y_{ti} and y_{ci} for the same student, it would be easy enough to calculate a unit level causal effect of coaching

$$b_i = y_{ti} - y_{ci}. \tag{2}$$

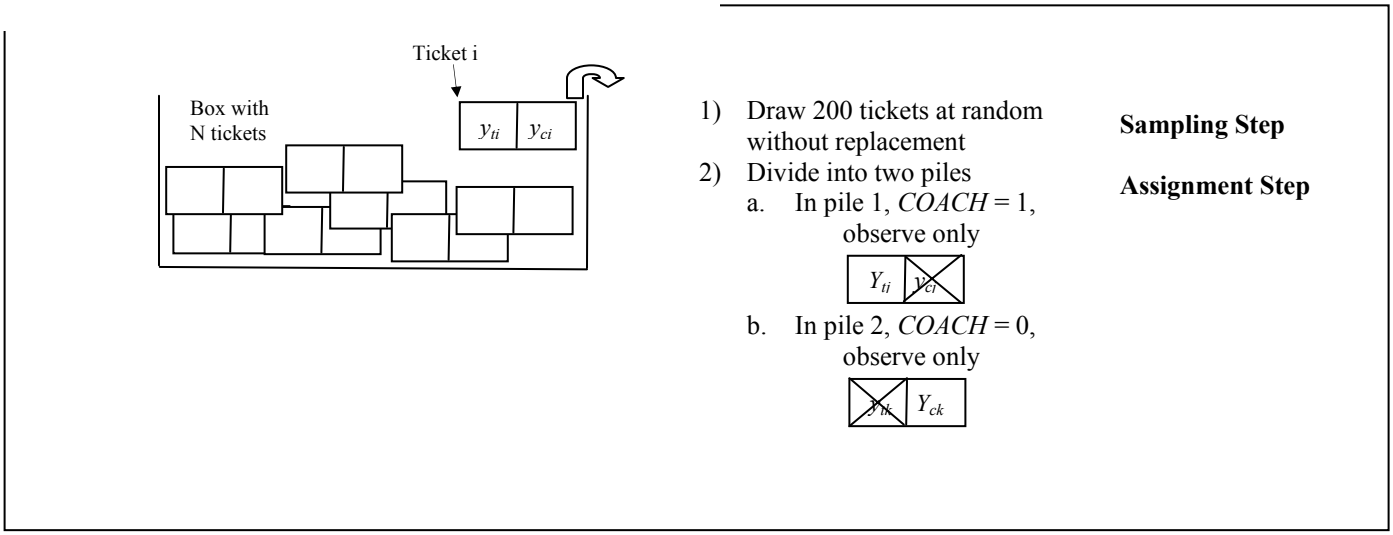
The parameter b_i is the amount by which student *i*'s SAT score increases because he or she was coached. The physical laws of nature being what they are, for student *i* we can observe y_{ci} or y_{ti}, but not both. While it is not possible to get a good estimate of a unit level causal effect, it is feasible to get an unbiased estimate of the average causal effect, *b*, where for a population of *N* students

$$b = \frac{1}{N}\sum_{i=1}^{N}\left(y_{ti} - y_{ci}\right) = \frac{1}{N}\sum_{i=1}^{N} b_i. \tag{3}$$

To make this more concrete, we can imagine an experiment with 200 students, sampled randomly from a population of *N* students. Half of the 200 students are assigned at random to a treatment condition, and the other half are assigned at random to a control condition. Those students assigned to the treatment are coached, those assigned to the control are not. After the treatment group has been coached, both groups of students take the math section of the SAT. The aim is to get an unbiased estimate of the average causal effect of coaching, as defined in Equation 3.

We model this experiment by having a box filled with $i = 1,..., N$ tickets, where each ticket represents a student in the population. Ticket i in the box contains two values on its face: y_{ti} on the left side, y_{ci} on the right. The values on the ticket represent the potentially observable responses for the corresponding student. Next, 200 tickets are drawn at random without replacement from the box, and separated into two piles. In the first pile there are 100 tickets whose treatment responses y_t are observed—these represent students that were assigned to coaching before taking the SAT. In the second pile there are 100 tickets whose control responses y_c are observed—these represent students not assigned to coaching before taking the SAT. Figure 2-1 illustrates the setup.[1]

Figure 2-1. Box Model for a Randomized Coaching Study



Let Y_{tj} be the response for the j^{th} ticket put in the treatment pile; let Y_{ck} be the response for the k^{th} ticket put in the control pile. These are observable random variables. Now, for all the tickets in the box, the average effect attributable to coaching is b,

[1] Holland has described a slightly different formulation of what follows (1986; 1988; 2001).

calculated as in (3). This value is not observable. For the sample of 200 tickets drawn from the box at random, *b* is estimated from the observable data as

$$\hat{b} = \frac{1}{100}\sum_{j=1}^{100} Y_{tj} - \frac{1}{100}\sum_{k=1}^{100} Y_{ck}. \quad (4)$$

Only a single test score is observable per ticket once it has been assigned to an experimental condition. The key point is that random assignment in an experimental design ensures that $\hat{b}$ will be an unbiased estimator of *b* because $\frac{1}{100}\sum_{j=1}^{100} Y_{tj}$ is an unbiased estimator of $\frac{1}{N}\sum_{i=1}^{N} y_{ti}$, and $\frac{1}{100}\sum_{k=1}^{100} Y_{ck}$ is an unbiased estimator of $\frac{1}{N}\sum_{i=1}^{N} y_{ci}$. Therefore, $E(\hat{b}) = b$.

In the setup presented here, the outcome of any individual person exposed to treatment or control conditions is not influenced by the assignment of other subjects to treatment or control conditions. Rubin (1986) has termed this the Stable Unit Treatment Value Assumption (SUTVA). The plausibility of SUTVA in coaching studies is often uncertain, particularly when the treatment is administered in group settings within a single school. The use of retired SAT exams and delayed treatment conditions in a number of randomized coaching studies (c.f. Roberts & Openheim, 1966; Alderman & Powers, 1980; Zuman, 1988) are examples of cases where SUTVA seems to have been violated. For example, because students are assigned to a control group taking an unofficial administration of the SAT, they are less motivated to do their best on the test relative to students assigned to a treatment group taking an official administration of the SAT.

In an observational study, subjects are not assigned to experimental conditions by the researcher. Instead, subjects are found in treatment and control groups for reasons that may be either overt or covert. Because of this, $\hat{b}$ will often be a biased estimator of b. For instance, if, for any number of reasons, the treatment response is observed for subjects (e.g. the tickets in Figure 2-1) with unusually large treatment response values, and the control response is observed for subjects with unusually small control response values, then the estimated effect will be biased upwards. If the converse is true, the estimated effect will be biased downwards. This is why causal effects estimated from observational studies may be biased. The bias may be the result of confounding due to omitted covariates, or it may result from self-selection among the subjects as a function of omitted covariates and an unmeasured latent variable. The linear regression model is presented as a statistical solution to the former problem, while the Heckman Model is presented as a solution to both. Of course, the "solution" is valid only under strong assumptions, which will be discussed.

2.2 Statistical Solutions to Bias

The Linear Regression Model

The objective in a randomized study is to demonstrate the strength of a hypothesized causal relationship between, for example, coaching status (*COACH*) and SAT scores (*Y*) as in Figure 2-2.

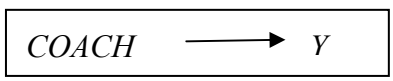


Figure 2-2. Causation

In an observational study with the same objective, it is usually conceivable, and often highly likely that other covariates may confound the relationship between the treatment and the outcome, as in figure 2-3.

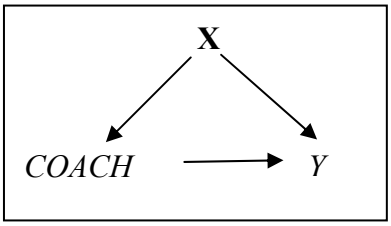


Figure 2-3. Confounding

In Figure 2-3, **X** represents a set of covariates that might include each student's pre-coaching SAT score and socioeconomic status. These covariates may influence post-coaching performance on the SAT and also be correlated with coaching status. The relationship between *Y* and *COACH* is thus confounded by **X**. A statistical approach frequently applied to correct for the possibility of bias due to confounding is linear regression.[2] In what follows a specialized version of linear regression is presented to facilitate a comparison with the Heckman Model.[3]

Consider the following behavioral model:

$$f_i(COACH) = a + bCOACH + \mathbf{X}_i\mathbf{c} + \sigma\varepsilon_i \tag{5}$$

$$COACH_i = 1 \Leftrightarrow \alpha + \mathbf{X}_i\boldsymbol{\gamma} + \delta_i > 0\,. \tag{6}$$

The model consists of a *response schedule* (5) and a *selection function* (6). In the response schedule, a student's potentially observable SAT score is a function of *COACH*.

[2] The analysis of variance (ANOVA) and analysis of covariance (ANCOVA) models used in previous coaching studies described in chapter 1 can be shown to be special cases of the linear regression model.
[3] The specialization comes primarily from restrictions on the distribution of the unobservable error terms. Linear regression could be used to make causal inferences under more general assumptions. See for example, Freedman, 2002 and Holland, 2001.

Two different scores are possible for student i, depending on whether $COACH = 1$ or 0. The variable *COACH* is in theory manipulable—if its value is changed, the SAT score subsequently observed for student i will change as well (unless, of course, there is no coaching effect). The observed covariates in the vector $\mathbf{X}_i$ are fixed characteristics of each student—they cannot be manipulated by the researcher. The response schedule assumes a linear relationship between the variable *COACH* and the SAT score, with a constant effect across individuals, represented by the parameter b. Likewise, the effect of $\mathbf{X}_i$ is linear, and $\mathbf{c}$ is the same for all students. The error term $\sigma\varepsilon_i$ represents the deviation of student i's SAT score from its expected value. In an experimental setting, the observed value of *COACH* for student i would be assigned by the researcher with a known probability. Here, the observed value of *COACH* is assumed to be governed by the selection function. This selection function will be described in more detail in the context of the Heckman Model. For now it suffices to note that the function implies that student i's decision to seek coaching depends on observable covariates in the vector $\mathbf{X}_i$, and on the unobservable (i.e. latent) covariate δ_i.

In an observational study, the researcher observes the triple $\{Y_i, COACH_i, \mathbf{X}_i\}$, where $COACH_i$ is determined by the selection function (6), and

$$Y_i = f_i(COACH_i) = a + bCOACH_i + \mathbf{X}_i\mathbf{c} + \sigma\varepsilon_i \qquad (7)$$

is determined by the response schedule (5). Further statistical assumptions must be made:

i) $(\varepsilon_i, \delta_i)$ are independently and identically distributed (iid) in i with a standard normal distribution;

ii) $\{\mathbf{X}_i: i = 1, \ldots, n\}$ is independent of $\{\varepsilon_i, \delta_i: i = 1, \ldots, n\}$.

iii) δ_i and ε_i are independent within student i.

According to these assumptions, the data generated from (5) and (6) have the feature that $\{(\mathbf{X}_i, \delta_i): i = 1, \ldots, n\}$ is independent of $\{\varepsilon_i: i = 1, \ldots, n\}$. It follows therefore that $\{(X_i, COACH_i) : i = 1, \ldots, n\}$ is independent of $\{\varepsilon_i: i = 1, \ldots, n\}$. Thus, $COACH_i$ and $\mathbf{X}_i$ are exogenous, so ordinary least squares (OLS) can be used to get unbiased estimates for the parameters a, b and $\mathbf{c}$, by running a linear regression of Y_i on a constant, $COACH_i$ and $\mathbf{X}_i$.

In making causal inferences about the effectiveness of coaching, b is the parameter of interest, with a causal interpretation because of Equation 5. In other presentations of unbiased parameter estimation using linear regression, it is assumed that

$$E(\varepsilon_i \mid COACH_i, \mathbf{X}_i) = 0. \tag{8}$$

This follows from assumptions i, ii and iii.

With respect to the ticket model of Figure 2-1, the linear regression adjustment for $\mathbf{X}_i$ is meant as a replacement for the random assignment step. However, the assumptions that coaching status and covariate values are independent of the error terms, and that error terms are independent within and across students, are clearly rather difficult to defend in the absence of a theoretical understanding of the causal mechanism at work. A common criticism among statisticians is that the plausibility of such assumptions in observational settings is seldom given adequate consideration.[4]

[4] Some exchanges along these lines can be found in Freedman (1987; 1995). For a different interpretation of the ε term in line with the Neyman-Rubin model for causal inference, see Holland, 2001.

Implicit in estimating the effect of coaching by linear regression is that any differences between coached and uncoached students related to SAT performance are accounted for by $\mathbf{X}_i$: bias is a function of variables omitted from the regression equation. To see this more clearly, consider Equation 7 presented in matrix format. Let **M** be a matrix containing the constant term and observed values of $COACH_i$ for $i = 1, \ldots, p$ students in a given study. Let the matrix **X** represent the collection of covariate values $\mathbf{X}_i$ for $i = 1, \ldots, p$. Similarly, the SAT score Y_i and the error term ε_i are collected into the vectors **Y** and **ε**. Then, in matrix format

$$\mathbf{Y} = \mathbf{M}\mathbf{b} + \mathbf{X}\mathbf{c} + \boldsymbol{\varepsilon}, \tag{9}$$

where $\mathbf{b} = [a \; b]$. If instead of the regression implied by Equation 9, the researcher regressed **Y** on **M**, omitting the confounding variables **X**, then the OLS estimate of the average coaching effect would be biased, since

$$\begin{aligned} \hat{\mathbf{b}} &= \left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{Y} \\ &= \left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{M}\mathbf{b} + \left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{X}\mathbf{c} + \left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\boldsymbol{\varepsilon} \\ E(\hat{\mathbf{b}} \mid \mathbf{M}, \mathbf{X}) &= \mathbf{b} + \left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{X}\mathbf{c}. \end{aligned} \tag{10}$$

The estimate of **b** is biased by $\left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{X}\mathbf{c}$. This is "omitted variable" bias.

Linear regression is useful because it reduces bias caused by confounding variables. For example, students who do well on the PSAT may be less likely to get coached, but more likely to do well on the SAT. If this is the case, omitting PSAT scores as a covariate in the regression equation will result in a biased coaching effect estimate. A key point is that omitted variable bias is not the same thing as "selection bias."

Selection bias occurs when the variable $COACH_i$ is endogenous—correlated to a latent covariate that has not been measured. If this is the case, the linear regression model generally will not produce unbiased estimates of the coaching effect—even if all the relevant observed covariates are included. The so-called "Heckman Model" (Heckman, 1978; 1979; Heckman & Robb, 1986; Greene, 1993), named after economist James Heckman who first developed the approach, has been applied in certain contexts as a general strategy for estimating a causal parameter in the presence of selection bias.

The Heckman Model

Under the Heckman Model, the variables in the regression equation are allowed to be correlated with the error term ε_i. In other words, the variables may be endogenous, so any causal parameter will suffer from selection bias.[5] In what follows the Heckman Model is illustrated in the context of an observational coaching study. I first describe the general approach and then present the mathematical details.

The motivation for the Heckman approach is a behavioral model similar to the one behind the use of linear regression:

$$f_i(COACH) = a + bCOACH + \mathbf{X}_i\mathbf{c} + \sigma\varepsilon_i \tag{11}$$

$$COACH_i = 1 \Leftrightarrow \alpha + \mathbf{X}_i\boldsymbol{\gamma} + \delta_i > 0. \tag{12}$$

[5] In this context, the term "selection bias" is being used synonymously with the term "endogeneity bias."

Everything in the causal relationship is the same as the one specified using the response schedule and selection function in (5) and (6). Observed SAT scores are again generated as

$$Y_i = f_i(COACH_i) = a + bCOACH_i + \mathbf{X}_i\mathbf{c} + \sigma\varepsilon_i, \quad (13)$$

where $COACH_i$ is determined by Equation 12. Assumptions i and ii are also retained:

i) $(\varepsilon_i, \delta_i)$ are iid in i with a standard normal distribution;

ii) $\{\mathbf{X}_i: i = 1, \ldots, n\}$ is independent of $\{\varepsilon_i, \delta_i: i = 1, \ldots, n\}$.

What has changed in the behavioral model? The critical change is that assumption iii is dropped. It is relaxed to allow ε_i and δ_i to be correlated. This introduces a new parameter, ρ, into the model. Under assumption iii of the linear regression model, the correlation ρ between ε_i and δ_i was restricted to 0. For the Heckman Model, ρ is allowed to take on any value between -1 and 1.

The causal parameter of interest is still b. Note that if ε_i and δ_i were not correlated, e.g. $\rho = 0$, then there would be no selection bias problem—linear regression could be used to correct for confounding and estimate an unbiased coaching effect. Intuitively, $\rho \neq 0$ will be the case if an unobserved reason why students decide to get coached is correlated with an unobserved reason that students perform well on the SAT. For example, suppose students with more "grit" are the ones most likely to get coached. At the same time, suppose students with more "moxie" will perform better on the SAT. (I offer no definition of grit and moxie; the two are distinguishable but latent.) While the linear regression model would assume that grit (i.e. δ_i) and moxie (i.e. ε_i) are independent, the Heckman Model allows for the possibility that they are correlated.

Given Equations 11-12 and assumptions i and ii, if $\rho \neq 0$ and the parameters a, b and **c** were estimated by regressing Y_i on a constant, $COACH_i$ and $\mathbf{X}_i$, the estimates would be biased. Because $\rho \neq 0$, the variable $COACH_i$ is endogenous, and $E(\varepsilon_i \mid COACH_i, \mathbf{X}_i) \neq 0$. The Heckman Model strategy is to get an estimate for this term, and then treat it as an observable confounder. Let $\lambda_i = E(\varepsilon_i \mid COACH_i, \mathbf{X}_i)$. If this value were known for student i, then regressing Y_i on a constant, $COACH_i$, $\mathbf{X}_i$ and λ_i would produce unbiased parameter estimates for a, b, **c** and h, where h is the regression coefficient associated with λ_i. Now, $E(\varepsilon_i - \lambda_i \mid COACH_i, \mathbf{X}_i) = 0$. If the assumptions of the Heckman Model are to be believed, we have controlled for selection bias in the estimate of b.

In practice, λ_i is not known, but given the assumption that ε_i and δ_i have standard normal distributions, $\hat{\lambda}_i$ can be calculated as a function of the estimated parameters $\hat{\alpha}$ and $\hat{\boldsymbol{\gamma}}$ in the selection function (12). Now, assuming that all confounding in the relationship between Y_i and $COACH_i$ is due to $\mathbf{X}_i$, and all selection bias is due to $\hat{\lambda}$, then by regressing Y_i on a constant, $COACH_i$, $\mathbf{X}_i$ and $\hat{\lambda}$ we have almost controlled for bias in the estimate of b due to both confounding and self-selection. Heckman (1979) has shown that $\hat{b}$ will converge to b asymptotically, so $\hat{b}$ will be biased but consistent. The details of the Heckman Model for the coaching application are sketched out below.

The starting point for the Heckman Model is the selection function describing the way students decide whether or not they will seek coaching. The vector $\mathbf{X}_i$ contains observable covariates related to the probability that a student is coached.[6] Latent covariates enter the picture through δ_i. The term δ_i is cast as an unmeasured latent continuous random variable with an assumed standard normal distribution. Student *i*'s decision to seek coaching is determined by a linear combination of the measured and unmeasured covariates represented by $\mathbf{X}_i$ and δ_i. The selection function specifies that if $\alpha + \mathbf{X}_i\boldsymbol{\gamma} + \delta_i > 0$, student *i* will be coached. Otherwise, student *i* will not be coached. Given assumptions i and ii, another way of writing the selection function is

$$\begin{aligned} P(COACH_i = 1 \mid \mathbf{X}_i) &= P(\alpha + \mathbf{X}_i\boldsymbol{\gamma} + \delta_i > 0 \mid \mathbf{X}_i) \\ &= P(-\delta_i < \alpha + \mathbf{X}_i\boldsymbol{\gamma} \mid \mathbf{X}_i) \\ &= \Phi(\alpha + \mathbf{X}_i\boldsymbol{\gamma}), \end{aligned} \tag{14}$$

where Φ represents the standard normal cumulative distribution function. Given all the $\mathbf{X}_i$'s, the $COACH_i$'s are assumed to be independent, so Equation 14 constitutes what is known as the probit model.

The following theorem[7] helps explain how the Heckman Model goes from specifying a selection function to getting an estimate for the bias term, $E(\varepsilon_i \mid \mathbf{X}_i, COACH_i)$.

[6] In this setup, for the sake of parsimony, the covariates represented in $\mathbf{X}_i$ are the same in both Equation 11 and 12. This is not a restriction of the Heckman Model. It is possible for the covariates in the selection function to contain unique covariates related to the probability a student is coached, but not to subsequent SAT performance. I discuss this issue more extensively in chapter 4.

[7] For a proof of a more general version of this theorem, see Johnson & Kotz, 1970, 112-113. For a description consistent with the Heckman Model, see Greene, 1990, 682-689.

Theorem I

Let t represent the point in the distribution at which a continuous random variable $v \sim N(0, 1)$ is truncated. When the truncation is from below

$$E(v \mid v > t) = \lambda(t) \tag{15}$$

$$Var(v \mid v > t) = 1 - \lambda(t)\left[\lambda(t) - t\right], \tag{16}$$

where

$$\lambda(t) = \frac{\phi(t)}{1 - \Phi(t)} \tag{17}$$

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} \tag{18}$$

$$\Phi(t) = \int_{-\infty}^{t} \phi\left(z\right) dz. \tag{19}$$

$\lambda(t)$ is commonly referred to as the Inverse Mills Ratio or Hazard Function. It is the ratio of the standard normal density function (18) to the normal cumulative distribution function (19). When the truncation in v is from above, then by symmetry of the normal distribution,

$$E(v \mid v \le t) = \lambda(t) = -\frac{\phi(t)}{\Phi(t)}. \tag{20}$$

Our goal is to estimate a value for the bias term $E(\varepsilon_i \mid \mathbf{X}_i, COACH_i)$ for student i. Fix a value $\mathbf{x}_i$ for $\mathbf{X}_i$. The selection bias term can be decomposed into two parts $E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, COACH_i = 1)$ and $E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, COACH_i = 0)$. Given our behavioral model, and the condition that $COACH_i = 1$, it follows that δ_i no longer has a normal distribution, but a truncated normal distribution. We use Theorem I to compute the

conditional expectation of δ_i, which will be $E(\delta_i \mid \alpha + \mathbf{X}_i\boldsymbol{\gamma} + \delta_i > 0)$. Similarly, under the condition that $COACH_i = 0$, it follows that δ_i again has a conditionally truncated distribution—this time the truncation is from above. Now the conditional expectation of δ_i is $E(\delta_i \mid \alpha + \mathbf{X}_i\boldsymbol{\gamma} + \delta_i \leq 0)$. The next step is to compute the conditional expectation of ε_i, given $\mathbf{X}_i$ and $COACH_i$.

