Causal Inference and the Heckman Model

Derek C. Briggs

University of Colorado, Boulder

derek.briggs@colorado.edu

Pre-print of paper published in the *Journal of Educational and Behavioral Statistics*, Winter 2004, Vol 29(4), 397-420.

Author's note:

Thanks to David Freedman for his helpful comments on earlier versions of this paper.

Abstract

In the social sciences, evaluating the effectiveness of a program or intervention often leads researchers to draw causal inferences from observational research designs. Bias in estimated causal effects becomes an obvious problem in such settings. I present the Heckman Model as an approach sometimes applied to observational data for the purpose of estimating an unbiased causal effect. I show how the Heckman Model can be used to correct for the problem of selection bias, and discuss in some detail the assumptions necessary before the approach can be used to make causal inferences. The Heckman Model makes assumptions about the relationship between two equations in an underlying behavioral model: a response schedule and a selection function. I show that the Heckman Model is particularly sensitive to the choice of variables included in the selection function. This is demonstrated empirically in the context of estimating the effect of commercial coaching programs on the SAT performance of high school students. Coaching effects for both sections of the SAT are estimated using data from the National Education Longitudinal Study of 1988 (NELS). Small changes in the selection function are shown to have a big impact on estimated coaching effects under the Heckman Model.

Introduction

A number of statistical methods may be used in observational settings to control for bias in the estimation of treatment effects. There is a common thread running through such approaches: 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 paper the focus is on a method of controlling for bias known as the Heckman Model (Heckman, 1978; 1979; Heckman & Robb, 1986; Greene, 1993)[1]. While the Heckman Model is a well-established approach among econometricians, its use is less common among educational statisticians. Much of what follows will serve as a didactic introduction to the Heckman Model for the benefit of this latter audience, but more generally, this paper presents the assumptions that would be necessary before the Heckman Model could be used to draw causal inferences in an observational setting.

To give this presentation an applied context, I use the Heckman Model to evaluate the effectiveness of coaching programs in improving performance on the SAT. The SAT is required for admission at almost all competitive four-year colleges in the United States, and has a math and verbal section, each scored on a scale that ranges from 200 to 800

[1] 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).

with standard deviation of about 110 points[2]. Each year about two million high school students take the test 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 have attributed to their programs average gains of 100-140 points on the combined math and verbal sections of the test (Schwartz, 1999). 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 (cf. Messick, 1980; Messick & Jungeblut, 1981; Becker, 1990; Powers, 1993, Powers & Rock, 1999, Briggs, 2001).

One problem for research on SAT coaching has been that coaching effect estimates are usually based on studies with observational designs, making causal inference about coaching effects equivocal. There is typically reason to believe that coached students self-select themselves on the basis of higher levels of motivation or academic ability. To the extent that such variables are themselves correlated with SAT performance, an estimated coaching effect will suffer from selection bias. When certain assumptions hold, the Heckman Model is a statistical approach that could be used in such a scenario to estimate an asymptotically unbiased effect of coaching. I discuss these assumptions and the form of the Heckman Model correction in the next section.

[2] 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.

The Heckman Model

Consider the following behavioral model for a student taking the SAT:

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

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

The model consists of a *response schedule* (1) and a *selection function* (2). In the response schedule, a student's potentially observable SAT score is a function of the variable *COACH*. 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 specified here assumes a linear relationship between the variable *COACH* and the potentially observable SAT score, with a constant effect across individuals, represented by the parameter *b*. Likewise, the effect of $\mathbf{X}_i$ is linear, and **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, which specifies that a student's decision to be coached is a function of the vector of observable covariates, $\mathbf{Z}_i$ and the latent covariate, δ_i.

According to the model, observed SAT scores are generated as

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

where $COACH_i$ is determined by Equation 2. Under the Heckman Model, three further statistical assumptions must be made:

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 $\{\varepsilon_i: i = 1, \ldots, N\}$

iii) $\{\mathbf{Z}_i: i = 1, \ldots, N\}$ is independent of $\{\delta_i: i = 1, \ldots, N\}$

No restrictions are imposed on the relationship between ε_i and δ_i, so the variable $COACH_i$ may be correlated with the error term ε_i. This relationship is captured by the parameter, ρ which may take on any value between -1 and 1. If $\rho \neq 0$, the variable $COACH_i$ will be endogenous, and the causal parameter b will suffer from selection bias[3].

Note that if ε_i and δ_i are not correlated, then $\rho = 0$, and there would be no selection bias problem. Linear regression could be used to 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 use of linear regression to draw causal inferences would require the assumption 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.