Under the Heckman Model, ε_i and δ_i have correlation ρ. Let ξ_i be a random variable equal to $(\varepsilon_i - \rho\delta_i)\Big/\sqrt{1-\rho^2}$. It follows from this definition that ξ_i has an expected value of 0 and is independent of δ_i. We can think of ξ_i as the random variable that picks up the variance left unexplained if ε_i is regressed on δ_i. Now we can relate ε_i to δ_i and ξ_i:

$$\varepsilon_i = \rho\delta_i + \sqrt{1-\rho^2}\xi_i. \tag{21}$$

Let $s_i = \alpha + \mathbf{X}_i\boldsymbol{\gamma}$. It follows from Equations 21 and 12 that

$$\begin{aligned} E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, COACH_i = 1) &= E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, s_i + \delta_i > 0) \\ &= \rho E(\delta_i \mid s_i + \delta_i > 0) \\ &= \rho E(\delta_i \mid \delta_i > -s_i). \end{aligned} \tag{22}$$

Note that ξ_i drops out of the equation because its conditional expectation is 0 by definition. The task is to evaluate the conditional expectation on the right side of (22). Taking advantage of the symmetry of the normal distribution and applying Theorem I leads to the Inverse Mills Ratio,

$$E(\delta_i \mid \delta_i > s_i) = \frac{\phi(s_i)}{1-\Phi(s_i)}. \tag{23}$$

Likewise,

$$\begin{aligned} E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, COACH_i = 0) &= E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, s_i + \delta_i \le 0) \\ &= \rho E(\delta_i \mid s_i + \delta_i \le 0) \\ &= \rho E(\delta_i \mid \delta_i \le -s_i). \end{aligned} \tag{24}$$

This again yields the Inverse Mills Ratio

$$E(\delta_i \mid \delta_i \le s_i) = -\frac{\phi(s_i)}{\Phi(s_i)}. \tag{25}$$

It follows from (22-25) that

$$E(\varepsilon_i \mid \mathbf{X}_i, COACH_i) = \rho\lambda_i(COACH_i, s_i), \tag{26}$$

where

$$\lambda_i(COACH_i, s_i) = COACH_i\left(\frac{\phi(s_i)}{1-\Phi(s_i)}\right) + (1 - COACH_i)\frac{-\phi(s_i)}{\Phi(s_i)}. \tag{27}$$

$\lambda_i(COACH_i, s_i)$ is a specific value for student *i*. While $\lambda_i(COACH_i, s_i)$ is not directly observable, it is estimable given the assumptions of the Heckman Model. $\lambda_i(COACH_i, \hat{s}_i)$ is computed using (23), (25) and (27) after estimating parameter values for α and $\boldsymbol{\gamma}$ in (14) via maximum likelihood.

The behavioral model of (11) and (12) leads to

$$Y_i = a + bCOACH_i + \mathbf{X}_i\mathbf{c} + h\lambda_i(COACH_i, \hat{s}_i) + \varepsilon_i^* \tag{28}$$

where $\varepsilon_i^* = \sigma\varepsilon_i - h\lambda_i(COACH_i, \hat{s}_i)$. The causal parameter of interest is still *b*. The parameter *h* associated with $\lambda_i(COACH_i, \hat{s}_i)$ in Equation 28 is equal to $\sigma\rho$. Consistent estimates for *b* and *h* will be obtained by regressing Y_i on a constant, $COACH_i$, $\mathbf{X}_i$ and $\lambda_i(COACH_i, \hat{s}_i)$. Note that while it is $\hat{\sigma}\hat{\rho}$ that is estimated by $\hat{h}$, if we wanted an

estimate for $\hat{\rho}$, we could obtain it by dividing $\hat{h}$ by $\hat{\sigma}$, where $\hat{\sigma}$ is estimated as a function of residuals from the regression equation. Because the conditional variance of ε_i^* depends on $\mathbf{X}_i$, a regression fit by OLS will be heteroskedastic. Estimates for *a*, *b*, **c** and *h* will be consistent, but inefficient. The standard errors estimated using OLS will be incorrect. A regression fit by Generalized Least Squares (GLS) would solve the latter problem (Greene, 1981). If the GLS estimate for *h* is statistically significant, this suggests that had *b* been estimated directly using linear regression without the Heckman correction, the estimate would have contained selection bias.

Finally, note that $\lambda_i(COACH_i, \hat{s}_i)$ essentially adds an interaction term consisting of $COACH_i$ and the Inverse Mills Ratio to the main effect for $COACH_i$ in the regression equation. The difference in expected SAT scores between coached and uncoached students will be $\hat{b} + \hat{h}\left[\frac{\phi(\hat{\alpha}_i + \mathbf{X}_i\hat{\boldsymbol{\gamma}})}{\Phi(\hat{\alpha}_i + \mathbf{X}_i\hat{\boldsymbol{\gamma}})(1-\Phi(\hat{\alpha}_i + \mathbf{X}_i\hat{\boldsymbol{\gamma}}))}\right]$. The effect of coaching estimated under the linear regression model is the combination of these two terms: the main coaching effect and the coaching by Inverse Mills Ratio interaction. The term in brackets will always be positive. The estimate $\hat{h}$ has been defined as the product of $\hat{\sigma}$ and $\hat{\rho}$. Since $\hat{\sigma}$ is always positive, if $\hat{\rho}$ is positive, this suggests that the coaching effect estimate from the linear regression model would be biased upwards. If $\hat{\rho}$ is negative, it suggests that the coaching effect estimate from the linear regression model would be biased downwards.

To summarize, the Heckman Model as applied to coaching studies has two main steps.

1. Specify a selection function for coaching status and estimate the parameters using maximum likelihood. Use these estimated parameters, and the assumed normal distributions of the response schedule and the selection function to compute the Inverse Mills Ratio when $COACH_i = 1$ and when $COACH_i = 0$.
2. Include $\lambda_i(COACH_i, \hat{s}_i)$ in a linear regression equation as a covariate. Estimate the coaching effect, $\hat{b}$ and the selection bias parameter, $\hat{h}$ (i.e. $\hat{\sigma}\hat{\rho}$) using OLS or GLS.

2.3 Comparing Linear Regression and the Heckman Model

When a causal effect is estimated in an observational study, its interpretation is always threatened by the possibility of bias. Linear regression and the Heckman Model are two statistical approaches that are commonly used to reduce bias in observational settings. Linear regression operates under the principal assumption that bias occurs because confounding variables were omitted from the regression equation. The Heckman Model assumes that bias comes from confounding caused by omitted variables, and more specifically, from endogeneity caused by the self-selection of subjects into treatment conditions. As presented here, the Heckman Model can be viewed as a two-step "correction" to the linear regression model in the presence of selection bias.[8]

[8] The Heckman Model can also be implemented as a one-step approach when estimation is done by maximum likelihood, but the two-step approach is more common in the applied literature (Vella, 1998).

Both linear regression and the Heckman Model assume that the functional form of the causal relationship between outcome, treatment and covariates is linear. In the context of observational studies where the coaching variable is dichotomous, the linearity assumption is violated if some or all of the covariates in $\mathbf{X}_i$ have a nonlinear relationship with Y_i. If the linearity assumption is incorrect, a coaching effect will be estimated as the difference between the wrong two regression surfaces. Both statistical approaches also typically make a constancy constraint, i.e. $b_i = b$, stipulating that person i = 1, ... , N is affected by the treatment in the same way. The constancy constraint is violated, for example, when certain types of students benefit significantly more or less from coaching. Indeed, interaction effects between coaching and student characteristics have been analyzed from the very earliest coaching study by Dyer (1953) to the more recent study by Briggs (2001). If the constancy constraint is wrong, then causal inferences about "the" coaching effect may be misleading. Parametric assumptions such as linearity and constancy have been discussed in more detail in the context of an alternative approach to causal inference in observational settings known as the Propensity Matching Model. For details, see Rosenbaum & Rubin, 1983; 1984 and Rosenbaum, 2002.

A key difference between the two approaches is the relaxation of the independence assumption between ε_i and δ_i when going from linear regression to the Heckman Model. Normality was assumed for ε_i and δ_i throughout in order to focus attention on this difference. If normality does not hold, then the Heckman Model as described here falls apart as a correction for the selection bias problem. Normality is a necessary condition for consistent estimation under the Heckman Model, but not for

linear regression. As long as the ε_i are iid, ε_i and δ_i are independent within student i, confounding covariates are included in the model, and the functional form is in fact linear, then linear regression will produce unbiased causal effect estimates even when the distribution of ε_i is non-normal.

The linear regression model can serve purely descriptive or predictive purposes, with the well-known disclaimer that association does not imply causation. This chapter has presented the rather strong assumptions necessary before association does imply causation. A clear problem in observational settings is that it is almost never realistic to assume that the bias in causal effect estimates is due solely to confounding from measured covariates available to the investigating researcher. Generally speaking, the use of linear regression with covariates will at best only reduce omitted variable bias by an unknown amount, not control or correct for it unequivocally.

Unlike linear regression, the Heckman Model is an approach specifically developed in the attempt to make unbiased causal inferences in observational settings. Because of the strong assumptions that underlie the model, its usefulness has been questioned by some statisticians (Wainer, 1986) and econometricians (Goldberger, 1983; Little, 1985). In one unusual case (Lalonde, 1986), the causal effect estimates from a Heckman Model were put to the empirical test—and the results were not encouraging. Lalonde gained access to data from a federally randomized experiment conducted to determine the average effect of a job training program. The effect was estimated by comparing the post-treatment incomes of subjects in an experimental treatment group to

the post-treatment incomes of an experimental control group. Based on the findings from the randomized experiment, the average effect of the program appeared to be a little over $800, with a standard error of about $300. Lalonde attempted to recreate these results by substituting non-experimental control groups for the experimental control, and using a Heckman Model with different specifications of the selection function to approximate the result of the randomized experiment. The results showed that when using four different selection function specifications while holding constant gender and type of non-experimental control groups, the estimated effect of the program varied from $10 to $670, and in few cases was the estimated effect within a standard error of the experimental estimate. Lalonde did not however, conclude that the Heckman Model's apparent sensitivity to alternate selection function specifications threatened the usefulness of the model, nor did he speculate as to what drove this sensitivity.

Powers & Rock (1999) employed both linear regression and the Heckman Model to estimate a causal effect for SAT coaching in an observational setting. The findings from this study, summarized in chapter 1, were that the two approaches produced relatively similar estimates of coaching effects, and that neither approach produced effect estimates considerably different from a baseline comparison with only pre-treatment test scores as covariates. In a footnote Powers & Rock reported that their Heckman Model estimates had been sensitive to specifications of the selection function, but no details were provided.

The relationship between the specification of the selection function and effect estimates would seem to merit closer attention, because as a procedure, the Heckman Model offers no guidance as to the covariates that should be included in its selection function. It is only assumed that $\{\mathbf{X}_i: i = 1, \ldots, n\}$ is independent of $\{\delta_i: i = 1, \ldots, n\}$. As a matter of identifiability, it does not matter whether the covariates in the selection function are different from those in the response schedule. The Inverse Mills Ratio is identified through its nonlinear relationship to $\mathbf{X}_i$. In some illustrations of the Heckman Model, it has been suggested that the covariates in the selection function should contain one or more variables related to the probability of treatment selection, but excluded from outcome prediction (e.g. Lalonde, 1986; Greene, 1993). In other illustrations, only covariates excluded from outcome prediction have been included in the selection function (e.g. STATA, 2000). Ideally, it would seem the choice of covariates should be based on some theoretical understanding of the selection mechanism.

In the following two chapters, I demonstrate that different choices of covariates for inclusion in the selection function can have a dramatic impact on the estimated coaching effect. I also compare coaching effects estimated by the Heckman Model to those estimated by linear regression, and consider whether either approach produces estimates different from a baseline model with no covariate adjustment. In chapter 3 I describe the data that will be used to estimate SAT coaching effects. In chapter 4 I present the resulting analysis.

CHAPTER 3: THE NELS DATA

In chapter 1 the history of SAT coaching studies was presented as a heterogeneous mix of investigations, varying sometimes dramatically in terms of samples, treatments and methodological designs. The most frequently used design has been the observational study, but causal inferences from such studies are threatened by estimation bias. In chapter 2, related statistical approaches used to control for bias in the context of coaching studies—linear regression and the Heckman Model—were described in some detail. In this chapter, I describe the data that will be used for empirical analyses of these approaches.

3.1 The Structure of NELS

The National Education Longitudinal Study of 1988 (NELS:88, hereafter referred to as "NELS") tracks a nationally representative sample of American students from the 8th grade through high school and beyond. The NELS data can be used for an observational evaluation of coaching effectiveness because it contains SAT scores and information about how students prepared for the SAT. Figure 3-1 summarizes the structure of NELS. A panel of nearly 15,000 students completed survey questionnaires in the second two waves of NELS in 1990 and 1992. In each wave of the survey, students were given questionnaires with hundreds of questions relating to their experiences in and outside of school. One of these questions asked students to select from a range of options

describing how they had prepared to take the SAT. In addition to student questionnaire responses, high school transcripts were collected. Each transcript includes information on student grades, course taking patterns, school demographics, and college admission test scores. Other sources of data found in NELS are standardized tests in math, reading, civics and science administered in each of the first three waves of the survey, as well as questionnaires given to parents, teachers and school administrators.

3.2 The NELS Sample

The sampling structure of NELS in the base year of the survey involved two stages. In the first stage, 1,052 schools (815 public, 237 private) with eighth grade students were sampled from about 40,000 existing middle schools nationally. These schools were selected with probabilities proportional to their estimated eighth grade enrollment from sampling strata that divided the United States into eight geographic regions.[1] In the second stage of the base year survey, 26 students were randomly selected per school. This resulted in a sample of roughly 26,000 eighth grade students from the 1987-88 school year.[2] The NELS sample has a stratified cluster design because private schools and Hispanic, Asian and American-Indian students were intentionally oversampled, and all sampled students are clustered within schools.

[1] Schools were excluded from the potential base year sample if they were already participating in the National Assessment of Educational Progress (NAEP) sample, were Bureau of Indian Affairs schools, special educational schools for the handicapped, area vocational schools, or schools for dependents of US personnel overseas. Other ineligible schools for sample selection were those that had closed or had enrolled no eighth graders as of the spring of 1988.

[2] Enrolled eighth grade students were excluded from the potential base year sample if they were deemed by school administrators as unable to complete the NELS questionnaires due to physical disabilities, mental disabilities, or problems with the English language.

Figure 3-1. Summary of NELS:88 Survey Waves

	BASE YEAR (BY)	FIRST FOLLOW-UP (F1)	SECOND FOLLOW-UP (F2)	THIRD FOLLOW-UP (F3)	FOURTH FOLLOW-UP (F4)
Data Collection:	Spring Term 1988	Spring Term 1990	Spring Term 1992	Spring 1994	Spring 2000
Grades Included:	Grade 8	Modal grade= 10	Modal grade= 12	H.S. + 2 years	H.S. + 8 years
Full Panel Samples:	17,424				
		16,749			
			16,489		
			13,120		
Cohort:	**students:** questionnaire (410 vars), tests (4)	**students:** questionnaire (694 vars), tests (4) **dropouts:** questionnaire (561 vars),tests (4)	**students:** questionnaire (786 vars, tests (4), H.S. transcripts **dropouts:** questionnaire, (577 vars), tests (4)	**all individuals:** questionnaire	**all individuals:** questionnaire
Parents:	questionnaire (331 vars)	none	questionnaire (423 vars)	none	none
School:	questionnaire (211 vars)	questionnaire (832 vars)	questionnaire (385 vars)	none	none
Teachers:	two teachers per student (taken from English, social studies, mathematics, or science). (238 vars)	two teachers per student (taken from English, social studies, mathematics, or science). (466 vars)	one teacher per student (taken from math or science). (420 vars)	none	none

The nationally representative nature of the base year sample was maintained in the first (F1) and second (F2) NELS follow-up surveys. As of the F1 follow-up, the 26,000 base year respondents had dispersed into roughly 4,000 high schools. About 75% of these students were attending one of 908 high schools, and these schools and students were sampled for the F1 follow-up with certainty. Another 600 schools were than selected with sampling probabilities proportional to the number of students from the base year sample attending the school. This, along with sample freshening, produced an F1 sample of about 21,000 students.[3] All students in the F1 sample were retained for the F2 sample, including those students who had dropped out of school. By the F2 follow-up, students from the F1 sample had dispersed—either because their families had moved or for other reasons—into a total of 2,258 schools. Financial and logistical constraints limited the amount of school-level contextual data (i.e. data from teachers, school administrators and transcripts) that could be collected for the F2 sample to a maximum of 1,500 schools. Of the total 2,258 schools available, 1,030 schools contained more than four students from the F1 sample. Contextual data from these schools was gathered with certainty. Contextual data from the remaining 470 schools was sampled with probabilities proportional to the number of F1 sample students in the school. After another round of sample freshening, the F2 follow-up produced a sample of 21,188 students.

[3] Freshening is a sampling procedure that takes into account the potential influx of new students into the cross-sectional population. For details, see NELS F2 User's Manual, p. 30-40.

For the analysis that follows, attention is focused on the NELS panel sample of students who completed surveys in the F1 and F2 follow-ups, and for whom transcript data was collected. This comprises an F1-F2 panel of 14,617 students. (For more information on the NELS sampling design, see the NELS Second Follow-up Student Component Data File User's Manual, 1995.)

Population Weights and Design Effects

Sample weights have been constructed and made available as part of the NELS database to allow for population inferences from the longitudinal and cross-sectional samples. To make the F1-F2 panel sample representative of the national population of 10^{th} to 12^{th} grade students during the 1990 to 1992 period, the NELS weight *F2TRP2WT* is applied to all statistical analyses that follow. Use of this weight indicates that the F1-F2 NELS panel is representative of an underlying population of about three million students.

Since the NELS F1-F2 panel is generated from a stratified cluster sample (SCS), the estimated standard errors of population parameters (e.g. the mean for a particular transcript variable or survey item response) will generally be larger than the standard errors that would be estimated had the panel been generated from a simple random sample (SRS). The ratio of these two standard error estimates for any given parameter corresponding to the variable *j* is known as a *design effect* ($DEFF_j$). That is

$$DEFF_j = \frac{SE_j(SCS)}{SE_j(SRS)}.$$

The standard errors estimated by typical statistical software packages such as SPSS, STATA or SAS are generally calculated under the assumption that the data has come from a SRS. The larger the design effect, the more that standard errors erroneously calculated under an SRS assumption will underestimate the standard errors that befit the SCS sample design of NELS. Essentially, the clustering of the NELS sample decreases the effective sample size because students sampled within the same school are not statistically independent. Note that this violates a common assumption of both linear regression and the Heckman Model, namely, that ε_i and δ_i are each independently distributed across students. If this lack of independence is not taken into account, tests of significance using estimated standard errors that are too small may well result in Type I errors.

A school identification code is available for 13,471 students (92%) in the NELS F1-F2 panel. These students were sampled from 974 different high schools. The mean and median size of the student clusters per school is 14. According to the NELS F2 manual this corresponds to a mean and median design effect across all variables of about 3.7 and 3. For subsamples of students in the F1-F2 panel, the mean and median cluster sizes, and presumably the corresponding design effects will be smaller. Finding out just how much smaller is outside the scope of this dissertation. For the analyses in this chapter and those that follow, all standard errors are estimated using proportional population weights that include a design effect correction to reduce the effective sample size. This amounts to a first order approximation of the standard errors that would be estimated under the assumption of a SCS.

More specifically, denote each student in the NELS F1-F2 panel sample with the subscript *i*. For any subset of *S* cases taken from the F1-F2 panel sample, the NELS variables that correspond to student *i* are weighted by the variable $DESWGT_i$, where

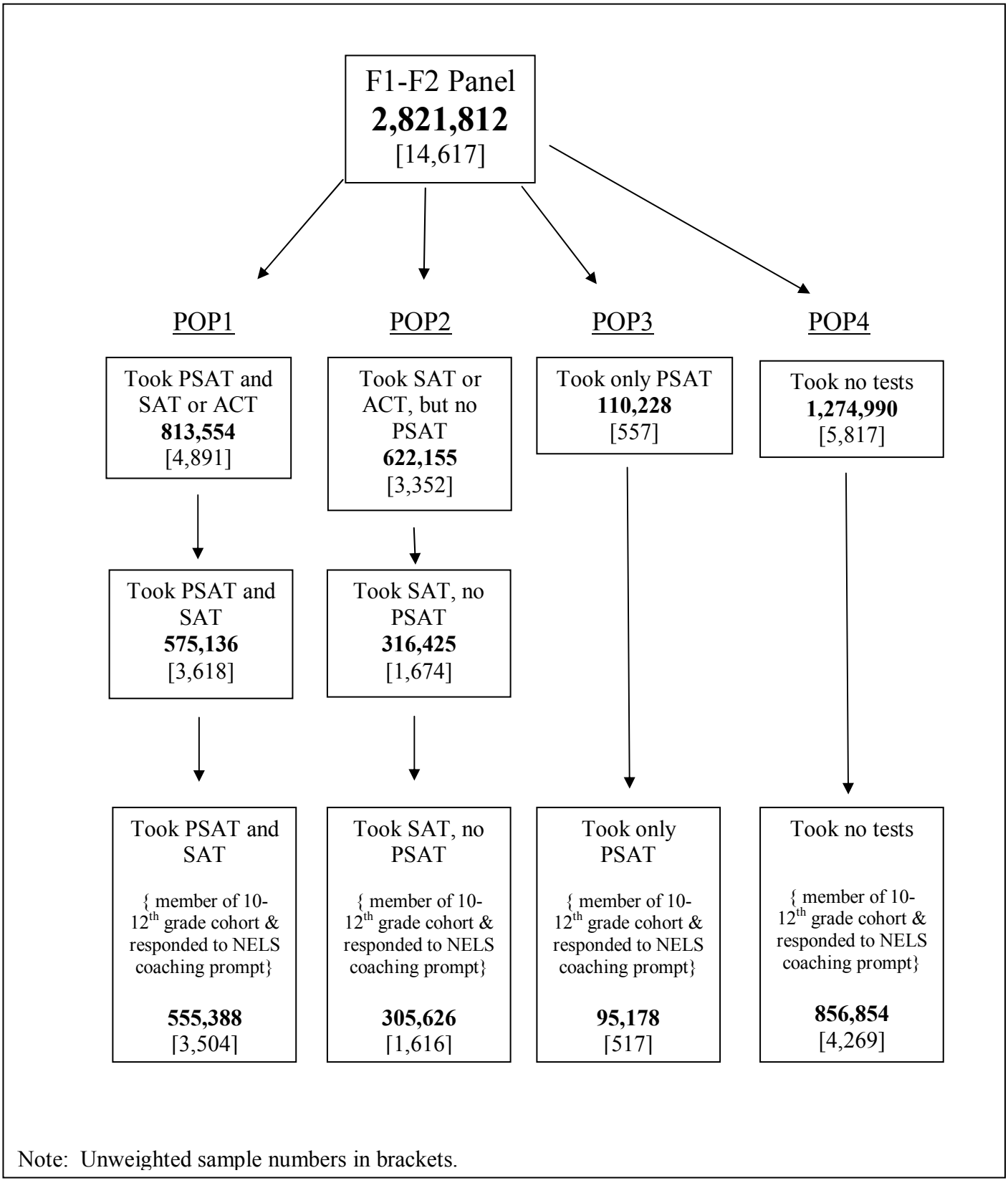
$$DESWGT_i = \frac{1}{DEFF}\left(\frac{F2TRP2WT_i}{\frac{1}{S}\sum_{i=1}^{S} F2TRP2WT_i}\right).$$

$F2TRP2WT_i$ is the population weight of cases in the F1-F2 panel sample for whom transcript data was collected, and *DEFF* is a postulated design effect that applies to all NELS variables. As an approximation of the design effect associated with each variable, it is assumed that $DEFF_j = DEFF$. The appropriate *DEFF* value for the F1-F2 subsamples I analyze below is probably somewhere between 1 (no design effect) and 3 (the median *DEFF* for all variables in the F1-F2 panel sample). I will generally take a conservative approach to standard error estimation, using *DEFF* = 3 for all tests of statistical significance done in SPSS or STATA unless otherwise specified. In all tests of statistical significance, I use a critical value of .05.

Test-Taking Populations in the F1-F2 Panel

Figure 3-2 presents a flow chart that details the different sub-populations of the students in the F1-F2 student panel with respect to their standardized admissions test-taking histories.

Figure 3-2. Different Test-taking Populations in the NELS F1-F2 Panel



Note: Unweighted sample numbers in brackets.

The focus here is on the two most widely taken admissions tests nationally, the SAT and the ACT. In addition we are interested in whether students took the PSAT,[4] because these students may comprise a different underlying population relative to those who only take the SAT and/or ACT. In Figure 3-2 both the weighted and unweighted numbers of students from the NELS F1-F2 panel sample are shown, with the unweighted numbers in brackets. The full population of students represented by the NELS sample can be divided into four groups. The first, POP1, represents 813,554 students who took both the PSAT and the SAT, or the PSAT and ACT. The second, POP2, represents 622,155 students who did not take the PSAT, but subsequently took the SAT or ACT. The third group, POP3, represents 110,228 students that took the PSAT but did not subsequently take either the SAT or ACT. The fourth group, POP4, represents the 1,274,990 students who took no tests at all. All students in the F1-F2 panel fall within one of these four groups.

This dissertation analyzes coaching effect estimates just for the SAT.[5] The emphasis in most SAT coaching studies has been on students like those in POP1, who have taken the SAT and for whom there is a prior SAT or PSAT score available before a test preparation treatment has been introduced. The analysis in this chapter and next is similarly restricted to this group of students. In chapter 5, I compare coaching effects estimated for the POP1 subsample to effects estimated for the POP2 subsample.

[4] The PSAT is essentially a pre-test for the SAT, but is also taken by most students who only take the ACT.
[5] See Briggs, 2001 for an evaluation of NELS coaching effects using ACT scores.

3.3 The NELS Variables

The NELS data is purely observational—no students were randomly assigned to experimental conditions before survey questions and test instruments were administered. To estimate a coaching effect from the NELS data using the kind of behavioral model introduced in chapter 2 requires three types of variables: an outcome variable (Y), a coaching variable ($COACH$) and covariates ($\mathbf{X}$). In what follows, these variables are described with respect to the 3,504 students from the POP1 sample who took both the PSAT and SAT, were members of the 10th grade and 12th grade cohorts as of the NELS F1 and F2 surveys, and indicated whether or not they had been coached as a means of preparing for the SAT. Thirteen students were excluded from the POP1 subsample because they were not members of the 10th to 12th grade cohort. Another 97 students were excluded because they did not answer the NELS prompt about their coaching status.[6] A detailed crosswalk of all variables presented below (including NELS variable mnemonics and survey source) is provided in Appendix Table A-1. Also reported in Table A-1 are the bivariate correlations of all covariates with SAT scores.

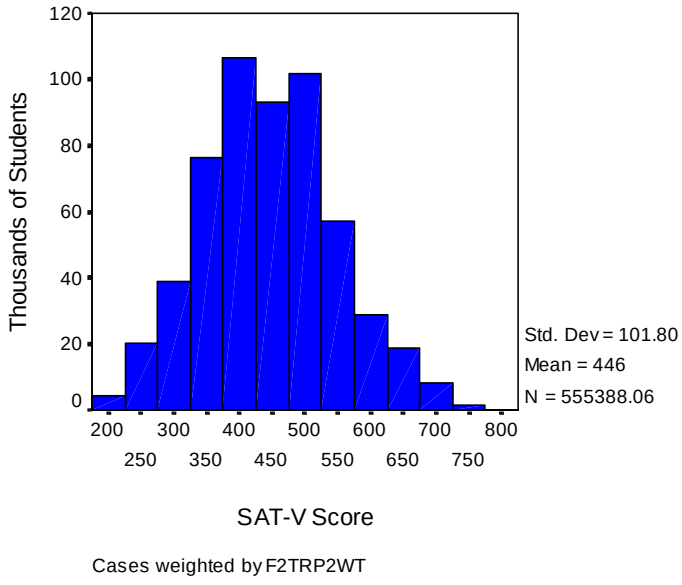
Math and Verbal SAT Scores

The outcome variable of interest is a score on either the math or verbal section of the SAT. As of the early 1990's, the SAT was a timed multiple choice test lasting for a

[6] The mean math and verbal SAT scores of those students who did not respond to the coaching prompt was 370 and 441, significantly different from the respective mean scores 446 and 501 of those who did respond. If the students who did not respond were more likely to be coached or uncoached, it introduces another source of bias into estimated effects. However, because the number of missing cases is small (< 3%), the potential bias is likely to be small as well.

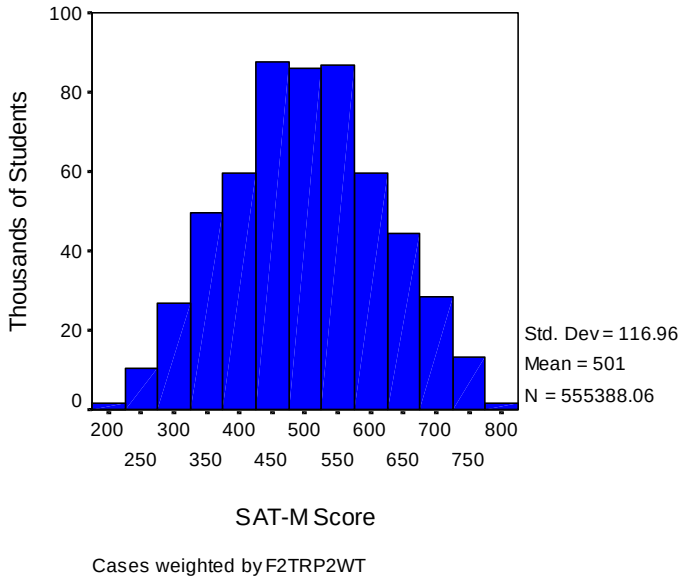
total of two and a half hours. The test was then, and is now, intended to measure the constructs of mathematical and verbal reasoning, and to this end students taking the test are given separate math and verbal scores on scales ranging from 200 to 800 points with a standard deviation that is usually about 110 points. Each score is based on student responses to about 85 verbal items and 60 math items on the SAT. Because student scores are based on a relatively large number of items, and these items are chosen with great care, the SAT has the desirable technical feature of high reliability. The reliability of SAT math and verbal scores using Cronbach's Alpha is about .9, and the standard error of measurement for each test section is usually about 30 points. Figures 3-3 and 3-4 plot the histogram distributions of the SAT math (SAT-M) and SAT verbal (SAT-V) scores for the students in the NELS POP1 subsample who took both the PSAT and the SAT[7]. The cases in these histograms represent over half a million high school students in the 12th grade during the 1991-1992 school year nationally.

Figure 3-3. SAT-V Scores of 10-12th Grade NELS Cohort



[7] The SAT scores of students in the NELS F1-F2 panel sample were gathered from high school transcripts in the summer *after* they had completed the 12th grade.

Figure 3-4. SAT-V Scores of 10-12th Grade NELS Cohort



The two variables appear to be normally distributed. The mean and standard deviation of SAT-V scores (446 and 102) are both slightly lower than the mean and standard deviation of SAT-M scores (501 and 117).[8] The mean scores for all college-bound seniors taking the test in 1991-92 was about 423 on the SAT-V, and 475 on the SAT-M. The mean SAT scores for the NELS POP1 subsample are slightly higher than those of the national population of test-takers because they are restricted to those students who had previously taken the PSAT.

[8] The SAT score scale was recentered as of 1995 (see Dorans, 2002 for details) . Historical tables with mean SAT scores are now expressed in this metric. The mean scores for the NELS POP1 subsample correspond to recentered scores of 543 on the SAT-V and 524 on the SAT-M.

The Coaching Variable

The treatment variable of interest is whether or not students have been coached before taking the SAT. The NELS F2 questionnaire asked students a targeted question about their test preparation activities. This question is replicated verbatim below.

> 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

With the exception of studying with a book, all of the methods listed above to prepare for the SAT have been classified as coaching in previous studies. In this analysis, students are classified as having been coached if they have enrolled in a commercial test preparation course. For a student answering question B above with a "yes", the dummy variable *COACH* is coded with a 1. For students answering with a "no", *COACH* is coded with a 0. 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 are excluded from the coaching definition because while the instruction may be systematic, it has no time constraint. Preparation with a tutor is excluded because while it may have a time constraint, it is difficult to tell if the instruction has been systematic. This definition of the term is consistent with that used by Powers & Rock (1999), and this makes the coaching effect estimates generated from the NELS data somewhat more comparable those generated from the nationally representative data in the Powers & Rock study. Also, commercial coaching is the most controversial means of test preparation, because it

is costly, widely available, and comes with published claims as to its efficacy. Table 3-1 presents the numbers of coached and uncoached students among those from the POP1 subsample who took both the PSAT and SAT. About 15% of the students indicated that they had taken a commercial course to prepare for the SAT.

Table 3-1. Proportion of Coached Students in POP1 Subsample

COACH	NELS POP1 subsample	National Population	% of Total Nat Pop
= 1	587	83,924	15
= 0	2,917	471,464	85
Total	3,504	555,388	100

It is of some interest whether coached students are more likely to prepare for the SAT in more ways than uncoached students. In fact, coached students are significantly more likely to prepare for the SAT by taking a course offered at their high school, employing a private tutor, and using books, videos and computers. On average, coached students report that they had prepared for the SAT with two different activities beyond a commercial course. Students not taking a commercial course report having prepared with just one activity. Unfortunately, these other test preparation variables cannot be considered covariates in a linear regression or Heckman Model because they are not temporally or logically antecedent to the variable *COACH*. For example, being coached might make it more likely that a student also prepares for the SAT with a tutor.[9] This raises the point that covariates in this analysis are restricted to characteristics of students taking the SAT that were either a) measured before the student was coached or b) not conceivably influenced by the coaching experience. This excludes certain variables

[9] In chapter 5, I take up the issue of whether alternate definitions of coaching in terms of combinations of the different test preparation indicators will change the magnitude and/or significance of the estimated coaching effect under the linear regression model.

available in the NELS database such as career and educational aspirations, because these variables may be influenced by whether or not a student has been coached.

Covariates

I group covariates that may confound the relationship between coaching and SAT performance into four categories: PSAT scores, demographic characteristics, academic background and intrinsic motivation. Variables from each of these categories, and their relationship to coaching status, are described and analyzed below. I also consider a small set of variables that may predict whether students are likely to be coached, but are unlikely to predict how well they will perform on the SAT. These variables, which seem to measure extrinsic motivation, should be particularly attractive candidates for inclusion in a selection function for coaching as part of the Heckman Model.

PSAT Scores

There is no information available in NELS on students who may have taken the SAT twice. However, for students in the POP1 sample, there are test scores available indicating prior performance on the PSAT. The PSAT, taken by most students in 10^{th} grade, 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 with scores on the corresponding sections of the SAT. The correlation of PSAT-M with SAT-M is .87; the correlation of PSAT-V with SAT-V is .88. Other

researchers have previously compared raw scores from the PSAT to SAT by multiplying PSAT scores by 10. The same tactic is taken here. Figures 5 and 6 plot the SAT and PSAT scores for the math and verbal sections of the test. The PSAT, SAT score pairs of coached students are denoted with solid circles, while those of uncoached students are denoted with empty circles.

Figure 3-5. Plot of SAT-M and PSAT-M Scores by Coaching Status

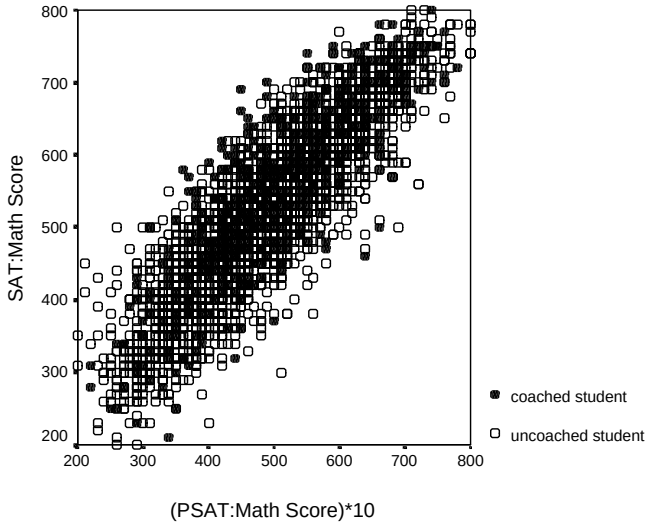


Figure 3-6. Plot of SAT-V and PSAT-V Scores by Coaching Status

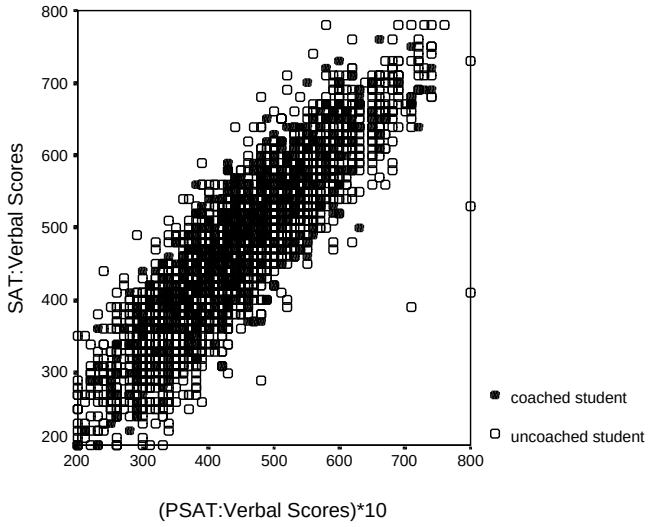


Table 3-2. PSAT Scores by Coaching Status

Covariate	Uncoached	Coached	Stat Sig (p-value)
PSAT-Verbal	422 (3.1)	427 (7.5)	.312
PSAT-Math	465 (3.6)	475 (8.2)	.490

standard errors in parentheses

Table 3-2 presents a cross-tabulation of PSAT scores by coaching status. The last column of this table, and the tables that follow, indicate the probability (p-value) that the observed differences in characteristics between coached and uncoached students could be explained by chance variation, given the null hypothesis that they are 0. When the covariate is categorical, the p-value is calculated using the Pearson Chi-Square test of independence. When the covariate is continuous, as in this case, the p-value is calculated from an F-test or two-tailed t-test. On average, coached students tend to score slightly higher on the PSAT-V and PSAT-M than their uncoached counterparts, but neither difference is statistically significant.

Demographic Characteristics

Demographic covariates include student characteristics such as age, gender, race/ethnicity, type of school attended and whether the school is in an urban, suburban or rural setting. The most important of the demographic covariates is an index of socioeconomic status (SES). The SES index was developed as part of the NELS database, and combines information about parental education, income and occupation into a single variable. Generally, students with higher SES values come from families

with parents that are better educated, wealthier and have jobs in more prestigious occupations. For students represented by the POP1 subsample, the SES variable has a mean of .44, a standard deviation of .73, and a range from –2.4 to 2.5. For the full F1-F2 panel, the variable has a wider range at the bottom end, because students who do not take admissions tests tend to have lower SES values. The SES index has also been formulated as an ordinal variable that categorizes students from the full NELS sample into four SES quartiles. The grouping of students into SES quartiles by coaching status provides a more absolute comparison of socioeconomic status with respect to the national population.

Table 3-3. Demographic Characteristics by Coaching Status

Covariate	Uncoached	Coached	Stat Sig (p value)
Age at time of NELS F2 survey	17.8 (.54)	17.9 (.43)	.005
Female %	56.4	57.5	.919
Race/Ethnicity %			.511
American Indian	<1	<1	
Asian	6.2	9.2	
Black	9.4	8.1	
Hispanic	7.6	8.7	
White	76.5	74	
Type of School			.237
Public	79.9	74.7	
Catholic	13.2	16.7	
Other Private	6.9	8.6	
Location of School			.001
Urban	36.1	46.6	
Suburban	43	44.3	
Rural	20.9	9.2	
SES Index	.39 (.71)	.73 (.78)	<.001
SES Quartile %			<.001
Top Quartile	45.7	71.8	
Second Quartile	29.2	16.7	
Third Quartile	17.5	6.3	
Bottom Quartile	7.6	5.2	

standard errors in parentheses

Table 3-3 presents the cross-tabulation of demographic covariates by coaching status. Coached and uncoached students are fairly similar with respect to their gender, race/ethnicity, and type of school. There are significant differences between the two groups in terms of their age, school setting and socioeconomic status. Coached students are on average about one month older and come from high schools in non-rural settings. Most strikingly, it is clear that coached students are far more socioeconomically advantaged than uncoached students. The mean SES index value for coached students is about half a standard deviation higher than the mean for uncoached students, and 72% of coached students are in the top socioeconomic quartile, compared to 46% of uncoached students.

Academic Background

Academic background covariates include the number of college-preparatory math courses taken in high school, the weighted grade point average obtained in those math courses, and scores on NELS standardized tests in math and reading administered in the 10^{th} grade. The standardized test scores have means of 58 and 57 for the math and reading subjects respectively, with standard deviations of about 8. In addition, there are four covariates that provide information about each student's high school curriculum. Academic background covariates are generally assumed to relate to SAT scores such that students who did better in high school in terms of grades, test scores and course-taking patterns will also do better on the SAT.

Table 3-4. Academic Achievement by Coaching Status

Covariate	Uncoached	Coached	Stat Sig (p value)
F1 Math Test Std Score (10th Grade)	57.4 (8.2)	57.9 (7.8)	.539
F1 Reading Test Std Score (10th Grade)	56.6 (8.6)	56.4 (8.2)	.783
Units of Math taken in high school	3.71 (.78)	3.75 (.64)	.529
Weighted GPA in Math Courses	2.6 (.76)	2.8 (.83)	.007
High School Program			.18
Rigorous Academic	39.2	41	
General	53.5	55.5	
Other	7.3	3.5	
Taken a remedial English course	6.1	9.4	.103
Taken a remedial Math course	8.8	10.1	.575
Taken an AP class	57.6	61.6	.270

standard errors in parentheses

Table 3-4 presents the cross-tabulation of academic achievement covariates by coaching status. For the most part, coached students do not appear to be academically "smarter" than their uncoached counterparts. Both groups have roughly the same mean score on the standardized NELS F1 tests in math and reading. Both coached and uncoached students were equally likely to have taken a rigorous academic course load, including four units of college-preparatory math and at least one AP class. The average weighted GPA of 2.8 that coached students attained in their math courses was slightly, but significantly higher than the GPA of 2.6 attained by uncoached students.

Student Intrinsic Motivation

The NELS survey asks students a wide variety of questions that might help gauge their intrinsic levels of motivation. Two of the more interesting ones are composite variables that measure, respectively, student self-esteem and locus of control (e.g.

whether students feel they have control of their lives). Like the SES index variable, these variables are constructed as part of the NELS survey from a number of related survey items. The self-esteem and locus of control variables are based on the responses to 7 and 6 Likert items from the F1 survey, before students were likely to have been exposed to the coaching treatment. For students represented by the POP1 subsample, the self-esteem and locus of control variables have means of .13 and .20, with standard deviations of .67 and .58, and values ranging from –3 to 1.4 and –2.7 to 1.5. Generally, students with higher values on these variables have responded to a small subset of NELS survey items in ways that suggest they have more positive feelings of self-esteem and locus of control in their lives. Each index variable was measured at two points in time using the same set of items: once from the F1 student survey, and then again from the F2 student survey. The correlation of the two variables between the 10^{th} and 12^{th} grade was .56 for the esteem construct, and .49 for the locus of control construct. Another covariate that might proxy for intrinsic student motivation derives from a retrospective NELS survey item that asked students to report the average number of hours they had spent each week on homework during high school. The operating theory behind the use of these variables as covariates is that "more" of each variable translates into students who are more intrinsically motivated to perform well on the SAT.

Table 3-5 presents the cross-tabulation of intrinsic motivation covariates by coaching status. In terms of their reported self-esteem and locus of control index scores as of the 10^{th} grade, there is no significant differences between coached and uncoached

students. Coached students appear to spend more of their spare time doing homework, than uncoached students, but again this difference is not statistically significant.

Table 3-5. Intrinsic Motivation by Coaching Status

Covariate	Uncoached	Coached	Stat Sig (p value)
Self-Esteem Index	.12 (.67)	.18 (.64)	.249
Locus of Control Index	.20 (.59)	.20 (.56)	.992
Homework done outside of school (hours per week) %			.200
16 or more hours	11.4	15.9	
10-15 hours	23.6	26.5	
1-9 hours	58.2	52.9	
<1 hour	6.8	4.7	
standard errors in parentheses			

Student Extrinsic Motivation

Students who are extrinsically motivated, for example, those that are pressured by their parents, may be especially likely to get coached, though this would seemingly have no direct bearing on SAT performance. Three NELS covariates that may proxy for extrinsic motivation are whether students had discussed plans and preparation for the SAT with their parents, whether their parents had encouraged them to take and prepare for the SAT, and whether students had at any time worked with a private tutor during high school. A fourth dummy variable was created from the PSAT and math GPA covariates. Any student scoring below the top quartile in their combined PSAT-V and PSAT-M scores (< 1010 points) among POP1 subsample students, but with a GPA in math courses in the top quartile of POP1 subsample students (> 3.25) gets a value of 1. The notion behind this variable is that it represents students who underachieved on the

PSAT relative to their high school standing. Such students are probably more extrinsically motivated to score higher on the SAT.

Table 3-6 presents the cross-tabulation of extrinsic motivation covariates by coaching status. Coached students were much more likely to have been encouraged by their parents to take the SAT, and to have discussed with them their test preparation plans. Coached students are more likely than uncoached students to have had a paid tutor that helped them with their homework during high school, and to have underachieved on the PSAT.