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

Given Equations 1-2 and assumptions i-iii, 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 of the relationship between coaching and SAT performance. 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, then selection bias has been purged from 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 (2). 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 regressing Y_i on a constant, $COACH_i$, $\mathbf{X}_i$ and $\hat{\lambda}$ will almost control 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 $\mathbf{Z}_i$ contains observable covariates related to the probability that a student is coached. 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 $\mathbf{Z}_i$ and δ_i. The selection function specifies that if $\alpha + \mathbf{Z}_i\boldsymbol{\gamma} + \delta_i > 0$, 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

$$\begin{aligned} P(COACH_i = 1 \mid \mathbf{Z}_i) &= P(\alpha + \mathbf{Z}_i\boldsymbol{\gamma} + \delta_i > 0 \mid \mathbf{Z}_i) \\ &= P(-\delta_i < \alpha + \mathbf{Z}_i\boldsymbol{\gamma} \mid \mathbf{Z}_i) \\ &= \Phi(\alpha + \mathbf{Z}_i\boldsymbol{\gamma}), \end{aligned} \tag{4}$$

where Φ represents the standard normal cumulative distribution function. Given all the $\mathbf{Z}_i$'s, the $COACH_i$'s are assumed to be independent, so Equation 4 constitutes what is known as the probit model.

The Heckman Model goes from specifying a selection function to getting an estimate for the bias term, $E(\varepsilon_i \mid \mathbf{X}_i, COACH_i)$ by estimating the expected value of a truncated normal random variable[4]. This estimate is known in the literature as the Mills Ratio or Hazzard Function, and can be expressed is the ratio of the standard normal density function, $\phi(t)$, to $\Phi(t)$:

$$\lambda(t) = \frac{\phi(t)}{1 - \Phi(t)} \tag{5}$$

where *t* is the point at which the distribution has been truncated. When the truncation is from above, then by symmetry of the normal distribution, the expected value of the random variable will be

[4] For details, see Johnson & Kotz, 1970, 112-113 and Greene, 1990, 682-689.

$$\lambda(t) = -\frac{\phi(t)}{\Phi(t)}. \tag{6}$$

The 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 the underlying behavioral model (Equations 1 and 2), and the condition that $COACH_i = 1$, it follows that δ_i no longer has a normal distribution, but a truncated normal distribution. The conditional expectation of δ_i will be $E(\delta_i \mid \alpha + \mathbf{Z}_i\boldsymbol{\gamma} + \delta_i > 0)$. 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 $E(\delta_i \mid \alpha + \mathbf{Z}_i\boldsymbol{\gamma} + \delta_i \le 0)$. 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. Think of ξ_i as the random variable that picks up the variance left unexplained if ε_i is regressed on δ_i. Now ε_i can be related to δ_i and ξ_i:

$$\varepsilon_i = \rho\delta_i + \sqrt{1-\rho^2}\xi_i. \tag{7}$$

Let $s_i = \alpha + \mathbf{Z}_i\boldsymbol{\gamma}$. It follows that

$$\begin{aligned} E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, COACH_i = 1) &= E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, s_i + \delta_i > 0) \\ &= \rho E(\delta_i \mid s_i + \delta_i > 0) \\ &= \rho E(\delta_i \mid \delta_i > -s_i). \end{aligned} \tag{8}$$

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 (8). Taking advantage of the symmetry of the normal distribution leads to the Inverse Mills Ratio,

$$E(\delta_i \mid \delta_i > -s_i) = \frac{\phi(s_i)}{1-\Phi(s_i)}. \tag{9}$$

Likewise,

$$\begin{aligned} E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, COACH_i = 0) &= E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, s_i + \delta_i \le 0) \\ &= \rho E(\delta_i \mid s_i + \delta_i \le 0) \\ &= \rho E(\delta_i \mid \delta_i \le -s_i). \end{aligned} \tag{10}$$

This again yields the Inverse Mills Ratio

$$E(\delta_i \mid \delta_i \le -s_i) = -\frac{\phi(s_i)}{\Phi(s_i)}. \tag{11}$$

It follows from (7-11) that

$$E(\varepsilon_i \mid \mathbf{X}_i, COACH_i) = \rho\lambda_i(COACH_i, s_i), \tag{12}$$

where

$$\lambda_i(COACH_i, s_i) = COACH_i\left(\frac{\phi(s_i)}{1-\Phi(s_i)}\right) + (1-COACH_i)\frac{-\phi(s_i)}{\Phi(s_i)}. \tag{13}$$

$\lambda_i(COACH_i, s_i)$ 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. The variable $\lambda_i(COACH_i, \hat{s}_i)$ can be computed after estimating parameter values for α and $\boldsymbol{\gamma}$ in the probit model of (4) via maximum likelihood.