Table 3-6. Extrinsic Motivation by Coaching Status

Covariate	Uncoached	Coached	Stat Sig (p value)
Private tutor helped w/ homework in high school %	10.7	17.1	.017
Student discussed plan to prepare for SAT w parents%			<.001
Often	20.1	44.8	
Sometimes	54.5	36.2	
Never	18.4	8.6	
....Missing Response	7.1	10.3	
Parents encouraged student to prepare for SAT %	87.3	98.0	<.001
Student scored below 1010 on PSAT, but has GPA > 3.25	10.9	22	<.001

The picture that emerges from Tables 3-3 through 3-6 is that of a coached group of students who are socioeconomically advantaged and more extrinsically motivated to take the SAT then uncoached students. It is not clear that the coached group is necessarily comprised of academically "smarter" or more intrinsically motivated students—both groups are enrolled in college-preparatory classes, both performed about the same on NELS standardized tests in reading and math, both report having comparable

levels of self-esteem and locus of control, and both report that they do about the same amount of homework per week.

3.4 Data Limitations

On the one hand, NELS is an ideal source of observational data for evaluating the effectiveness of coaching on standardized admissions tests. First, the data in NELS are nationally representative, and the target population is well-defined. It is not simply students taking standardized admissions tests in American high schools at the national level but rather all American high school students who *could* have taken these tests. Second, unlike data used in previous coaching studies, which is often exclusively student-reported, most of the NELS data on student academic performance is collected from official transcripts. There is evidence that self-reported data on student grades and test scores tend to overstate student performance (Morgan, 1990; Frucot & Cook, 1994). This criticism has been levied at the Powers & Rock study, which relied on student responses to a mailed survey for information about high school grade and course-taking patterns. Such criticism is not applicable to the NELS academic achievement and test score variables.

On the other hand, there is no getting around the fact that NELS was not designed specifically to address the issue of coaching effectiveness. While students were asked to indicate how they prepared for admissions tests, subsequent questions were not asked about the quality, intensity and duration of their test preparation. As a result, any student

claiming to have received coaching to prepare for a test, is assumed to have received the same quality, intensity and duration of coaching as any other student making the same claim. This is simply the constancy assumption, discussed briefly in chapter 2, revisited. As I suggested then, this is often a questionable assumption, and I consider (to the extent that this is possible given the data) the sensitivity of it later in chapter 5. In addition, while data on student admissions test performance is gathered from transcripts, only one set of scores exists for students who have taken a standardized admissions test. There is no indication whether students have taken these tests multiple times, even though we can be certain that many students have done so. Finally, no information is available as to precisely when students were coached. In a re-analysis of the Powers & Rock data, Hansen (2002) finds evidence that as much as one quarter of the student sample had participated in a commercial coaching program at a reported date *before* taking the PSAT (or the SAT for the first time). If this is true in the NELS data as well, it throws into question the use of PSAT scores as covariates. The assumption made here is that the SAT scores available from student transcripts reflect the highest score a student attained on the test, and that the coaching treatment occurred before taking the SAT, but after taking the PSAT.

CHAPTER 4: ANALYSIS

Our interest is in estimating the effect of coaching on SAT scores for students in the NELS POP1 subsample. In chapter 2 I described a behavioral model for SAT performance under which the coaching parameter *b* has a causal interpretation. This model is revisited below.

$$f_i(COACH) = a + bCOACH + \mathbf{X}_i\mathbf{c} + \sigma\varepsilon_i \tag{1}$$

$$COACH_i = 1 \Leftrightarrow \alpha + \mathbf{Z}_i\boldsymbol{\gamma} + \delta_i > 0\,. \tag{2}$$

$$Y_i = f_i(COACH_i) = a + bCOACH_i + \mathbf{X}_i\mathbf{c} + \sigma\varepsilon_i\,. \tag{3}$$

The selection function (2) has now been modified so that the covariates in the selection function ($\mathbf{Z}_i$) are allowed to be different from those in the response schedule ($\mathbf{X}_i$). If ε_i and δ_i are iid, and the $\mathbf{X}_i$'s and $\mathbf{Z}_i$'s are independent of the ε_i's and δ_i's respectively, and if ε_i and δ_i are independent within student *i*, the only bias in *b* will be due to confounding. So $Y_i = f_i(COACH_i)$ and the coaching effect can be estimated by regressing Y_i on a constant, $COACH_i$ and $\mathbf{X}_i$. If $\mathbf{X}_i$ includes all measured covariates that confound the relationship between coaching and SAT performance, then the OLS estimate $\hat{b}$ will be unbiased.

If the assumption that ε_i and δ_i are independent is dropped, such that the error terms are allowed to be correlated by the parameter ρ, then the OLS estimate $\hat{b}$ may

contain bias due to self-selection, even if all the available confounding variables have been included in the model. Assuming that ε_i and δ_i are both normally distributed, the Heckman Model is a possible correction for this problem. The parameters $\hat{\alpha}$ and $\hat{\boldsymbol{\gamma}}$ in the selection function are estimated using maximum likelihood. These parameters in turn are used to estimate $\lambda_i(COACH_i, \hat{\alpha}+\mathbf{Z}_i\hat{\boldsymbol{\gamma}})$, an interaction term between $COACH_i$ and the Inverse Mills Ratio for each student. Now either OLS or GLS can be used to get a theoretically unbiased $\hat{b}$ by regressing Y_i on a constant, $COACH_i$, $\mathbf{X}_i$ and $\lambda_i(COACH_i, \hat{\alpha}+\mathbf{Z}_i\hat{\boldsymbol{\gamma}})$.

Coaching effects can be estimated from the NELS data using both the linear regression model and Heckman Model. Estimates from each model can be compared to the simplest alternative: the average SAT section score for coached students minus the average SAT score for uncoached students. For the SAT-V, this difference is 20 points (463 – 443); for the SAT-M, the difference is 30 points (526 – 496). If coached and uncoached students had been assigned randomly, these would be unbiased estimates of the coaching effects, and the usual method of determining the statistical significance of these differences could be used. Of course, we know that students in NELS were not randomly assigned, so these estimates are almost surely biased to some degree. What do linear regression and the Heckman Model suggest about the magnitude of this bias?

4.1 Coaching Effects and the Linear Regression Model

To control for confounding in the estimation of coaching effects, an appropriate set of covariates must be chosen for $\mathbf{X}_i$. The choice of covariates can be guided to a great extent by previous investigations of coaching effectiveness. My review of the literature indicated that previous SAT or PSAT scores, demographic characteristics, academic background and student motivation may serve to confound coaching effect estimates. These potential confounders were described and analyzed sequentially relative to coaching status in chapter 3. Now I establish a linear regression model with all covariates that theoretically confound the coaching estimate. If this model is to have a causal interpretation, it should hold not only for the NELS sample, but for other samples as well. Hence all covariates with a theoretical relationship with both coaching status and SAT performance must be included in the model, even if the empirical relationship with respect to the NELS sample suggests otherwise.

I start by including each student's PSAT section score (*PSAT-V* and *PSAT-M*) as a covariate in the regression model. Demographic characteristics included are a student's age in years (*AGE*), socioeconomic status (*SES*), dummy variables for gender (*FEMALE*), race/ethnicity (*ASIAN*, *BLACK*, *HISPANIC*, *NATIVE*, *WHITE*), and whether the student's high school was public or private (*PRIVATE*), or located in a suburban, rural or urban locations (*SCH_URB*, *SCH_RUR*, *SCH_SUB*). Academic background covariates are dummy variables for whether or not a student reports having taken an Advanced Placement class (*AP*) or remedial classes in math (*REMATH*) or English (*RE_ENG*). A

strongpoint of the NELS data is the availability of transcript-based academic background variables. Such variables include a dummy variable indicating whether or not the student has been enrolled in a rigorous academic program while in high school (*RIGHSP*), scores on standardized achievement tests in math (*F1MATH*) and reading (*F1READ*), the number of units a student has taken in college preparatory math courses[1] (*MTHCRD*), and his or her weighted grade point average in those courses (*MTHGRD*). Finally, three covariates were added to proxy for intrinsic student motivation: the NELS self-esteem (*F1ESTEEM*) and locus of control (*F1LOCUS*) indices, and a dummy variable indicating whether the student reported averaging more than 10 hours per week on homework during high school (*HOMEWORK*). There are a total of 21 covariates in the linear regression model. The reference categories are *WHITE* and *SCH_SUB* for the racial/ethnic and school location dummy variables respectively.

Table 4-1 reports the results of separate linear regressions of student SAT-V and SAT-M scores on a constant, *COACH*, and the full set of 21 covariates in $\mathbf{X}_i$ listed above. Each regression was weighted by the variable *DESWGT* (described in chapter 3) to account for the stratification and clustering of students in the NELS POP1 sample. Regressions were run with two different versions of *DESWGT*; one where the design effect (*DEFF*) is set equal to 1, the other with a *DEFF* set equal to 3. A *DEFF* of 1 assumes no design effect, and as such probably underestimates standard errors. A *DEFF* of 3 reflects the median design effect reported in the NELS F2 Student Survey User's Manual for the full NELS F1-F2 panel. The clustering of students in the POP1 subsample, amounts to a mean of 4 and median of 6 students per school—relative to a

[1] College preparatory math courses consist of algebra, geometry, trigonometry, pre-calculus and calculus.

mean and median of 14 for the full F1-F2 panel sample. In the POP1 subsample there is on average just one coached student per sampled school. Given this, a *DEFF* of 3 probably tends to overestimate standard errors. Hence, all else being equal, the standard errors of parameter estimates associated with each version of the *DESWGT* variable should reflect lower and upper bounds in tests of statistical significance, and to give a sense for this range, both are reported for the regression coefficient estimates in Table 4-1.

Table 4-1. Coaching Effects using the Linear Regression Model

	SAT-V (mean = 447, sd = 101)			SAT-M (mean = 504, sd = 116)		
R^2	.788			.822		
adj R^2	.787			.818		
Coached/Total	503/3144			503/3144		
Variables in Regression Eqn	$\hat{a},\hat{b},\hat{\mathbf{c}}$	Std Error Range *DEFF* = 1	Std Error Range *DEFF* = 3	$\hat{a},\hat{b},\hat{\mathbf{c}}$	Std Error Range *DEFF* = 1	Std Error Range *DEFF* = 3
Constant	144.1	36.1	63.6	-7.6	37.5	66.1
COACH	11.1*	2.4	4.3	19.2*	2.5	4.5
PSAT-M	.05*	.02	.03	.41*	.02	.03
PSAT-V	.61*	.01	.02	.09*	.01	.02
AGE	-8.7*	1.9	3.4	-2.7	2.0	3.5
SES	3.8	1.4	2.4	10.2*	1.4	2.5
FEMALE	-5.0	1.9	3.3	-16.1*	1.9	3.4
ASIAN	7.9	3.5	6.2	4.8	3.6	6.4
BLACK	-3.5	3.2	5.6	-14.3*	3.3	5.8
HISPANIC	-3.1	3.4	6.1	-4.6	3.6	6.3
NATIVE	-6.2	14.4	25.4	-26.2	15.0	26.4
PRIVATE	8.9*	2.4	4.2	-0.9	2.5	4.4
SCH_RUR	-6.6	2.3	4.0	-3.5	2.4	4.1
SCH_URB	1.1	2.0	3.6	1.3	2.1	3.7
AP	12.4*	1.9	3.3	8.8*	2.0	3.5
RE_ENG	-11.4	4.2	7.4	8.2	4.4	7.7
REMATH	1.7	4.0	7.1	-19.1*	4.2	7.3
RIG_HSP	-1.2	1.7	3.1	2.8	1.8	3.2
F1READ	2.5*	0.2	0.3	-0.5	0.2	0.3
F1MATH	0.4	0.2	0.4	4.9*	0.2	0.4
MTHCRD	-1.3	1.3	2.3	8.8*	1.3	2.4
MTHGRD	3.6	1.4	2.4	14.8*	1.4	2.5
F1ESTEEM	5.2	1.6	2.8	-1.9	1.6	2.9
F1LOCUS	-6.2	1.8	3.2	-2.1	1.9	3.4
HOMEWORK	3.5	1.8	3.1	1.4	1.9	3.3

* p-value for two-sided t-test < .05 across SE range

Under the linear regression model, the estimated effect for *COACH* is 11 and 19 points respectively on the SAT-V and SAT-M. Expressed as a proportion of a standard deviation in SAT scores, this amounts to effect sizes of .11 and .16 for each estimate. Both effects are statistically significant whether tested using the standard errors based on the lower or upper *DEFF* bounds. Using the more conservative standard error estimate, the 95% confidence intervals for the estimated SAT-V and SAT-M coaching effects are [3, 20] and [10, 28]. These estimated effects suggest that the linear regression model does reduce bias due to confounding by inclusion of the covariates in $\mathbf{X}_i$. The estimated SAT-V coaching effect decreased by 9 points from 20 to 11, and the estimated SAT-M coaching effect decreased by 11 points from 30 to 19.

No covariates included in the full regression model were missing for more than 5% of the sample. Listwise deletion of missing data in the linear regression model reduced the POP1 subsample size from 3,504 to 3,144. This is a concern if data with missing values is not missing at random among coached and uncoached students (Little & Rubin, 1987). If the missing at random assumption is wrong, the exclusion of cases with missing data becomes another potential source of bias in the model. The cross-tabulation of coaching status by cases included and excluded from the linear regression model revealed no evidence to reject the missing at random assumption: the proportions of coached and uncoached students were the same for included and excluded cases, and a Chi-Square test for independence was not statistically significant (p-value = .69).

Under the behavioral model for linear regression considered here, only the estimated coefficient for the variable *COACH* has a causal interpretation. Nonetheless, the estimated coefficients for the covariates in $\mathbf{X}_i$ merit some scrutiny. If the estimates suggest associations with SAT performance that contradict the theory under which they were chosen for inclusion (e.g. students with better grades in math perform worse on the SAT-M), or simply defy common sense, we have a clear reason to worry about whether the model has been properly specified.

When SAT-V scores are the outcome variable, five covariates have positive, statistically significant associations: *PSAT-M*, *PSAT-V*, *AP*, *F1READ* and *MTHGRD*. Every 100 points a student scores on the PSAT-M is associated with an SAT-V score increase of 5 points; every 100 points scored on the PSAT-V is associated with an SAT-V score increase of 61 points. Taking at least one Advanced Placement course during high school is associated with a 12 point score increase. Scoring one standard deviation (8 points) higher on the F1 NELS standardized reading test is associated with a 20 point score increase. One covariate, *AGE*, has a negative, statistically significant association with SAT-V scores. The age of students in the POP1 subsample as of the spring of their senior years ranged from about 16 to 19, with a mean and median of about 18 years. According to the linear regression model, an additional year in age is associated with a SAT-V score decrease of about 9 points. A possible explanation for this is that younger students who are taking the SAT as high school seniors are better test-takers than older students taking the SAT, particularly if they have skipped a grade on the basis of test performance at an early age. In any case, the association is not very strong—one

standard deviation on the age variable is equivalent to six months, and relatively few students in the sample are separated in age by more than a year.

When SAT-M scores are the outcome variable, seven covariates have positive, statistically significant associations: *PSAT-M*, *PSAT-V*, *SES*, *AP*, *F1MATH*, *MTHCRD*, and *MTHGRD*. Every 100 points a student scores on the PSAT-M and PSAT-V is associated with an SAT-M score increase of 41 and 9 points respectively. Being one standard deviation higher on the SES index and taking at least one Advanced Placement course is associated with 10 and 9 point SAT-M score increases respectively. Scoring one standard deviation higher on the F1 NELS standardized math test, taking one more unit of college-preparatory math and the difference between an "A" or a "B" grade point average in such math courses is associated with 40, 9 and 15 point score increases. Three covariates have negative, statistically significant associations with SAT-M scores: *FEMALE*, *BLACK*, and *REMATH*. Being a female student is associated with SAT-M scores that are 16 points lower than being a male student. Being a black student is associated with SAT-M scores that are 14 points lower than those of white students. Having taken a remedial math course is associated with scoring 19 points worse on the SAT-M.

On the whole, the associations of covariates with SAT-V and SAT-M scores in the linear regression model seem reasonable. No estimated coefficients are wildly implausible, though we cannot rule out the possibility that one or more is biased. One possible source of bias may be additional covariates that have been mistakenly omitted

from the regression model. For example, perhaps the correct model would include a series of interaction terms with the coaching variable. I take up this issue in the next chapter. Another possibility is that bias exists of a very specific nature due to the endogeneity of the variable *COACH*. This latter problem is one that the Heckman Model has been designed to solve.

4.2 Coaching Effects and the Heckman Model

Specifying a Selection Function

In order to estimate an effect for *COACH* using the Heckman Model, I start by specifying a selection function that, given a set of covariates $\mathbf{Z}_i$, predicts whether student *i* will be coached or not. The specification decision hinges upon what covariates are included in $\mathbf{Z}_i$. Optimally, students in the NELS survey would have been asked questions about why they did or did not enroll in coaching programs, but as NELS was not designed with the Heckman Model in mind, such data is not available. This is a fairly typical situation in an observational study. A consequence of this is that the specification of a selection function is seldom guided by theory. In many empirical applications of the Heckman Model, the decision of what covariates to include in $\mathbf{Z}_i$ is largely a matter of ensuring that the model is well identified.

Figure 4-1. Five Selection Function Specification

SF1	$\mathbf{Z}_i = \{\mathbf{X}_i\}$
SF2	$\mathbf{Z}_i = \{\mathbf{X}_i, PARENT_i\}$
SF3	$\mathbf{Z}_i = \{PARENT_i, PPRESS_i, HWTUTOR_i, HI\_MOT_i\}$
SF4	$\mathbf{Z}_i = \{SES_i, SCH\_RUR_i, REMATH_i, MTHCRD_i, PPRESS_i, HWTUTOR_i, HI\_MOT_i\}$
SF5	$\mathbf{Z}_i = \{AGE_i, SES_i, SCH\_RUR_i, MTHGRD_i, PARENT_i, PPRESS_i, HWTUTOR_i, HI\_MOT_i\}$

I consider five plausible specifications of a selection function for coaching: SF1, SF2, SF3, SF4 and SF5. The predictors in each specification are listed in Figure 4-1. Which of these is the "right" specification of the selection function? A reasonable case could be made for each of the five. In SF1, all the covariates specified as possible confounders in the regression equation are included as predictors in the selection function, and this represents the kind of mechanical use of the Heckman Model we might expect to see when the data analyst has no operating theory for how students select themselves into coaching. Note that the Heckman Model in this case is identified only by the nonlinearity of the selection function. Some have referred to this as "weak" identification (Breen, 1996; Vella, 1998). In SF2, one additional predictor, the dummy variable *PARENT*—which takes a value of 1 if a student was strongly encouraged by his or her parents to prepare for the SAT—has been added to the selection function. Now the model is overidentified, since *PARENT* is not a covariate in the response schedule. Here we imagine the data analyst has access to at least one variable thought to predict coaching status, but not SAT performance. This is known as a single exclusion restriction. SF2 doesn't constitute a theory per se, but it is the simplest possible improvement over SF1. For SF3, only covariates excluded from $\mathbf{X}_i$ in the linear regression equation are included as predictors in the selection function[2], where *PPRESS*, *HWTUTOR* and *HI_MOT* are dummy variables that take values of 1 if the student's test preparation plans were "often" discussed with his or her parents, if the student had a private tutor that helped with

[2] Values for the predictors *PARENT*, *PPRESS* and *HWTUTOR* were missing for anywhere from 2 to 10% of the POP1 subsample of 3,144 students. To ensure that subsequent Heckman Model parameter estimates will be based on the same sample of students as those produced by linear regression, missing values for these predictors were coded as three unique dummy variables which took the value of 1 if a student's response was missing, and 0 otherwise. For any selection function specification including one or more of these three variables, the associated missing value dummy variable *MPARENT*, *MPPRESS* or *MHWTUTOR* was also included.

homework during high school, and if the student did poorly on the PSAT relative to his high school GPA in math courses. Under SF3, SF2 is augmented such that there are now four variables thought to predict coaching status, but not SAT performance. In addition, the strong and questionable assumption is made that no covariates in $\mathbf{X}_i$ should be used to predict coaching status. The specification SF3 is meant as an extreme contrast with SF1. In SF1, all covariates in $\mathbf{X}_i$ are also in $\mathbf{Z}_i$; in SF3, no covariates[3] in $\mathbf{X}_i$ are also in $\mathbf{Z}_i$. In SF4, all predictors included in the selection function are chosen by a stepwise selection algorithm. SF4 is another example of a mechanical approach a data analyst might take in specifying the selection function: all possible covariates are thrown into an algorithm, and an optimal subset emerges. Finally, for SF5, predictors are chosen for two reasons: because they have some theoretical relationship to coaching status (*SES*, *PARENT*, *PPRESS*, *HWTUTOR*, *HI_MOT*) or because they have an empirical relationship to coaching status (*AGE*, *SCH_RUR*, *MTHGRD*). SF5 is a fairly crude approximation of a theory-based specification approach. Here the data analyst has taken some care in choosing predictors with a hypothesized relationship to coaching status (i.e. it is well-established that coaching programs can be expensive, and hence high-SES students are more likely to enroll in them). In addition, the data analyst has analyzed the pairwise cross-tabulations of all covariates with coaching status, and included three for which there was evidence of a statistically significant relationship. SF5 has four exclusion restrictions as in SF3, but includes in $\mathbf{Z}_i$ a subset of covariates from $\mathbf{X}_i$, as in SF4.

Table 4-2 presents the parameter estimates generated from a weighted probit model (weighted by the variable $DESWGT_i$ with $DEFF = 3$) for each of the five SF

[3] Strictly speaking this is not true since *HI_MOT* is itself a function of *PSAT-V*, *PSAT-M* and *MTHGRD*.

specifications. It is not at all obvious on statistical grounds that any one of the five specifications is the best choice for use in the Heckman Model. Unlike linear regression, where model fit is often assessed on the basis of R^2, there is no such measure of absolute fit for the probit model.

Table 4-2. Selection Function Parameters Estimated using Weighted Probit Model

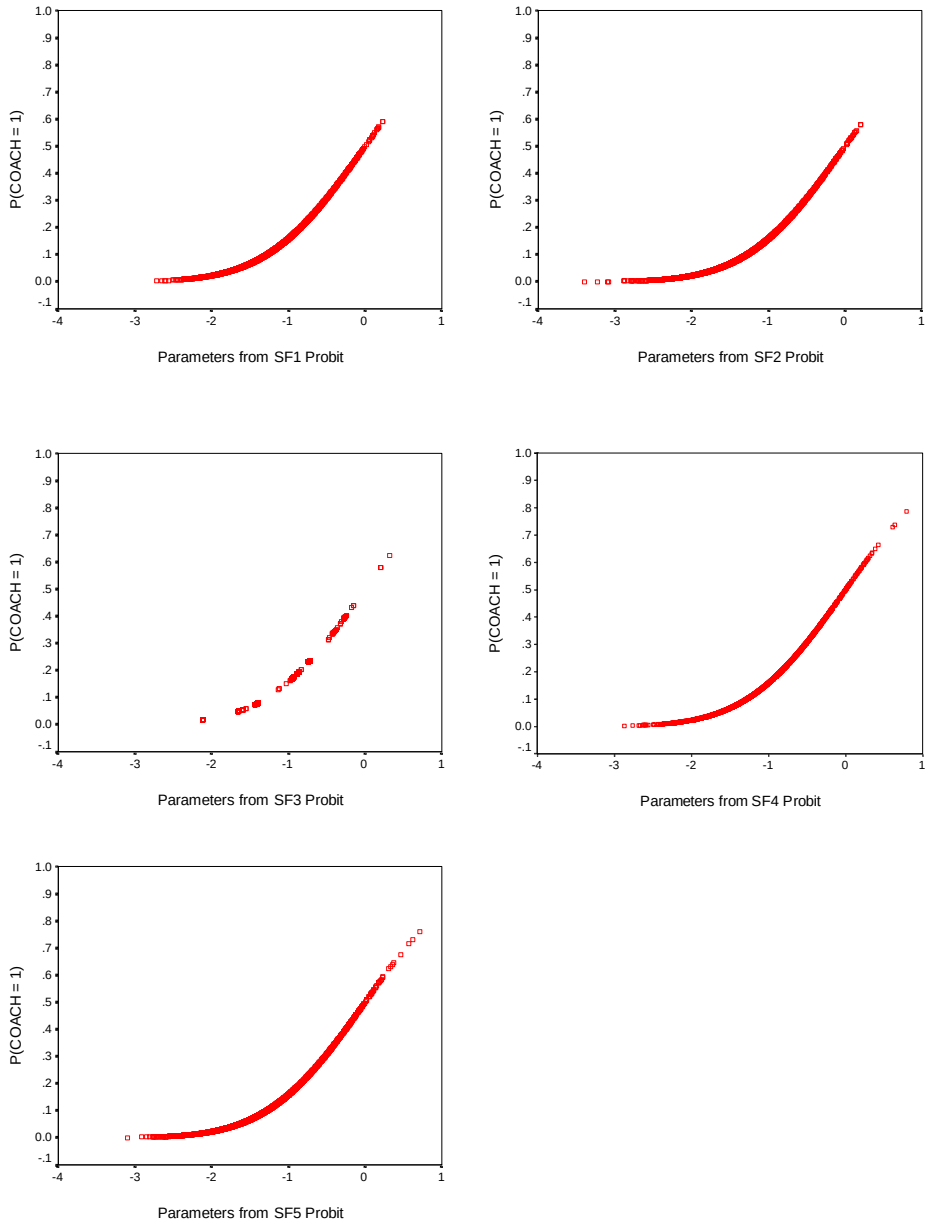
	SF1		**SF2**		**SF3**		**SF4**		**SF5**	
Log Likelihood	-1175.3		-1163.6		-1187.3		-1119.2		-1119.2	
dof	23		25		7		8		11	
Pseudo R^2	.0994		.1084		.0902		.1424		.1423	
% sig covariates	13% (3/23)		20% (5/25)		86% (6/7)		100% (8/8)		72% (8/11)	
Variables in Selection Fcn	$\hat{\alpha},\hat{\gamma}$	se	$\hat{\alpha},\hat{\gamma}$	se	$\hat{\alpha},\hat{\gamma}$	se	$\hat{\alpha},\hat{\gamma}$	se	$\hat{\alpha},\hat{\gamma}$	se
Constant	-3.984*	1.886	-4.712*	1.921	-2.115*	.187	-2.146*	.234	-4.202*	1.870
PSAT-M	-.0006	.0007	-.0006	.0007						
PSAT-V	-.0004	.0006	-.0003	.0006						
AGE	.142	.099	.142	.100					.112	.102
SES	.563*	.091	.548*	.091			.441*	.078	.439*	.079
FEMALE	.084	.096	.084	.096						
ASIAN	.128	.153	.138	.154						
BLACK	.078	.170	.097	.170						
HISPANIC	-.031	.163	-.028	.166						
NATIVE	-.326	.518	-.342	.518						
PRIVATE	.058	.146	.061	.148						
SCH_RUR	-.390*	.116	-.374*	.117			-.429*	.124	-.416*	.120
SCH_URB	.065	.159	.066	.159						
AP	-.052	.142	-.049	.143						
RE_ENG	.151	.200	.149	.199						
REMATH	.300	.199	.307	.194			.471*	.161		
RIG_HSP	.093	.108	.092	.108						
F1READ	.001	.008	.001	.008						
F1MATH	-.010	.009	-.010	.009						
MTHCRD	.143*	.058	.139*	.058			.138*	.055		
MTHGRD	.159	.113	.161	.113					.009	.057
F1ESTEEM	.114	.078	.117	.077						
F1LOCUS	-.093	.093	-.097	.093						
HOMEWORK	.006	.097	-.003	.097						
PARENT[a]			.695*	.191	.702*	.187			.602*	.188
MPARENT[a]			.745*	.220	.721*	.230			.688*	.231
PPRESS[a]					.677*	.130	.652*	.115	.628*	.115
MPPRESS[a]					.529*	.145	.552*	.149	.526*	.143
HWTUTOR[a]					.459*	.113	.333*	.121	.334*	.121
MHWTUTOR[a]					.560	.394			.592	.370
HI_MOT[a]					.472*	.233	.424*	.210	.447*	.205

* p-value for two-sided t-test < .05 (DEFF = 3)
N = 3,144
[a] These covariates are excluded from the regression equation

When compared using a likelihood ratio (LR) test to a baseline specification with just a constant and no predictors, all five SF specifications would be considered a statistical improvement. A variant of this approach is represented by the "Pseudo R^2" values in the third row of Table 4-2. The Pseudo R^2 for each specification is calculated as $(1 - L)/L_0$, where L is the log likelihood for a given specification of the selection function, and L_0 is the log likelihood for the baseline specification. According to this criterion, the SF4 and SF5 specifications improve model fit the best relative to the baseline model, but not by much—all five specifications are within about .04 of one another. Of the five specifications, only SF1 and SF2 are nested and can be compared directly using a likelihood ratio test. The difference in deviance between SF2 and SF1 is 11.7 with an approximate Chi-Square distribution on 2 degrees of freedom. On this basis SF1 can be rejected in favor of SF2, but no LR test can recommend SF2 over SF3, SF4 or SF5.

Another criterion we might consider in picking a "best fitting" specification is one with the largest proportion of statistically significant probit coefficient estimates. This is fairly important, since the next step of the Heckman Model is to calculate an Inverse Mills Ratio as a function of the estimated coefficients, whether they are significant or not. Naturally, the SF4 specification comes out on top here—all of its coefficients are statistically significant, because its predictors were selected with this criterion in mind. The SF3 and SF5 specifications are not far behind, with 86% and 72% of estimated coefficients statistically significant. As we might expect, SF1 and SF2 are particularly weak relative to this criterion, with only 13% and 20% of estimated coefficients statistically significant.

Figure 4-2. Predicted Probabilities of COACH = 1 for SF Specifications



For each of the k = 1 through 5 SF specifications, let $\hat{s}_{ik} = \hat{\alpha}_k + \hat{\boldsymbol{\gamma}}_k \mathbf{Z}_i$. Figure 4-2 shows the plots of the predicted probabilities of being coached as a function of $\hat{s}_{ik}$. The shape of the five curves is generally quite similar, though for SF4 and SF5 the highest estimated probability is about .2 higher at the maximum value of $\hat{s}_{ik}$. Table 4-3

compares the actual and predicted number of coached students for each specification. With the exception of SF1, all the specifications tend to underpredict the number of coached students. None of these models predicts correctly the coaching status for more than about 20% of those students who were actually coached.

Table 4-3. Predicted Coaching Status by Selection Function

		Specifications of Selection Function				
		SF1	SF2	SF3	SF4	SF5
COACH = 0	N	2641	2641	2641	2641	2641
	Sum of P(C=1)	400.2	371.3	352.3	365.4	360.2
	Mean P(C=1)	0.15	0.14	0.13	0.14	0.14
	Median P(C=1)	0.13	0.12	0.08	0.11	0.11
	Max P(C=1)	0.59	0.58	0.62	0.79	0.76
COACH = 1	N	503	503	503	503	503
	Sum of P(C=1)	114.0	111.4	91.6	119.6	116.1
	Mean P(C=1)	0.23	0.22	0.18	0.24	0.23
	Median P(C=1)	0.21	0.21	0.17	0.21	0.21
	Max P(C=1)	0.59	0.58	0.58	0.74	0.73
TOTAL	N	3144	3144	3144	3144	3144
	Sum of P(C=1)	514.2	482.7	443.9	484.9	476.2
	Mean P(C=1)	0.16	0.15	0.14	0.15	0.15
	Median P(C=1)	0.15	0.13	0.08	0.12	0.12
	Max P(C=1)	0.59	0.58	0.62	0.79	0.76

P(C) = 1 is the estimated probability of being coached for student *i* given $\mathbf{Z}_i$

The point of these model comparisons is that in most applications of the Heckman Model, precious little ink has been spent validating selection function specifications. Seldom are alternate specifications compared, and it is even more seldom that there is any theory to bolster the specification ultimately chosen. The decision of what predictors to include or exclude from the selection function is a non-trivial one, and can have substantial ramifications on the estimated parameters generated by the Heckman Model, as I demonstrate below.

Heckman Model Estimates

In chapter 2 I showed how parameter estimates from the selection function are used to calculate estimated values for $\lambda_i(COACH_i, \hat{s}_i)$, where

$$\lambda_i(COACH_i, \hat{s}_i) = COACH_i\left(\frac{\phi(\hat{s}_i)}{1-\Phi(\hat{s}_i)}\right) + (1-COACH_i)\frac{-\phi(\hat{s}_i)}{\Phi(\hat{s}_i)}. \quad (4)$$

The histogram of $\lambda_i(COACH_i, \hat{s}_i)$ estimated for SF5 is shown in Figure 4-3. It is bimodal, comprised of two different Inverse Mills Ratios for uncoached and coached students. Uncoached students are represented by the large cluster of values on the left, and coached students are represented by the smaller cluster of values on the right. According to the Heckman Model, coached students have $E(\delta_i \mid \delta_i > s_i) = \frac{\phi(s_i)}{1-\Phi(s_i)}$, while uncoached students have $E(\delta_i \mid \delta_i \le s_i) = -\frac{\phi(s_i)}{\Phi(s_i)}$, and this is reflected by the histogram of estimated values.

Figure 4-3. Histogram of Inverse Mills Ratio Estimated for SF5

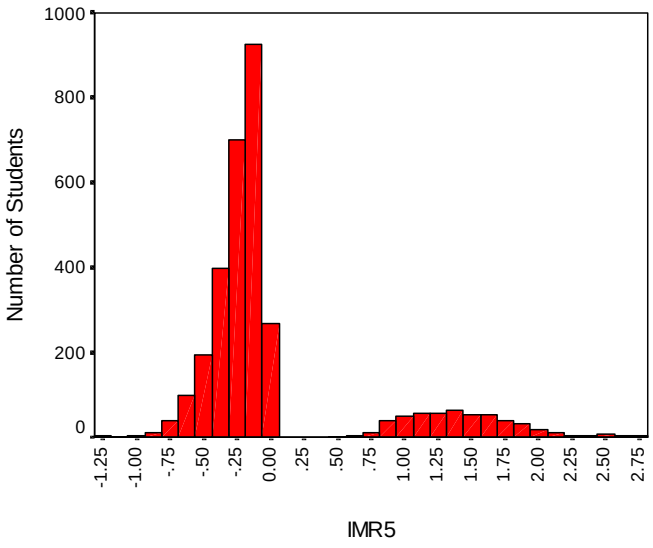


Table 4-4. SAT-V Coaching Effects using the Heckman Model

	SF1		SF2		SF3		SF4		SF5	
R^2	.776		.777		.776		.776		.776	
adj R^2	.771		.772		.770		.770		.770	
Variables in Regression Eqn	$\hat{a},\hat{b},\hat{\mathbf{c}},\hat{h}$	se	$\hat{a},\hat{b},\hat{\mathbf{c}},\hat{h}$	se	$\hat{a},\hat{b},\hat{\mathbf{c}},\hat{h}$	se	$\hat{a},\hat{b},\hat{\mathbf{c}},\hat{h}$	se	$\hat{a},\hat{b},\hat{\mathbf{c}},\hat{h}$	se
Constant	177.0*	65.4	168.3*	64.8	144.9*	63.6	144.8*	63.6	144.7*	63.9
COACH	69.2*	29.6	58.2*	26.2	0.1	14.9	16.6	15.4	12.4	14.9
PSAT-M	0.07*	0.03	.06*	.03	.05*	.03	.06*	.03	.05*	.03
PSAT-V	0.61*	0.02	.61*	.02	.61*	.02	.61*	.02	.61*	.02
AGE	-10.7*	3.5	-10.2*	3.5	-8.6*	3.4	-8.7*	3.4	-8.7*	3.4
SES	-2.5	4.3	-1.2	3.8	4.2	2.4	3.2	2.9	3.7	2.9
FEMALE	-6.4	3.4	-6.1	3.3	-4.7	3.3	-5.1	3.3	-5.0	3.3
ASIAN	6.4	6.1	6.7	6.1	7.8	6.2	8.0	6.2	7.9	6.2
BLACK	-4.1	5.6	-3.8	5.6	-3.5	5.6	-3.4	5.6	-3.4	5.6
HISPANIC	-3.1	6.1	-2.8	6.1	-3.1	6.1	-3.1	6.1	-3.1	6.1
NATIVE	-2.9	26.3	-3.8	25.8	-6.7	25.4	-6.0	25.4	-6.2	25.4
PRIVATE	8.6*	4.2	8.7*	4.2	8.7*	4.3	9.0*	4.3	8.9*	4.3
SCH_RUR	-2.5	4.6	-3.5	4.4	-6.6	4.0	-6.2	4.1	-6.5	4.1
SCH_URB	-1.0	3.7	-0.3	3.7	1.3	3.6	1.0	3.6	1.1	3.6
AP	12.5*	3.4	12.8*	3.4	12.5*	3.3	12.4*	3.3	12.4*	3.3
RE_ENG	-13.9	7.4	-13.2	7.4	-11.2	7.4	-11.5	7.4	-11.5	7.4
REMATH	-2.5	7.4	-1.8	7.3	1.5	7.1	1.1	7.2	1.7	7.1
RIG_HSP	-2.7	3.1	-2.3	3.1	-1.0	3.1	-1.3	3.1	-1.2	3.1
F1READ	2.5*	0.3	2.5*	0.3	2.5	0.3	2.5	0.3	2.5	0.3
F1MATH	0.4	0.4	0.4	0.4	0.3	0.4	0.4	0.4	0.4	0.4
MTHCRD	-3.0	2.4	-2.6	2.4	-1.3	2.3	-1.4	2.3	-1.3	2.3
MTHGRD	1.1	2.7	1.7	2.6	4.0	2.5	3.3	2.5	3.5	2.5
F1ESTEEM	3.3	2.9	3.8	2.9	5.2	2.8	5.1	2.8	5.1	2.8
F1LOCUS	-4.2	3.3	-4.8	3.3	-6.1	3.2	-6.1	3.2	-6.1	3.2
HOMEWORK	3.8	3.1	3.6	3.1	3.6	3.2	3.4	3.2	3.5	3.2
IMR1	-32.3*	16.2								
IMR2			-26.2	14.4						
IMR3					6.5	8.4				
IMR4							-3.2	8.7		
IMR5									-0.8	8.4
$\hat{\rho}$ of (δ_i, ε_i)	-.60		-.42		.15		-.05		-.01	

N = 3,144 [effective sample size = 1,015]
* p-value < .05 (based standard errors with DEFF = 3)

IMR1 = all covariates in regression eqn used in selection eqn
IMR2 = all covariates in regression eqn + 1 covariate (*PARENT*) not used in reg eqn
IMR3 = only covariates not used in reg eqn, all dummies (*HWTUTOR*, *PARENT*, *PPRESS*, *HI_MOT*)
IMR4 = covariates chosen by stepwise selection (*SCH_RUR*, *PPRESS*, *HWTUTOR*, *REMATH*, *HI_MOT*, *SES*, *MTHCRD*
IMR5 = covariates that were stat sig in coaching crosstabs (*AGE*, *SES*, *MTHGRD*, *SCH_RUR*, *HWTUTOR*, *PARENT*, *PPRESS*, *HI_MOT*)

Table 4-5. SAT-M Coaching Effects using the Heckman Model

	SF1		SF2		SF3		SF4		SF5	
R^2 adj R^2	.801 .796									
Variables in Regression Eqn	$\hat{a},\hat{b},\hat{\mathbf{c}},\hat{h}$	se	$\hat{a},\hat{b},\hat{\mathbf{c}},\hat{h}$	se	$\hat{a},\hat{b},\hat{\mathbf{c}},\hat{h}$	se	$\hat{a},\hat{b},\hat{\mathbf{c}},\hat{h}$	se	$\hat{a},\hat{b},\hat{\mathbf{c}},\hat{h}$	se
Constant	24.2	67.9	10.5	67.3	-8.3	66.1	-4.2	66.0	1.7	66.3
COACH	78.7*	30.2	58.8*	27.5	30.1	15.5	46.4*	15.9	41.6*	15.4
PSAT-M	.42*	.03	.42*	.03	.42*	.03	.42*	.03	.42*	.03
PSAT-V	.09*	.02	.09*	.02	.09*	.02	.09*	.02	.09*	.02
AGE	-4.8	3.6	-3.9	3.6	-2.7	3.5	-2.9	3.5	-3.4	3.5
SES	3.7	4.4	6.0	3.9	9.9*	2.5	7.3*	3.0	7.7*	3.0
FEMALE	-17.3*	3.5	-16.9*	3.5	-16.4*	3.4	-16.8*	3.4	-16.6*	3.4
ASIAN	3.1	6.3	3.6	6.4	4.9	6.4	5.1	6.4	5.0	6.4
BLACK	-14.4*	5.8	-14.4*	5.8	-14.2*	5.8	-14.1*	5.8	-14.1*	5.8
HISPANIC	-3.4	6.3	-4.0	6.3	-4.6	6.3	-4.5	6.3	-4.6	6.3
NATIVE	-21.2	27.5	-23.9	26.6	-25.7	26.4	-24.7	26.4	-25.1	26.4
PRIVATE	-1.8	4.3	-1.3	4.4	-0.7	4.4	-0.4	4.4	-0.4	4.4
SCH_RUR	0.1	4.8	-1.0	4.5	-3.5	4.1	-1.5	4.3	-1.8	4.3
SCH_URB	-0.1	3.8	0.3	3.8	1.1	3.8	0.6	3.8	0.8	3.8
AP	9.4*	3.5	9.2*	3.5	8.7*	3.5	8.6*	3.5	8.7*	3.5
RE_ENG	6.4	7.6	6.9	7.7	7.9	7.7	7.7	7.7	7.9	7.7
REMATH	-24.2*	7.7	-22.2*	7.6	-18.9*	7.3	-21.8*	7.5	-18.8*	7.3
RIG_HSP	1.1	3.2	1.8	3.2	2.6	3.2	2.4	3.2	2.4	3.2
F1READ	-0.5	0.3	-0.5	0.3	-0.5	0.3	-0.5	0.3	-0.5	0.3
F1MATH	5.1*	0.4	5.0*	0.4	5.0*	0.4	5.0*	0.4	5.0*	0.4
MTHCRD	7.2*	2.5	7.8*	2.5	8.8*	2.4	8.1*	2.4	8.8*	2.4
MTHGRD	12.2*	2.8	13.3*	2.7	14.4*	2.6	13.8*	2.6	14.0*	2.6
F1ESTEEM	-3.5	3.0	-2.9	3.0	-1.9	2.9	-2.2	2.9	-2.0	2.9
F1LOCUS	-0.6	3.4	-1.2	3.4	-2.1	3.4	-1.9	3.4	-2.0	3.4
HOMEWORK	1.4	3.3	1.4	3.3	1.3	3.3	1.2	3.3	1.3	3.3
IMR1	-33.1*	16.5								
IMR2			-22.0	15.1						
IMR3					-6.4	8.7				
IMR4							-16.0	8.9		
IMR5									-13.3	8.7
$\hat{\rho}$ for (δ_i, ε_i)	-.64		-.36		-.10		-.25		-.20	

N = 3,144 [effective sample size = 1,015]
* p-value < .05 (based standard errors with DEFF = 3)

IMR1 = all covariates in regression eqn used in selection eqn
IMR2 = all covariates in regression eqn + 1 covariate (*PARENT*) not used in reg eqn
IMR3 = only covariates not used in reg eqn, all dummies (*HWTUTOR*, *PARENT*, *PPRESS*, *HI_MOT*)
IMR4 = covariates chosen by stepwise selection (*SCH_RUR*, *PPRESS*, *HWTUTOR*, *REMATH*, *HI_MOT*, *SES*, *MTHCRD*
IMR5 = covariates that were stat sig in coaching crosstabs (*AGE*, *SES*, *MTHGRD*, *SCH_RUR*, *HWTUTOR*, *PARENT*, *PPRESS*, *HI_MOT*)

Using Equation 4, $\lambda_{ik}(COACH_i, \hat{s}_{ik})$ can be estimated for the k = 1, ..., 5 SF specifications. For the second step of the Heckman Model I proceed by including $\lambda_{ik}(COACH_i, \hat{s}_{ik})$ as a covariate in the regression of Y_i on a constant, $COACH_i$, and $\mathbf{X}_i$. All cases are weighted by $DESWGT_i$ with a *DEFF* of 3. In addition, because the conditional variance of ε_i under the Heckman Model is heteroskedastic, a generalized least squares fitting procedure (Greene, 1981) is used to get efficient standard error estimates for the regression coefficients. Tables 4-4 and 4-5 report the results of these regressions for SAT-V and SAT-M test scores.

The estimated effects for *COACH* vary, sometimes dramatically, depending upon which version of $\lambda_{ik}(COACH_i, \hat{s}_{ik})$ is included in the Heckman Model. For specifications with *SAT-V* as the dependent variable, the estimated coaching effect ranges from a low of 0 points to a high of 69 points. For specifications with *SAT-M* as the dependent variable, the estimated coaching effect ranges from a low of 30 points, to a high of 80 points. Parameter estimates for covariates under all five specifications of the Heckman Model with either *SAT-V* or *SAT-M* as the dependent variable are generally similar to those from the linear regression model.

Depending upon which selection function specification we consider, the Heckman Model tells us a different story about the nature of selection bias in SAT coaching. In models with SAT-V as the dependent variable, the estimated correlation $\hat{\rho}$ between δ_i and ε_i is -.60 and -.42 for SF1 and SF2, but close to zero for SF4 and SF5. When SAT-M is the dependent variable, the estimated correlation is -.64 for SF1, but between -.36 and

-.10 for SF2 through SF5.

Only in the SF1 specification of the model is the parameter estimate for $\lambda_{ik}(COACH_i, \hat{s}_{ik})$ also statistically significant, indicating the presence of selection bias. For these (as well as most other) specifications, the estimated negative correlations between δ_i and ε_i would suggest that the students who are more likely to get coached are the ones who are *less* likely to perform well on a particular section of the SAT. If these versions of the Heckman Model were to be believed, we would expect the coaching effects estimated by the linear regression model to be biased downwards. On the other hand, most specifications of the Heckman Model considered here suggest that any selection bias in the data is not statistically significant.

Multicollinearity helps explain why coaching effect estimates vary so dramatically, with large standard errors, under different specifications of the Heckman Model selection function. In particular, the variable $COACH_i$ and $\lambda_{ik}(COACH_i, \hat{s}_{ik})$ are strongly correlated, which follows from the fact that the latter is defined as an interaction with the former. When the $\lambda_{ik}(COACH_i, \hat{s}_{ik})$ based on SF1 and SF2 are regressed on a constant, $COACH_i$ and $\mathbf{X}_i$, the respective adjusted R^2's are .98 and .97. Likewise, the regressions based on SF3, SF4 and SF5 have adjusted R^2's of .92, .94 and .92 . Greene (1993) points out three symptoms typically associated with multicollinearity in regression models.

1) Small changes in the data structure can produce wide swings in parameter estimates.

2) Coefficients may have very high standard errors and low significance levels in spite of the fact that they are jointly highly significant.
3) Coefficients will have the wrong sign or implausible magnitude.

These symptoms are evident in the estimates of coaching effects generated under the Heckman Model:

- adding a single predictor to the selection function from SF1 to SF2 decreases the Heckman Model SAT-V coaching estimate from 79 to 59 points;
- across all Heckman Model specifications, standard error estimates for the coaching effect are consistently high (the lowest is 15 points);
- and at least two of the SAT effect estimates (SAT-V and SAT-M effect under SF1) are of an arguably implausible magnitude relative to previous observational SAT coaching studies.

To see more clearly the collinear relationship between the variable $COACH_i$ and $\lambda_{i5}(COACH_i, \hat{s}_{i5})$, I subtract from each variable its predicted value when regressed on $\mathbf{X}_i$. The resulting variable is the residual component not predicted by $\mathbf{X}_i$. The two residualized variables— $COACH_i r$ and $\lambda_{i5}(COACH_i, \hat{s}_{i5})r$ —are plotted in Figures 4-4 and 4-5 for the conditions $COACH_i = 1$ and $COACH_i = 0$. The correlation between the residualized variables is still about .73.

Figure 4-4. Collinearity when COACH = 1 (ρ = .72)

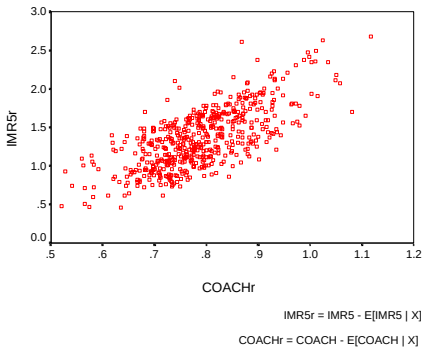
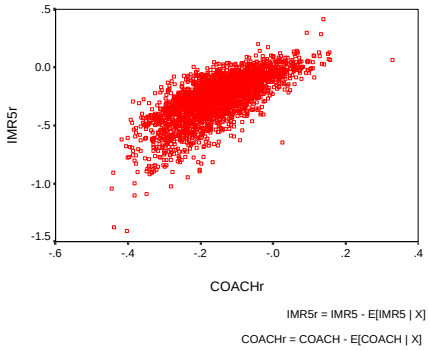


Figure 4-5. Collinearity when COACH = 0 (ρ = .74)



The easiest solution to the multicollinearity problem is to omit one or more covariates from the regression equation. But this is no real solution to the problem because we have now violated our behavioral model—any decrease in multicollinearity will come with a potential increase in bias. Other solutions have been proposed and applied to handle collinear data without omitting variables (c.f. ridge regression and principal components analysis described in Greene, 1993, p. 270-273). A detailed discussion of these methods is outside the scope of this dissertation, but it is important to note that "solutions" to multicollinearity have their own associated problems. To the extent that such methods change the structure and relationship of the data under

consideration, they will almost certainly change the causal interpretation of the Heckman Model as presented here.

4.3 Comparisons

Figures 4-6 and 4-7 compare the estimated SAT-V and SAT-M coaching effects estimated by 1) taking the difference in average scores between coached and uncoached students, 2) using linear regression and 3) using the five Heckman Model specifications. I include around each point estimate the corresponding 95% confidence interval.

Figure 4-6. Comparison of SAT-V Coaching Effect Estimates

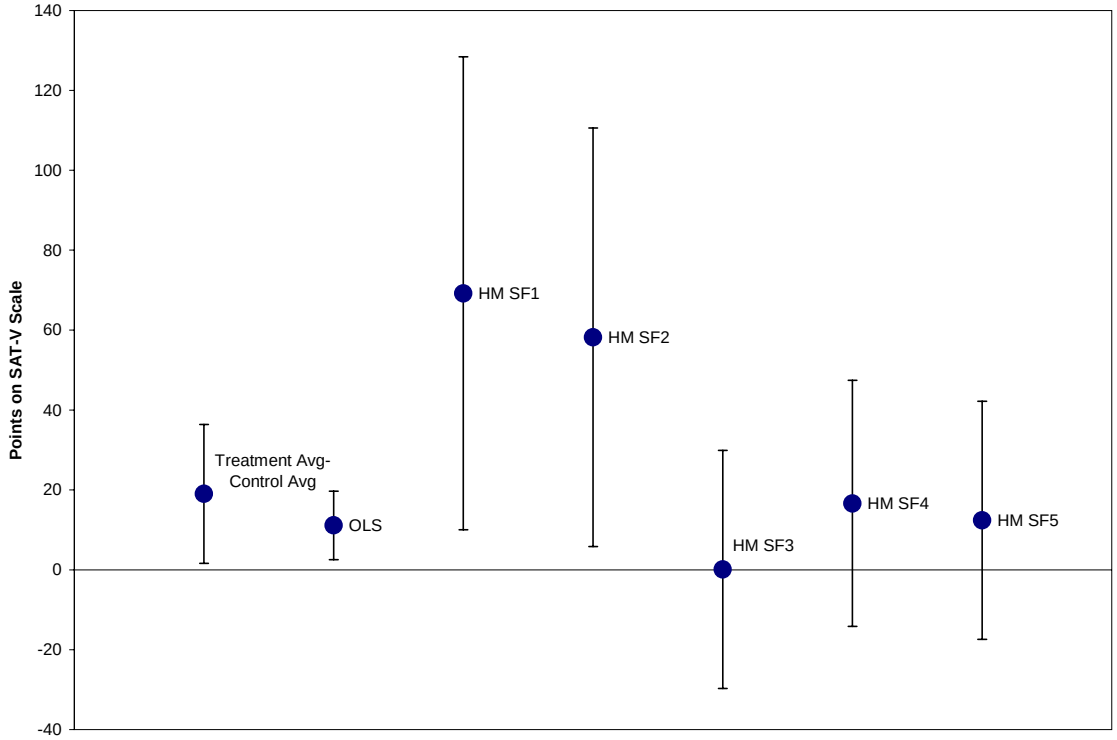
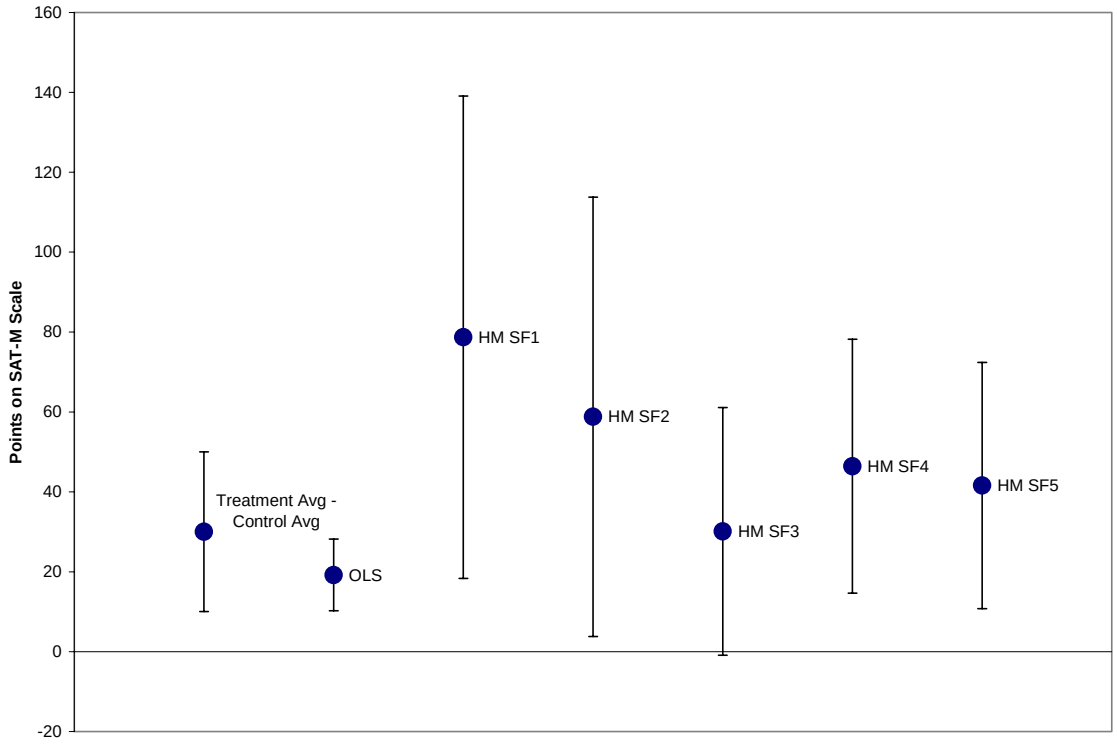


Figure 4-7. Comparison of SAT-M Coaching Effect Estimates



For the SAT-V, the linear regression model produces a statistically significant point estimates of about 11 points for the coaching effect. The Heckman Model produces effect estimates ranging from 0 to 70 points, only two of which (SF1 and SF2) are statistically significant. If the SF1 and SF2 specification of the Heckman Model are ignored, the SAT-V effect estimates from both models are smaller than what would be estimated by simply taking the average difference in SAT-V scores for coached and uncoached students. For the SAT-M, the Heckman Model produces coaching effect estimates ranging from 30 to 70 points—estimates that are generally more than twice as large as the 19 point estimate produced under linear regression. The SAT-M coaching effect estimates are generally statistically significant under both models. Under the Heckman Model the estimates tend to be larger (SF 3 is the exception) than what would

be estimated by simply taking the difference in the average SAT-M scores for coached and uncoached students, while under linear regression the estimate is smaller.

Table 4-6. Comparing Commercial Coaching Effects by Model and Study

Source of Effect Estimate	SAT-V Effect	SAT-M Effect
Linear Regression		
NELS	11*	19*
Powers & Rock (1999)	6*	16*
Smyth (1990)	9*	18*
Zuman (1988)	52*	58*
FTC (1978) [Company A]	28*	24*
FTC (1978) [Company B]	2	4
Heckman Model		
NELS (SF1)	69*	79*
NELS (SF2)	58*	59*
NELS (SF3)	0	30
NELS (SF4)	17	46*
NELS (SF5)	12	42*
Powers & Rock (1999)	12*	13*

* p-value < .05
NOTE: All effect estimates are based on observational data for commercial coaching programs.

There is no absolute criterion against which to compare the coaching effects estimated by linear regression and the Heckman Model. Only one randomized study has been conducted for commercial coaching programs (Zuman, 1988), but the sample considered was quite small, and other methodological problems make the results equivocal. Relative comparisons to the effects estimated in other observational studies of commercial coaching may be of interest nonetheless. Table 4-6 makes this comparison for effects estimated in studies using the linear regression model and the Heckman Model. Note that while only the Powers & Rock data is similar in its national representativeness to the data in the NELS, if the principal assumptions of the linear regression model hold, the estimated coaching effects should only differ as a function of

the bias induced by the omission of certain covariates from $\mathbf{X}_i$. In theory, if all studies had collected the same covariates, and used sound sample designs, by virtue of the constancy constraint we would expect the estimated coaching effects to be the generally the same, regardless of the specific sample used.

According to Table 4-6, the two most recent observational coaching studies using linear regression (Powers & Rock, 1999; Smyth 1990) both generate estimated coaching effects that are within 3 to 8 points of the NELS-based estimate. The Zuman and FTC studies suggest evidence of larger coaching effects based on the linear regression model, and both would seem to call into question the constancy assumption: the Zuman study because it was based on a very specific sample of high SES students, the FTC study because different effects were found for two different coaching programs. The validity of the constancy assumption would seem to be a bigger threat to the linear regression model than the threat of bias due to confounding variables. This is probably because all of the studies using linear regression in Table 4-6 included pre-coaching SAT or PSAT scores in $\mathbf{X}_i$. The covariates are so strongly correlated to post-coaching SAT scores, that the addition of other covariates to control for confounding, while theoretically important, may have little to no empirical impact on the model. This is certainly the case for the NELS data. The addition of all other covariates to a model with only PSAT-V and PSAT-M scores in $\mathbf{X}_i$ decreases the estimated SAT-M and SAT-V coaching effects by just 3 points.

Comparisons of the Heckman Model are harder to make in Table 4-6 since from study to study both $\mathbf{X}_i$ and $\mathbf{Z}_i$ will vary, and I have already demonstrated the sensitivity of Heckman Model estimates to different selection function specifications. Furthermore, outside of the analysis in this dissertation, only the Powers & Rock study has used the Heckman Model to estimate coaching effects. The covariates and predictors available in the Powers & Rock data, while not quite of the same quality as some of those available from NELS, were fairly similar. In their regression equation Powers & Rock included covariates for PSAT or first SAT scores, father's education, student high school GPA, math GPA, race/ethnicity and two measures of student motivation.[4] Their selection function included all the same variables, and also included student's GPA in high school social science courses. This specification of the Heckman Model is probably most comparable to my SF2. Powers & Rock's SAT-V coaching effect estimates produced using the Heckman Model were similar only to those produced under SF4 and SF5 with the NELS data; for the SAT-M their effect estimate was generally less than a third of the NELS-based estimates. Powers & Rock also estimated standard errors that were on the whole much smaller than those found in the analysis of the NELS data, in part perhaps because their data structure did not require a design effect correction.

The empirical analysis in this chapter has hopefully shed some light on the use of the linear regression and Heckman Model approaches to estimate unbiased effects of SAT coaching with observational data. Researchers must be quite cautious in using these methods to draw causal conclusions, particularly given the types of assumptions

[4] This information was not included in their published study of 1999, but was provided to me in a personal communication (Rock, 2002).

described in chapter 2. In the social sciences, bias in estimated effects can never be completely ruled out. However, it may be reduced, possibly to trivial amounts, in the rare event when a researcher has access to covariates that are highly correlated with the outcome measure. This is not an established theorem, but seems like a reasonable conjecture (see Holland, 2001). Extreme caution should be exercised before applying the Heckman Model as a means of drawing causal inferences about a treatment effect. There is seldom any theory to guide the specification of the selection function, and if the selection function is specified just with the objective of identifying the model (e.g. SF1 and SF2), the resulting effect estimates will probably be highly questionable, if not completely out of whack. Once a selection function has been specified, estimated, and used to calculate $\lambda_i(COACH_i, \hat{s}_i)$, the next concern should be the potential for multicollinearity between the covariates, the treatment variable, and $\lambda_i(COACH_i, \hat{s}_i)$, with most of the problem stemming from the collinearity among the latter terms. When multicollinearity is a problem, it may cast doubt on both the estimated treatment effect and the standard errors around the treatment effect. All too often the Heckman Model has been applied in the social science with little to no discussion of these issues. With access to the right software (e.g. STATA, LIMDEP), the Heckman Model is easily implemented with seemingly obvious causal conclusions. I would suggest that when this is takes place without a compelling theoretical rationale and a careful scrutiny of the data, such conclusions are of dubious value.

CHAPTER 5: IS THERE ONE COACHING EFFECT?

A constancy constraint is made explicitly when a single coaching effect is estimated as the parameter $\hat{b}$. When all the assumptions necessary for using the linear regression model to make causal inferences hold, including the constancy constraint, it makes sense to interpret $\hat{b}$ as the number of points by which we can expect a student's SAT score to change after exposure to commercial coaching. If we relax the constancy constraint, then this kind of a causal conclusion would need to be amended, because the estimated coaching effect will be higher or lower for certain types of students and certain types of coaching. I take up this issue in the context of the linear regression model. First, I consider whether the way coaching is defined changes the estimated effect. Next, I consider whether there are interactions between coaching and student characteristics. Finally, I check whether the same coaching effect applies to different samples of students.

5.1 Alternate Definitions of the Coaching Treatment

In chapters 3 and 4, coaching was defined as participation in a course offered by a commercial test preparation service. Students were also asked if they had prepared for the SAT with other test preparation activities, many of which have also been classified as coaching in previous studies. Table 5-1 presents the percentage of students from the NELS POP1 subsample by each of the six possible modes of test preparation. While a

fairly small number of students report having prepared for the SAT with a private tutor (7%) or video (6%), a more substantial number of students report having taken a commercial course (15%), a high school course (23%) or having studied with computer software (17%). More than 60% of students taking both the PSAT and SAT prepared for the test by using a book. Close to three quarters of all students prepared for the SAT with at least one of these six activities. It is thus important to recognize that when it comes to preparing for the SAT, only a minority of students appear to do "nothing." If each test preparation activity is assumed to have a positive effect on SAT performance, then because students enrolled in commercial courses (i.e. $COACH_i = 1$) are also more likely to prepare for the SAT with other preparatory activities, the coaching effect estimated in chapter 4 may be confounded.

Table 5-1. Six Forms of Preparation for the SAT

Category	Percent of POP1 Subsample by Test Preparation Activity					
	Commercial Course	High School Course	Private Tutor	Book	Video	Computer
0 ("no")	85	77	93	38	94	83
1 ("yes")	15	23	7	62	6	17

NOTE: Listwise deletion across six test prep categories reduced N from 3,504 to 3,479

We can define coaching differently if we apply the information about all six forms of test preparation. Let six dummy variables take a value of 1 respectively if a student prepared for the SAT by i) taking a commercial course (*COACH*)[1], ii) taking a class offered by their high school (*PREPHS*), iii) studying with a private tutor

[1] To be consistent with Chapter 4, I still use the label *COACH* to represent a student who enrolled in a commercial course, but in this context, the variable label should not be confused with the full coaching treatment, defined here are as any combination of the six possible test preparation activities.

(*PREPTUT*), iv) studying with a book (*PREPBOOK*), v) studying with a video tape (*PREPVID*), or vi) studying with computer software (*PREPPC*). I now include these six variables along with all possible two-way interactions between the variables, with the exception of *PREPBOOK*. Two-way interactions with *PREPBOOK* are excluded because they add little new information and add a strong amount of collinearity to the model. For example, 95% of students who prepared for the SAT by using a private tutor also reported that they had prepared by using a book. Generally, if we know a student has prepared for the SAT by any of the five other preparation activities, it is very likely they have also used a book. Higher-order interactions among the preparatory activities are certainly possible, but beyond two-way interactions the numbers of cases become very sparse, so for this reason and for the sake of parsimony, I only consider two-way interactions. The effect of coaching now has an entirely different interpretation. Before, the treatment effect was equal to the parameter b. Now, the treatment effect is equal to the sum of the parameters for as many as six main effects and ten interactions, depending upon the particular combination of test preparation activities practiced by a given student. The results of estimating these new parameters using the linear regression model are shown in Table 5-2.

The main SAT-V and SAT-M effect estimates for the variable *COACH* have not changed from the linear regression model in Chapter 4. The broader definition of the coaching treatment has not served to confound the marginal interpretation for the effect of a commercial course. The standard errors of the estimates are somewhat larger, primarily because *COACH* has a collinear relationship with its four interaction terms.

The main effect estimates of the variables *PREPHS*, *PREPTUT*, *PREPBOOK*, *PREPVID* and *PREPPC* tend to be fairly small. The exceptions are the SAT-M effect estimates for *PREPHS* and *PREPTUT*, which are a statistically significant 9 and 28 points respectively. It appears that the single most effective way for a student to score higher on the SAT-M is by preparing with a private tutor. Not coincidentally, it is also the most expensive means of preparation, to which the fewest number of students have access.

Table 5-2. Linear Regression with Test Prep Interactions

	SAT-V (mean = 447, sd = 101)			SAT-M (mean = 504, sd = 116)		
R^2	.779			.805		
adjusted R^2	.776			.803		
COACH=1/Total	493/3128			493/3128		
Treatment	Effect Estimate	Std Error Range *DEFF*=1	*DEFF*=3	Effect Estimate	Std Error Range *DEFF*=1	*DEFF*=3
COACH	11.4*	3.4	6.1	19.8*	3.5	6.3
PREPHS	-0.1	2.5	4.5	9.3*	2.6	4.6
PREPTUT	8.7	4.8	8.5	27.6*	5.0	8.8
PREPBOOK	4.0	1.8	3.2	1.7	1.9	3.4
PREPVID	2.6	5.2	9.3	-3.6	5.4	9.6
PREPPC	-9.1	3.2	5.6	-6.6	3.3	5.8
COACH x PREPHS	-3.6	5.6	10.0	-18.3	5.8	10.3
COACH x PREPTUT	-13.0	7.3	12.9	-8.3	7.5	13.3
COACH x PREPVID	-0.3	10.4	18.3	-6.0	10.7	19.0
COACH x PREPPC	17.2	6.0	10.7	15.3	6.2	11.0
PREPHS x PREPTUT	10.4	7.8	13.7	-4.1	8.0	14.2
PREPHS x PREPVID	-1.1	8.0	14.3	5.1	8.3	14.7
PREPHS x PREPPC	5.5	5.0	8.8	1.5	5.1	9.1
PREPTUT x PREPVID	-29.5	11.5	20.4	3.7	11.9	21.1
PREPTUT x PREPPC	-6.7	8.4	14.8	-12.1	8.6	15.3
PREPVID x PREPPC	-7.4	8.9	15.7	-12.2	9.2	16.3

* p-value for two-sided t-test < .05 across standard error range
Covariates (**X**) are the same as those specified in chapter 4.

None of the parameters for the two-way interaction terms in Table 5-2 are statistically significant across the range of standard error estimates, but some are significant at the lower bound. The negative parameter estimate for the *COACH x PREP_TUT* interaction term suggests that for students who are helped by a private tutor

and also enroll in a commercial preparatory course, the coaching effect is less than the sum of the two main effects. Also of interest are positive and fairly large SAT-V and SAT-M interaction parameter estimates for *COACH x PREP_PC*. This suggests that while the main effect of preparing with a computer is negative, the combination of preparing with a computer and a commercial course adds about 8 points per SAT section over preparing with a commercial course alone.

If coaching is given this broader definition as any combination of six preparatory activities, the largest coaching effect estimate is about 40 points, all else held constant, for students preparing for the SAT-M with commercial coaching and a private tutor. On the SAT-V the largest estimated coaching effect is about 18 points for students preparing with a commercial course and computer software. Note that according to the specification of the linear regression model reflected in Table 5-2, the coaching effect does not necessarily increase with the quantity of test preparation activities. In fact, the estimated effect for a student doing all six activities is -11 points on the SAT-V and 13 points on the SAT-M.

In chapter 4, a principal constancy constraint being made was that the effect of coaching is the same for all commercial program types. There is no good way to test this constraint directly with the NELS data, because students were not asked to provide the names of the commercial programs they had taken. The notion that all commercial programs are of the same quality may be problematic. Seppy Basili, vice president for learning and assessment at Kaplan Inc., was quoted in the *New York Times* (Kolata, 2001)

as saying "What we've seen over the past 15 years is this huge increase in weekend courses and one-day courses.[2] The whole notion of grouping commercial courses with this broad brush causes a problem for us." This suggests that there are two types of commercial coaching programs offering either high or low quality test preparation. The proposition is that companies offering high quality coaching for the SAT are misrepresented by the NELS data because they cannot be differentiated from companies offering low quality coaching.

This proposition can be tested in part with the NELS data, given certain assumptions. In two different national surveys of the test preparation activities of students taking the SAT, Powers (1988; 1998) found that about 12% of all students reported that they had been coached by commercial companies. This is not far off the 15% estimate for the NELS POP1 subsample. Of students who reported taking commercial courses in the Powers survey, 40% indicated that they had been coached specifically by one of the two largest companies offering these services: Kaplan or The Princeton Review. These percentages did not change much over the 10 year period between 1986 to 1996. It seems reasonable then to presume that about 40% of the coached students in the NELS sample were enrolled in Kaplan or The Princeton Review. That is, about 201 of the 503 coached students in the NELS POP1 subsample were probably coached by one of the two largest coaching companies. Both companies claim that the effects of their coaching program are in the range of about 50-70 points per section. If this were true, how low must the effects for the other 302 coached students

[2] As I showed in chapter 1, though it seems intuitively reasonable, a clear empirical relationship between coaching duration and effect has not been established.

have been in order to arrive at the estimated average SAT-V and SAT-M effects of 11 and 19 points found for the full sample?

Some quick algebra answers the question. If the SAT-V effect for just the students coached by Kaplan and The Princeton Review is actually 50 points, then to arrive at an effect estimate for all coached students of 11 points, the effect for students coached by all other programs would have to be -15 points: (i.e. solve for "x" in [50*201+ x*302])/503 = 11). Likewise, the SAT-M effect estimate for students coached by all other programs would have to be -2 points. In other words, if the claims by Kaplan and The Princeton Review are to be believed, then the coaching effects of all other commercial companies must be *negative*—that is, students taking these courses do worse on the SAT than they would have done if they had not taken the course. Such a scenario is not impossible to imagine, but seems highly implausible.

5.2 Testing Interactions with the Coaching Treatment

Whether we accept the assumption that commercial coaching has a constant effect across program types or not, it seems reasonable to suspect that certain types of students will benefit more from coaching than others. To this end I consider all possible two-way interactions of covariates and *COACH* in the linear regression model. When these interaction effects are tested with standard errors based on a conservative design effect correction (*DEFF* = 3), there are no statistically significant interactions with coaching for

either the SAT-V or SAT-M. Only when standard errors are based upon no design effect correction (*DEFF* = 1) is there any evidence of statistically significant interactions.

Table 5-3. Linear Regression Model with Covariate Interactions

	SAT-V (mean = 447, sd = 101)			SAT-M (mean = 504, sd = 116)		
R^2	.788			.822		
adj R^2	.783			.818		
COACH=1/Total	503/3144			503/3144		
Variables in Regression Eqn	$\hat{a},\hat{b},\hat{\mathbf{c}}$	Std Error Range DEFF=1	DEFF=3	$\hat{a},\hat{b},\hat{\mathbf{c}}$	Std Error Range DEFF=1	DEFF=3
Constant	145.7	36.0	63.5	4.8	37.6	66.3
COACH	12.5	4.8	8.4	14.1	3.3	5.9
COACH x SES	10.4	3.8	6.8	---	---	---
COACH x HISPANIC	-24.1	10.7	18.8	---	---	---
COACH x AP	-13.0	4.9	8.6	---	---	---
COACH x SCH_URB	---	---	---	12.3	5.0	8.8
COACH x F1MATHr	---	---	---	1.5	0.5	0.9
COACH x PSAT-Mr	---	---	---	-.08	.04	.07
PSAT-M	.06	.02	.03	.43	.02	.03
PSAT-V	.61	.01	.02	.09	.01	.02
AGE	-5.2	1.9	3.4	-2.9	2.0	3.5
SES	8.7	3.5	2.5	9.9	1.4	2.5
FEMALE	-2.9	3.2	3.3	-16.3	1.9	3.4
ASIAN	-0.7	3.6	6.2	5.0	3.6	6.4
BLACK	-24.1	10.7	5.6	-13.1	3.3	5.8
HISPANIC	-6.8	14.4	6.4	-4.2	3.6	6.3
NATIVE AMERICAN	10.1	2.4	25.3	-25.9	14.9	26.3
PRIVATE	-6.6	2.3	4.3	-0.8	2.5	4.4
SCH_RUR	0.2	2.1	4.0	-3.8	2.4	4.2
SCH_URB	14.5	2.0	3.6	-1.1	2.3	4.0
AP	.06	.02	3.6	9.1	2.0	3.5
RE_ENG	-10.6	4.2	7.4	8.3	4.4	7.7
REMATH	1.1	4.0	7.0	-19.0	4.2	7.3
RIG_HSP	-1.8	1.7	3.1	2.4	1.8	3.2
F1READ	2.5	0.2	0.3	-0.5	0.2	0.3
F1MATH	0.3	0.2	0.4	4.7	0.2	0.4
MTHCRD	-1.1	1.3	2.3	8.7	1.3	2.4
MTHGRD	3.2	1.4	2.4	14.4	1.4	2.5
F1ESTEEM	4.7	1.6	2.8	-2.1	1.6	2.9
F1LOCUS	-6.0	1.8	3.2	-1.9	1.9	3.4
HOMEWORK	3.8	1.8	3.1	1.8	1.9	3.3

* p-value for two-sided t-test < .05 across standard error range

The statistically significant interactions at the lower SE bound can be summarized as follows:

- When *SAT-V* is the outcome variable, there is a positive interaction effect estimate between coaching and socioeconomic status. There are negative interaction effect estimates between coaching and being Hispanic, and between coaching and having taken an AP course in high school.
- When *SAT-M* is the outcome variable, there are positive interaction effect estimates between coaching and living in a urban location, and between coaching and scoring higher on the NELS standardized test in math. There is a negative interaction effect estimate between coaching and scoring higher on the PSAT-M.

Table 5-3 indicates the size of these estimated interaction effects. I first consider the interpretation of estimated interaction effects on the SAT-V. Imagine that two students are coached and have identical values on all covariates in **X**, but one has an SES index value of 1 (in the top quartile of all students in the F1-F2 panel sample), and the other student has an SES index value of 0 (somewhere in the 2nd or 3rd quartile of all students in the F1-F2 panel sample). The linear regression model indicates that the estimated effect of coaching for the student with the high SES (25 points) is about 13 points higher on the SAT-V than that of the student with the low SES (12 points). One explanation for this might be that high SES students are more likely to enroll in costlier, more intensive coaching programs that are better suited to improve verbal reasoning skills. Nonetheless, the difference in estimated coaching effects for high and low SES students is still fairly small—roughly the equivalent of answering one more SAT-V item correctly. The linear regression model also indicates that ceteris paribus, the estimated effect of coaching for Hispanic students (-18 points) is about 31 points less than that of

coaching for white students (13 points), and the estimated effect for a coached student who has taken an AP course in high school (0) is 13 points less than the estimated effect for a coached student who has not taken an AP course (13 points).

When *SAT-M* is the outcome variable, the estimated interaction effect between *COACH* and *SCH_URB* indicates that the coaching effect for students in urban settings is about 11 points higher than the coaching effect for students in suburban settings, and about 15 points higher than the effect for students in rural settings. This might be interpreted as an indication that students from schools in urban locations have access to more effective coaching programs than students in other high school settings, though again, the difference in estimated coaching effects is fairly small.

The estimated interaction effects between *COACH* and the covariates *F1MATH* and *PSAT-M* require some care in their interpretation. The two interaction terms are strongly collinear with *COACH*, so including them in the model blows up the associated standard errors of parameter estimates. To circumvent this problem, I introduce residualized versions of the variables *COACH x F1MATH* and *COACH x PSAT-M* by regressing each variable on *COACH* and saving the predicted value. For each variable I then subtract the actual value from the predicted value. The two new variables are *COACH x F1MATHr*, which ranges from -24 to 14, and *COACH x PSAT-Mr*, which ranges from -255 to 355 . These variables are orthogonal to *COACH* by construction, and are interpretable as the amount scored by a student above or below the average for all coached students on *F1MATH* or *PSAT-M*. Positive values for these residualized

interaction variables indicate a student that scored better than expected, given that he or she was coached. Negative values indicate the opposite.

For example, imagine a coached student that has scored a 66 on *F1MATH*. This is eight points (1 standard deviation) more than the mean score for all coached students, so the variable *COACH x F1MATHr* equals 8. What is the effect of coaching for this student compared to one with the same *F1MATH* score who is uncoached? According to the estimates of the linear regression model reported in Table 5-3, ceteris paribus, the effect of coaching is 26 points. This effect is 50 points higher than the effect that would be estimated for a coached student scoring at the mean of *F1MATH*. It appears that students who are good at math, as assessed by the NELS standardized test, are the ones who benefit the most from coaching on the SAT-M. The opposite relationship holds for coached students with respect to their prior performance on the PSAT-M. In this case, for a student scoring 1 standard deviation above the mean for coached students (*PSAT-M*= 582, *COACH x PSAT-Mr* = 108), the estimated effect of coaching relative to an uncoached student with the same PSAT-M score is 5 points. This is 31 points less than the estimated effect for a coached student scoring at the mean of the PSAT-M distribution for all coached students. Students who did well on the PSAT-M apparently do not get the same benefit from coaching as students who did more poorly. Perhaps this latter group of students are more apt to benefit from instruction that emphasizes test-taking strategies and general testwiseness.

5.3 The Coaching Effect in the POP2 Subsample

A good test of the constancy constraint specifically, and the linear regression model more generally, would be to estimate SAT coaching effects using the same variables on a different sample of students. Can the effect estimates found for the POP1 subsample be replicated? While I do not have access to a parallel sample of students, the POP2 subsample described in chapter 3 may now be of some use. These are students who took the SAT but not the PSAT. There are 1,616 such students in the POP2 subsample, and 267 (16.5%) reported that they were coached prior to taking the SAT. A coaching effect cannot be estimated for this subsample with the same set of linear regression covariates specified in chapter 4 because no PSAT scores are available for these students by definition. A saving grace for the NELS data is that an excellent substitute for PSAT scores is available in the variables *F1READ* and *F1MATH*. The variable *F1READ* has a .69 correlation with PSAT-V scores, and a .75 correlation with SAT-V scores; the variable *F1MATH* has a .81 correlation with PSAT-M scores, and a .83 correlation with SAT-M scores. As a result, while dropping PSAT scores does reduce the predictive strength of the linear regression model, it does not seem to dramatically confound coaching effect estimates so long as NELS standardized test scores are included in the model.

The results of estimating the effects of coaching using the linear regression model with the POP1, POP2 and POP1 + POP2 combined samples are shown in Table 5-4. The linear regression covariates (**X**) specified for these different samples are the same as those used in chapter 4, with the exception that PSAT scores have been omitted. For the

POP1 subsample, this omission has reduced the SAT-V effect estimate from 11 to 9 points while increasing the corresponding standard errors. There is little impact on the SAT-M effect estimate, which decreases from 19 to 17 points without much change in the corresponding range of standard errors. Relative to the POP1 subsample, the estimated coaching effects are not dramatically different for the POP2 subsample: the SAT-V effect (3 points) is a little smaller, the SAT-M effect (20 points) is somewhat bigger. The standard errors around these effect estimates are larger, reflecting the fact that this subsample is about half the size of the POP1 subsample. Note that the POP1 and POP2 coaching effect estimates are roughly within one or two standard errors of each other. This suggests that we might do well to summarize the coaching effect estimates by applying the linear regression model to the combined subsamples. In this case, the estimated coaching effects are 7 points for the SAT-V, and 18 points for the SAT-M. The former estimate is only statistically significant at the lower standard error bound; the latter is significant at both the lower and upper standard error bound.

Table 5-4. Coaching Effects for POP1 and POP2 Subsamples

	SAT-V			SAT-M		
	POP1	POP2	POP1 + POP2	POP1	POP2	POP1 + POP2
COACH=1/Total	503/3144	228/1404	731/4548	503/3144	228/1404	731/4548
adjusted R^2	.61	.72	.65	.77	.76	.77
COACH effect Std Error Range	9.2 (3.4 to 6)	2.6 (5.1 to 9)	6.7 (2.7 to 4.8)	16.6 (2.9 to 5.1)	20.4 (5.1 to 9)	18.3 (2.5 to 4.4)

NOTES:
Covariates (**X**) are the same as those in chapter 4 with the exception of *PSAT-M* and *PSAT-V*.
COACH=1/Total before listwise deletion:
POP1 = (587/3504)
POP2 = (267/1616)
POP1 + POP2 = (854/5120)

5.4 Discussion

Is there a single coaching effect for the SAT-V and SAT-M? Certainly if previous research on the topic is any indication, the answer would seem to be no. Putting aside the issue of methodological design, coaching effect estimates have been shown to vary with respect to how coaching is defined and the characteristics of students who are coached. With respect to the NELS data, there is some evidence that coaching is more effective when it is defined more loosely as the combination of six possible test preparation activities. The estimated coaching effect is largest on the SAT-M for students who prepare by taking both a commercial course and hiring a private tutor. If coaching is defined more restrictively as taking a commercial preparatory course, the validity of the constancy constraint is less clear. I suggest that it is unlikely that the coaching effect for certain companies is dramatically different from the SAT-V and SAT-M confidence intervals of [3, 20] and [10, 28]. There is, however, evidence in the NELS data to bolster the notion that coaching is more effective for certain types of students. Students with high SES backgrounds benefit the most from coaching on the SAT-V. Hispanic students and students who have taken at least one AP course in high school benefit less from coaching on the SAT-V relative to white students and students who haven't taken an AP course respectively. Students in urban settings benefit most from coaching on the SAT-M. Students who are above average at math but do poorly on the PSAT-M benefit more from coaching than students who are below average at math but do fairly well on the PSAT-M. It should be emphasized that these interaction effects are

not unequivocal. They are only statistically significant under the assumption that there is no clustering of students within schools in the POP1 subsample, an assumption we know is wrong. When standard errors are estimated more conservatively (*DEFF* = 3), none of the estimated interaction effects are statistically significant. Lastly, the effect of coaching does not seem much different when estimated for the subsample of the NELS students who have not taken the PSAT prior to taking the SAT.

CHAPTER 6. SUMMARY AND CONCLUSION

6.1 Summary of Results

I have shown that coaching for the SAT differs in terms of instructional strategies, duration and setting. Previous studies of SAT coaching also differ with respect to their methodological designs. A small number of studies have no control groups, a large number are observational, and some are based on randomized experiments. While the total number of SAT coaching studies conducted since the 1950s is large, once studies are grouped by methodological design (uncontrolled, observational, randomized experimental) and coaching setting (school-based, commercial-based and computer-based), it becomes clear that the number of studies per condition is fairly sparse. A related problem is that coaching studies are seldom, if ever, replicated.

This is one reason why an unequivocal causal effect for coaching has proven difficult to establish. Another reason is that bias clouds the interpretation of estimated coaching effects. Randomized studies could, in theory, be used to estimate unbiased effects, but such studies in the context of SAT coaching have had limited success. Randomization with blind treatment groups is almost impossible to establish in school settings, and problems with student motivation and attrition lead to the same sorts of problems researchers encounter with effects estimated from observational designs: bias due to confounding. Confounding may be due to measured or unmeasured covariates; in

the specific case when confounding is due to an unmeasured latent covariate, it has been presented as selection bias.

Statistical approaches can be taken to estimate coaching effects while controlling for bias. Two approaches of particular interest are the linear regression model and the Heckman Model. The former has been applied most often to estimate coaching effects. The latter is a relatively new technique, and has only been applied once before in the context of an SAT coaching study. The Heckman Model merits close attention because as a model it purports to purge an estimated coaching effect of selection bias, something the linear regression model in its basic form cannot do. Either approach can be used to estimate the causal effect of coaching once an underlying behavioral model consisting of a selection function and a response schedule has been established. Without this underlying behavioral model, an estimated coaching effect will not have a clear causal interpretation—whether it is estimated using the linear regression approach or the Heckman Model approach.

As described in chapter 2, the critical difference between the linear regression and Heckman Model approaches lies in the assumption made about the relationship between the error term ε_i and latent covariate δ_i within student i. If ε_i and δ_i are assumed to be independent, the linear regression model is a feasible estimation approach: if the right covariates are included in the regression equation, the estimated coaching effect will be unbiased. If ε_i and δ_i are allowed to be correlated, the Heckman Model can be used to control for bias due to the endogeneity of the coaching variable.

If its assumptions are to be believed, the Heckman Model is a seemingly attractive solution to the problem of bias in estimated effects. It essentially turns the problem of confounding due to a latent covariate into that of confounding due to a measured covariate omitted from a regression equation. Putting aside the validity of distributional assumptions, much hinges upon the specification of the selection function. In chapter 4, I specified five different selection functions for coaching. All the specifications were reasonable choices, none was clearly superior in terms of model fit, yet they led to quite different causal inferences about the effect of coaching. One big reason for this was the collinear relationship between $COACH_i$ and $\lambda_i(COACH_i, \hat{\alpha}+\mathbf{Z}_i\hat{\boldsymbol{\gamma}})$.

Relative to an effect of coaching estimated as the difference in average SAT scores for coached and uncoached students, both the linear regression model and the Heckman Model appear to reduce bias. Including covariates in the linear regression model reduces the estimated SAT-V coaching effect by 9 points, and the estimated SAT-M coaching effect by 11 points. According to what seem to me the best of the Heckman Model specifications (SF4 and SF5), there is some indication that selection bias may cause the linear regression model to underestimate the SAT-M coaching effect by 20 to 25 points, but it seems to have no impact on the estimated SAT-V effect.

One threat to causal inferences common to both linear regression and the Heckman Model is the constancy constraint. In chapter 5, I considered this issue in the context of the linear regression model. When coaching was defined more generally as

any combination of six modes of test preparation, the marginal interpretation of the commercial coaching effect stayed about the same. The largest estimated effect for any single mode of test preparation was getting help from a private tutor on the SAT-M. There was some evidence that coaching is more effective for certain types of students. When SAT-M and SAT-V coaching effects were estimated for the NELS POP1 and POP2 student subsamples using the same linear regression model, the estimated effects were quite similar.

6.2 Conclusion

I return now to the fundamental questions I posed in the introduction to this dissertation:

1. How can the linear regression model and the Heckman Model be used to make unbiased causal inferences in observational settings?
2. Using these models, what can be concluded about the effect of coaching on SAT performance?

For the first question, it should be clear that both these models require very strong assumptions about the mechanism by which student data is generated. Establishing an underlying behavioral model to allow for causal inferences is akin to establishing a "social law of physics." This requires a theoretical understanding of causation that probably never exists in the social sciences. Because of its sensitivity to misspecification of the selection function, the Heckman Model is particularly worrisome as a tool for

making causal inferences. There may, however, be some hope for linear regression when the covariates included in the model are strongly predictive (i.e. correlation ≈ .85) of the outcome. Estimated causal effects will still almost certainly be biased, but the degree of bias is probably small. Given strongly predictive covariates, the constancy constraint is probably the biggest stumbling block to this use of linear regression for purposes of drawing causal inferences, but this is at least fairly easy to test. Holland (2001) has suggested that ideally, given certain assumptions, a treatment effect estimated using linear regression is best interpreted as an *average* causal effect.

A point worth emphasizing is that the best way to establish a causal effect from observational data, irrespective of the statistical model being used, is to replicate the results with a different sample. There was no single study or statistical model that established from observational data the deleterious effects of smoking on a range of health outcomes. Rather it was the consistent replication of these findings over a long period of time that led the way to what is now an accepted causal relationship.

Does coaching have an effect on SAT performance? It seems clear that an effect does exist, but the magnitude of this effect probably varies as a function of coaching and student characteristics. Based on the NELS data and the linear regression model, the best guess for the effect of commercial coaching is somewhere between 3 and 20 points on the SAT-V, and between 10 and 28 points for the SAT-M. Students with high socioeconomic backgrounds who have not taken an Advanced Placement course during high school appear to benefit most from coaching on the SAT-V. Students in urban

settings who are above average in math appear to benefit the most from coaching on the SAT-M. If coaching is defined more broadly as all possible combinations of six test preparation modes, the potential effect ranges from about -9 to 30 points on the SAT-V, and from about -7 to 44 points on the SAT-M. Regardless of how it is defined and modeled, coaching appears to be more effective for the SAT-M than it is for the SAT-V.

Table 6-1. Proportions of NELS Subsamples Engaging in Test Prep Activities

Test Preparation Activity	NELS F1-F2 Panel Subsamples			
	POP1	POP2	POP3	POP4
High School Course	23	18	15	15
Commercial Course	15	11	7	8
Private Tutor	7	10	7	6
Book	62	59	50	37
Video	6	6	7	7
Computer	17	14	13	10

POP1: Student took PSAT and SAT
POP2: Student took SAT, no PSAT
POP3: Student took PSAT, no SAT or ACT
POP4: Student took no tests

These estimates are still probably biased to some extent. Note that selection bias remains a potential problem even if the Heckman Model is considered the proper approach and is applied as in chapter 4. This is because there are actually two levels of selection taking place in the NELS data: students self-select how they will prepare for the SAT, but they also self-select whether they will actually take the SAT, regardless of how they choose to prepare. As is apparent in Table 6-1, many students in the NELS POP3 and POP4 samples who reported that they had prepared to take the SAT or ACT ultimately did not take either test before the end of their senior year in high school. If students in the POP3 and POP4 samples decided not to take an admissions test because of a negative coaching experience, leading them to believe they would perform poorly on

the SAT or ACT, then the omission of these students might serve to bias upwards the coaching effects estimated for students in the POP1 and POP2 subsamples. An extension of this study for a Heckman Model enthusiast would be to model the two different levels of selection before estimating a coaching effect.

6.3 Directions for Further Research

The SAT is in the process of changing its format by adding a writing section and eliminating analogy items from the verbal section. It would be advisable to evaluate the effectiveness of coaching for this new version of the test. A good place to start would be with a medium-scale (~500 students) randomized study where the coaching treatment is carefully defined and well understood. I have pointed out the logistical and ethical difficulties of conducting a coaching study with a randomized experimental design. Two strategies taken together might circumvent these difficulties.

First, a randomized study should work with junior students assigned to "coached" or control conditions taking an official administration of the SAT. A key aspect would be that the control condition would not be defined simply as the absence of coaching, but as systematic practice for the test without direct instruction. Few students do nothing to prepare for the SAT, and this should be taken into account up front in the study design. Students would not be told which condition was "coaching;" each would simply be considered alternate methods of test preparation. Both groups would subsequently be

given the opportunity to prepare with the other condition and re-take the SAT by the fall of their senior years.

Second, though the study would be designed as a randomized experiment, data would be collected as if the study had an observational design. This might include collecting data on a nonrandomly assigned control group. In the best case scenario, this might facilitate the kinds of comparisons between statistical approaches and the empirical "truth" made by Lalonde in his evaluation of the Heckman Model and other econometric approaches. At worst, if the randomized design did not hold, we would have another observational study to compare to previous estimates of coaching effects.

One potentially troubling finding from the NELS data is that there appear 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 (Briggs, 2001). 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. It would be worthwhile for a future study to survey students who decide not to take college admissions tests after having been coached in order to find out why this decision was made.

It has been suggested (Schwartz, 1999) that the benefits of coaching 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 could be true. It might be interesting to design a longitudinal coaching study that follows students beyond high school. Does coaching have a long-term effect on variables other than SAT performance? Perhaps most importantly, does coaching increase the likelihood of a student being accepted at his or her top choice for college? This study has not considered these other potential benefits of coaching. Further, the data used here is from the early 1990s, and may not reflect the state of the world more than ten years later. It is possible that there are specific programs and tutors capable of producing score gains substantially larger than one standard error of measurement on a section of the SAT. However, the empirical evidence for this is often anecdotal at best. With respect to the NELS data, for the average student there is little evidence that commercial coaching makes a substantial difference in SAT performance. Students and their parents should be careful before investing their resources into coaching programs with the expectation of dramatic improvements in SAT scores.

REFERENCES

Alderman, D. L. and Powers, D. E. (1980). The effects of special preparation on SAT-verbal scores. American Educational Research Journal 17: 239-253.

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. American Psychologist 36(10): 1086-1093.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: further synthesis and appraisal. Review of Educational Research 60(3): 373-417.

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

Berk, R. A. and Freedman, D. A. (2001). Statistical assumptions as empirical commitments. Unpublished manuscript: 24.

Bond, L. (1989). The effects of special preparation on measures of scholastic ability. Educational Measurement. R. L. Linn. New York, American Council on Education and Macmillan Publishing Company.

Breen, R. (1996). Regression Models: Censored, Sample Selected or Truncated Data. Thousand Oaks, SAGE Publications.

Briggs, D. C. (2001). The Effect of Admissions Test Preparation: Evidence from NELS:88. Chance 14(1): 10-18.

Briggs, D. C. (2002). Comment: Jack Kaplan's 'A new study of SAT coaching'. Chance 15(1): 7-8.

Burke, K. B. (1986). A model reading course and its effect on the verbal scores of eleventh and twelfth grade students on the Nelson Denny Test, the Preliminary Scholastic Aptitude Test, and the Scholastic Aptitude Test. Doctoral dissertation, Georgia State University.

Cochran, W. G. (1969). The use of covariance in observational studies. Applied Statistics 18: 270-275.

Coffin, G. C. (1987). Computer as a tool in SAT preparation. Paper presented at the Florida Instructional Computing Conference, Orlando, FL.

Coffman, W. E. and Parry, M. E. (1967). Effects of an accelerated reading course on SAT-V scores. Personnel and Guidance Journal 46: 292-296.

Cole, N. (1982). The implications of coaching for ability testing. Ability testing: Uses, consequences, and controversies. Part II: Documentation section. A. K. Wigdor and W. R. Gardner. Washington, DC, National Academy Press.

Curran, R. G. (1988). The effectiveness of computerized coaching for the Preliminary Scholastic Aptitude Test (PSAT/NMSQT) and the Scholastic Aptitude Test (SAT). Doctoral dissertation, Boston University.

Davis, W. D. (1985). An empirical assessment of selected computer software purported to raise SAT scores significantly when utilized with short-term computer-assisted instruction on the microcomputer. Doctoral dissertation, Florida State University.

Dear, R. E. (1958). The effect of intensive coaching on SAT scores. Princeton, NJ, Educational Testing Service.

DerSimonian, R. and Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: a meta-analysis. Harvard Educational Review 53: 1-15.

Dorans, N. J. (2002). Recentering and realigning the SAT score distributions: how and why. Journal of Educational Measurement 39(1): 59-84.

Dyer, H. S. (1953). Does Coaching Help? The College Board Review 19: 331-335.

Evans, F. and L. Pike (1973). The effects of instruction for three mathematics item formats. Journal of Educational Measurement 10(4): 257-272.

Federal Trade Commission, Boston Regional Office. (1978). Staff memorandum of the Boston Regional Office of the Federal Trade Commission: The effects of coaching on standardized admission examinations. Boston, Federal Trade Commission, Boston Regional Office.

Federal Trade Commission, Bureau of Consumer Protection. (1979). Effects of coaching standardized admission examinations: Revised statistical analyses of data gathered by the Boston Regional Office of the Federal Trade Commission. Washington, DC, Federal Trade Commission, Bureau of Consumer Protection.

Fraker, G. A. (1987). The Princeton Review reviewed. The Newsletter. Deerfield, MA: Deerfield Academy.

Frankel, E. (1960). Effects of growth, practice, and coaching on Scholastic Aptitude Test scores. Personnel and Guidance Journal 38: 713-719.

Freedman, D. (1987). As others see us: a case study in path analysis (with discussion). Journal of Educational Statistics 12(101-223).

Freedman, D. (1995). Some issues in the foundations of statistics (with discussion). Foundations of Science 1, 19-83.

Freedman, D. (1997). From association to causation via regression. In V. R. McKim & S. P. Turner (Eds.), Causality in crisis? : statistical methods and the search for causal knowledge in the social sciences (pp. 113-161). Notre Dame, Ind.: University of Notre Dame Press.

Freedman, D. (2002). On specifying graphical models for causation, and the identification problem (Technical Report 601). Berkeley: University of California, Berkeley, Department of Statistics.

French, J. W. (1955). The coachability of the SAT in public schools. Princeton, NJ, Educational Testing Service.

Frucot, V. and Cook, G. (1994). Further research on the accuracy of students' self-reported grade point averages, SAT scores, and course grades. Perceptual and Motor Skills **79**: 743-746.

Glass, G. V., McGaw, B. and Smith, W. L. (1981). Meta-analysis of social research. Beverly Hills, CA, Sage.

Goldberger, A. (1983). Abnormal selection bias. In S. Karlin, T. Amemiya & L. Goodman (Eds.), Studies in econometrics, time series and multivariate statistics. New York: Academic Press.

Greene, W. (1981). Sample selection bias as a specification error: comment. Econometrica 49, 795-798.

Greene, W. H. (1993). Econometric Analysis. New York, Macmillan Publishing Company.

Hansen, B. (2002). The promise of full matching for observational studies: evidence from a quasiexperiment assessing coaching for the SAT. Unpublished manuscript.

Harvey, K. S. (1988). Videotaped versus live instruction as a coaching method for the mathematics portion of the scholastic aptitude test. Doctoral dissertation, University of Georgia.

Heckman, J. (1979). Sample selection bias as a specification error. Econometrica 47: 153-161.

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equations system. Econometrica 46, 931-961.

Heckman, J. and Robb, R. (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), Drawing Inferences from Self-Selected Samples (pp. 63-107). Mahwah, NJ: Lawrence Erlbaum Associates.

Heckman, J. J. and Hotz, J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training (with comments). Journal of the American Statistical Association, 84, 862-880.

Hedges, L. V. and Olkin, I. (1985). Statistical methods for meta-analysis. New York, Academic Press, Inc.

Holland, P. W. (1986). Statistics and Causal Inference. Journal of the American Statistical Association 81(396), 945-960.

Holland, P.W. (1988). Causal inference, path analysis and recursive structural equations models. In C. Clogg (Ed.), Sociological Methodology (pp. 449-484).

Holland, P. W. (2001). The causal interpretation of regression coefficients. In M. C. Galavotti, P. Suppes, & D. Costantini (Eds.), Stochastic Causality (pp. 173-187): CSLI Publications.

Holmes, C. T. and Keffer, R. (1995). A computerized method to teach Latin and Greek root words: Effect on verbal SAT scores. Journal of Educational Research 89(1): 47-50.

Hopmeier, G. H. (1984). The effectiveness of computerized coaching for scholastic aptitude test in individual and group modes. Doctoral dissertation, Florida State University: 69.

Jackson, R. (1980). The Scholastic Aptitude Test: a response to Slack and Porter's 'critical appraisal'. Harvard Educational Review 50(3): 382-391.

Johnson, S. T. (1984). Preparing Black students for the SAT—Does it make a difference? (An evaluation report of the NAACP Test Preparation Project). New York, National Association for the Advancement for Colored People.

Jöreskog, K. and D. Sörbom (1996). LISREL 8: User's Reference Guide. Chicago, Scientific Software International.

Kaplan, J. (2001). A new study of SAT coaching. Chance 14(4): 1-6.

Keefauver, L. W. (1976). The effects of a program of coaching on Scholastic Aptitude Test scores of high school seniors tested as juniors. Doctoral dissertation, University of Tennessee at Knoxville.

Kintisch, L. S. (1979). Classroom techniques for improving Scholastic Aptitude Test scores. Journal of Reading 22: 416-419.

Kolata, G. (2001). Admissions test courses help, but not so much, study finds. The New York Times. New York.

Kulik, J. A., Bangert-Drowns, R. L. and Kulik, C. C. (1984). Effectiveness of coaching for aptitude tests. Psychological Bulletin 95: 179-188.

Lalonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review 76(4), 604-620.

Little, R. (1985). A note about models for selectivity bias. Econometrica 53(6), 1469-1474.

Little, R. J. and Rubin, D. B. (1987). Statistical analysis with missing data. New York, John Wiley.

Lawrence, I., Rigol, G., Van Essen, T. and Jackson, C. (2001). A historical perspective on the SAT. Rethinking the use of the SAT in University Admissions Conference, Santa Barbara, CA. November 17-18, 2001.

Lemann, N. (1999). The big test : the secret history of the American meritocracy. New York, Farrar Straus and Giroux.

Laschewer, A. (1986). The effect of computer assisted instruction as a coaching technique for the scholastic aptitude test preparation of high school juniors. Doctoral dissertation, Hofstra University.

Lass, A. H. (1961). Unpublished report. (Cited in Pike, 1978.)

Marron, J. E. (1965). Preparatory school test preparation: Special test preparation, its effect on College Board scores and the relationship of affected scores to subsequent college performance. West Point, N.Y., Research Division, Office of the Director of Admissions and Registrar, United States Military Academy.

McClain, B. (1999). The impact of computer-assisted coaching on the elevation of twelfth-grade students' SAT scores. Doctoral dissertation, Morgan State University.

Messick, S. (1980). The effectiveness of coaching for the SAT: review and reanalysis of research from the fifties to the FTC. Princeton, Educational Testing Service: 135.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: implications for educational and testing practice. Educational Psychologist 17(2): 67-91.

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

Pallone, N. J. (1961). Effects of short- and long-term developmental reading courses upon the S.A.T. verbal scores. Personnel and Guidance Journal 39(654-657).

Pike, L. W. (1978). Short-term instruction, testwiseness, and the Scholastic Aptitude Test: A literature review with research recommendations. Princeton, NJ, Educational Testing Service.

Powers, D. E. (1988). Preparing for the SAT: A survey of programs and resources. Princeton, NJ, Educational Testing Service.

Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. Educational Measurement: Issues and Practice(Summer): 24-39.

Powers, D. E. (1998). Preparing for the SAT I: Reasoning Test—an update. New York, The College Board.

Powers, D. E. and Rock, D. A. (1999). Effects of Coaching on SAT I: Reasoning Test Scores. Journal of Educational Measurement 36(2): 93-118.

Roberts, S. O. and Openheim, D. B. (1966). The effect of special instruction upon test performance of high school students in Tennessee. Princeton, NJ, Educational Testing Service.

Rock, D. (1980). Disentangling coaching effects and differential growth in the FTC commercial coaching study. Princeton, NJ, Educational Testing Service.

Rock, D. (2002). Personal communication. July 24, 2002.

Rosenbaum, P. R. (1995). Observational Studies. New York, Springer-Verlag.

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

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79, 516-524.

Rubin, D. B. (1986). Comment: which ifs have causal answers. Journal of the American Statistical Association 81(396), 961-962.

Schroeder, B. (1992). Problem-solving strategies and the mathematics SAT: A study of enhanced performance. Doctoral dissertation, Columbia University.

Sesnowitz, M., Bernhardt, K. and Knain, D. M. (1982). An analysis of the impact of commercial test preparation on SAT scores. American Educational Research Journal 19(3): 429-441.

Shaw, E. (1992). The effects of short-term coaching on the Scholastic Aptitude Test. Doctoral dissertation, University of La Verne.

Schwartz, T. (1999). The test under stress. The New York Times. New York**:** Section 6, Page 30, Column 1.

Slack, W. V. and Porter, D. (1980). The Scholastic Aptitude Test: A critical appraisal. Harvard Education Review 50: 154-175.

Smyth, F. L. (1989). Commercial coaching and SAT scores. The Journal of College Admissions 123(Spring): 2-9.

Smyth, F. L. (1990). SAT Coaching: what really happens and how we are led to expect more. The Journal of College Admissions 129(Fall): 7-17.

Snedecor, P. J. (1989). Coaching: does it pay—revisited. The Journal of College Admissions 125: 15-18.

Stroud, T. W. F. (1980). Reanalysis of the Federal Trade Commission study of commercial coaching for the SAT. Princeton, NJ, Educational Testing Service.

United States Department of Education (1995). NELS Second Follow-up Student Component Data File User's Manual. Washington, D.C., National Center for Educational Statistics. Available on the world wide web at http://www.nces.ed.gov/surveys/nels88/.

Warch, K. L. (1996). The impact of computer-assisted coaching on high school students' SAT I scores. Doctoral dissertation, California State University, Long Beach.

Whitla, D. K. (1962). Effect of tutoring on Scholastic Aptitude Test scores. Personnel and Guidance Journal 41: 32-37.

Whitla, D. K. (1988). Coaching: does it pay? Not for Harvard students. The College Board Review 148(Summer 1998): 32-35.

Wing, C. W. (1987). Some field observations of the impact of test preparatory programs on high school students' Scholastic Aptitude Test scores. A report to the Awards Committee for Education and Wake Forest University, Winston-Salem, NC.

Winokur, H. (1983). The effects of special preparation for the verbal section of the SAT. Doctoral dissertation, Virginia Polytechnic and State University.

Wrinkle, G. W. (1996). A Scholastic Assessment Test preparation class and its effect on Scholastic Assessment Test scores. Doctoral dissertation, University of Houston.

Statacorp. (2001). Stata reference manual set, vol. 4. Stata Press. (www.stata.com).

Vella, F. (1998). Estimating models with sample selection bias: a survey. The Journal of Human Resources 33(1), 127-169.

Wainer, H. ed. (1986). Drawing Inferences from Self-Selected Samples. Mahwah, NJ: Lawrence Erlbaum Associates.

Zuman, J. P. (1988). The effectiveness of special preparation for the SAT: An evaluation of a commercial coaching school. Doctoral dissertation, Harvard University.

Zwick, R. (2002). Fair game? the use of standardized admissions tests in higher education. New York: RoutledgeFalmer.

APPENDIX

Variable Name	Variable Type	Description	NELS Variable Source	Correlations (POP1)	
				SAT-V	SAT-M
AGE	Continuous	Student's age in years relative to month of 1992 NELS F2 survey	F2S5AMO, F2BIRTHM, F2BIRTHY	-.052	-.022
AP	Dummy	= 1 if student reported taking an AP class during high school	F2S13E	.379*	.380*
ASIAN	Dummy	= 1 if student was of Asian or Pacific Island descent	F2RACE1	.027	.101*
ASP	Dummy	=1 if both parents and student indicate high level of college aspirations.	F2S43, F2S42A, F2S42B	.178**	.21*
BLACK	Dummy	=1 if student is African-American	F2RACE1	-.184*	-.246*
COACH	Dummy	=1 if student took a course offered by a commercial test preparation service	F2S45B	.07*	.09*
COLREP	Dummy	=1 if student indicated that reputation of college he/she hoped to attend was "very" important	F2S59L	.154*	.144*
DADASP	Dummy	=1 student reports that father expects him/her to complete 4 years of college or beyond	F2S42A	.181*	.154*
DESWGT	Continuous	Design Effect correction applied for all significance tests, = (1/DEFF)*(F2TRP2WT/meanF2TRP2WT)	F2TRP2WT	NA	NA
EASYADMT	Dummy	=1 if student indicated that easy admissions standards at college he/she hoped to attend were "very" important	F2S59M	-.258*	-.257*
F1ESTEEM	Continuous	NELS composite variable computed from responses to 6 Likert items in F1 survey. Higher values indicate higher levels of self-esteem.	F1CNCPT2-	.095*	.095*
F1LOCUS	Continuous	NELS composite variable computed from responses to 7 Likert items in F1 survey. Higher values indicate that student feels more "in control" of his or her life.	F1LOCUS2	.177*	.189*
F1MATH	Continuous	NELS standardized math test administered to students in F1 survey (10^{th} grade).	F12XMSTD	.657*	.837**
F1M_TOP	Dummy	=1 if student scored in top quartile of F1MATH relative to full NELS F1 sample taking test	F12XMQ	.54**	.685**
F1READ	Continuous	NELS standardized reading test administered to students in F1 survey (10^{th} grade).	F12XRSTD	.72*	.568**
F1R_TOP	Dummy	=1 if student scored in top quartile of F1READ relative to full NELS F1 sample taking test	F12XRQ	.597**	.447**

APPENDIX

Variable Name	Variable Type	Description	NELS Variable Source	Correlations (POP1)	
				SAT-V	SAT-M
F2MATH	Continuous	NELS standardized math test administered to students in F2 survey (12th grade).	F22XMSTD	.669*	.873*
F2M_TOP	Dummy	=1 if student scored in top quartile of F2MATH relative to full NELS F2 sample taking test	F22XMQ	.482*	.644*
F2READ	Continuous	NELS standardized reading test administered to students in F2 survey (12th grade).	F22XRSTD	.71*	.57*
F2R_TOP	Dummy	=1 if student scored in top quartile of F2READ relative to full NELS F2 sample taking test	F22XRQ	.559*	.445*
F2TRP2WT	Continuous	NELS population weight applied to F1 – F2 panel sample for whom transcript information is available	F2TRP2WT	NA	NA
FEMALE	Dummy	=1 if student is female	F2SEX	-.057	-.194*
HISPANIC	Dummy	=1 if student is Hispanic	F2RACE1	-.161*	-.146**
HOMEWORK	Dummy	=1 if student claims to have averaged more than 10 hours per week on homework (outside of school) during high school.	F2S25F2	.161*	.14*
HWTUTOR	Dummy	=1 if student had a private tutor to help with homework assignments during high school	F2S26B	-.066*	-.133*
MIDWEST	Dummy	=1 if student's high school is located in the midwest	TRNREGON	.051	.061*
MINORITY	Dummy	=1 if student is either African-American or Hispanic	F2RACE1	-.263*	-.302*
MOMASP	Dummy	=1 student reports that mother expects him/her to complete 4 years of college or beyond	F2S42B	.182*	.183*
MTHCRD	Continuous	Number of units of college-preparatory math courses student took while in high school	MTHPIP8	.120*	.236*
MTHGRD	Continuous	Student's weighted grade point average in college-preparatory math courses taken while in high school	MTHGRD	.427*	.595*
NORTHEA	Dummy	=1 if student's high school is located in the northeast	TRNREGON	.039	-.025
NUMAPPS	Continuous	The number of post-secondary institutions to which student has applied	F2S60A	.288*	.318*
PARENT	Dummy	=1 if student reports that parent encouraged him/her to prepare for the SAT	F2P62A	.074*	.103*
PREPBOOK	Dummy	=1 if student prepared for the SAT by studying with a book	F2S45B	-.013	-.012
PREPHS	Dummy	=1 if student prepared for the SAT by taking a special course offered by his/her high school	F2S45B	-.035	.000
PREPPC	Dummy	=1 if student prepared for the SAT by studying with a computer program	F2S45B	-.008	.047

APPENDIX

Variable Name	Variable Type	Description	NELS Variable Source	Correlations (POP1)	
				SAT-V	SAT-M
PREPTUT	Dummy	=1 if student prepared for the SAT by receiving one-to-one tutoring	F2S45B	.000	.009
PREPVID	Dummy	=1 if student prepared for the SAT by using a video tape	F2S45B	-.064	-.065*
PPRESS	Dummy	=1 if student "often" discussed plans and preparation for the SAT with parent	F2S99E	.01	-.015
PRIVATE	Dummy	=1 if student attended a private high school	G12CTRL2	.07*	.004
PSATM	Continuous	Student's score on the math section of the PSAT	F2RPSATV[1]	.675*	.868*
PSATV	Continuous	Student's score on the verbal section of the PSAT	F2RPSATM[1]	.854*	.637*
RE_ENG	Dummy	=1 if student reported participation in remedial English class during high school	F2S13A	.185*	.13*
REMATH	Dummy	=1 if student reported participation in remedial Math class during high school	F2S13B	.232*	.295*
RIGPROG	Dummy	=1 if student was enrolled in a rigorous academic curriculum during high school	F2RTRPR6	.079*	.128**
SAMESCH	Dummy	=1 if student was in same school during F1 and F2 surveys	F2F1SCFL	.077*	.077*
SATM	Continuous	Student's score on the math section of the SAT	F2RPSATV[1]	NA	NA
SATV	Continuous	Student's score on the verbal section of the SAT	F2RPSATM[1]	NA	NA
SCH_URB	Dummy	=1 if school was in a urban location	G12URBN3	-.021	-.049
SCH_RUR	Dummy	=1 if school was in a rural location	G12URBN3	-.060*	-.042
SCH_SUB	Dummy	=1 if school was in a suburban location	G12URBN3	.066*	.079*
SES	Continuous	Socioeconomic status composite variable. Combines information on parent education, income and occupation into single index.	F2SES1	.353*	.374*
SES_TOP	Dummy	=1 if student is in top quartile of SES composite variable relative to all students in NELS F2 sample.	F2SESQ	.28*	.304*
SOUTH	Dummy	=1 if student's high school is located in the south	TRNREGON	-.066*	-.058
STUDASP	Dummy	=1 student reports that he/she expects to complete 4 years of college or beyond	F2S43	.189*	.23*
WEST	Dummy	=1 if student's high school is located in the west	TRNREGON	-.011	.051

All correlations calculated with design effect adjustment, DEFF = 3

[1] The math and verbal SAT and PSAT scores were mistakenly inverted in the source NELS:88 data.