The behavioral model of (1) and (2) now leads to

$$Y_i = a + bCOACH_i + \mathbf{X}_i\mathbf{c} + h\lambda_i(COACH_i, \hat{s}_i) + \varepsilon_i^* \quad (14)$$

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 14 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 an estimate for $\hat{\rho}$ is desired, it can be obtained 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 $\mathbf{Z}_i$, 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) will 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.

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 by regressing Y_i on a constant, $COACH_i$, $\mathbf{X}_i$ without the Heckman Model correction 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 estimated without the Heckman Model correction would be biased upwards. If $\hat{\rho}$ is negative, it suggests that the coaching effect estimate without the Heckman Model correction would be biased downwards.

To summarize, the Heckman Model as applied to coaching studies has two main steps[5].

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.

The Heckman Model in Practice

As presented here, the Heckman Model assumes 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

[5] 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).

with Y_i. If the linearity assumption is incorrect, a coaching effect will be estimated as the difference between the wrong two regression surfaces. A constancy constraint, i.e. $b_i = b$, is also typically stipulated, such that person $i = 1, \ldots, N$ 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.

Because of the strong assumptions that underlie the Heckman Model, its usefulness has been questioned by some statisticians (Wainer, 1986; Little, 1985) and econometricians (Goldberger, 1983). In one unusual case (Lalonde, 1986), the causal 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 group, 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 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 subsequent 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 $\{\mathbf{Z}_i: i = 1, \ldots, n\}$ 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. Statacorp, 2001). In either case, it is typically assumed that the additional variables included in the selection function are strong predictors of treatment assignment, yet uncorrelated with the outcome of interest. These are known as instrumental variables in the econometric literature.

In the context of SAT coaching, an ideal instance of an instrumental variable would occur if participation in coaching programs was assigned to interested students using a lottery system. One could imagine some sample of students in which each student was given a randomly generated number. Subsequently, a researcher would decide, based on some cutoff value, whether each student would be assigned to a coaching program or not. The value of the random number generator would be perfectly correlated with coaching assignment, but presumably uncorrelated with SAT performance. Specifying a selection function for use in a Heckman Model would seem quite credible in this instance.

As in the case above, the ideal choice of covariates to include in the selection function should be based on some theoretical understanding of the selection mechanism. Of course, observational studies seldom present the researcher with this sort of scenario.

The NELS Data

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. A panel of nearly 15,000 students completed survey questionnaires in the second two waves of NELS in 1990 and 1992. 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 included information on student grades, course taking patterns, school demographics, and college admission test scores.

For the analysis that follows, attention is focused on the NELS panel sample of students who completed surveys in the first (F1) and second (F2) follow-ups, and for whom transcript data was collected. This comprises an F1-F2 panel of 14,617 students. The emphasis in most SAT coaching studies has been on students 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. I similarly restrict attention to the 3,504 students from the NELS subsample 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.

The NELS Variables

To estimate a coaching effect from the NELS data using the Heckman Model requires three types of variables: an outcome variable (*Y*), a coaching variable (*COACH*), and covariates to be included in **X** and **Z**. I briefly describe each in turn.

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 total of two and a half hours. The test was then, and is now, intended to measure the constructs of mathematical and verbal reasoning, with scores from two different test sections. Each score was based on student responses to about 85 verbal items and 60 math items on the SAT. Because this is a relatively large number of items, and the items are chosen with great care, the SAT has the desirable technical feature of high internal consistency. 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. The mean and standard deviation of SAT-V scores (446 and 102) for the NELS subsample 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

[8] That mean SAT-V scores are higher than mean SAT-M scores is an artifact of the original samples used to create the original SAT score scales. The SAT 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.

was about 423 on the SAT-V, and 475 on the SAT-M. The mean SAT scores for the NELS 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.

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[6] 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. About 15% of the students in the NELS subsample indicated that they had taken a commercial course to prepare for the SAT.

Covariates

Insert Table 1 about here

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. A review of the research literature on SAT coaching (see Briggs, 2002) indicates that previous SAT or PSAT scores, demographic characteristics, academic background and student motivation may serve to confound coaching effect estimates. These variables, and their relationship to coaching status, are shown in Table 1. Student motivation can be further divided into variables that proxy for intrinsic motivation (e.g. self-esteem) and extrinsic motivation (e.g. parental pressure). The latter variables may predict whether students are likely to be

[6] Other ways of defining the coaching treatment with respect to the NELS prompt on test preparation activities are certainly possible. Many of these are described and analyzed in Briggs (2004).

coached, but are unlikely to have a direct influence on how students will perform on the SAT. Variables measuring extrinsic motivation, shown separately in Table 2, might be particularly attractive candidates to include within the matrix **Z** for a coaching selection function.

Insert Table 2 about here

When coached and uncoached students are compared along these sets of covariates in the NELS data, it appears that the coached group is more socioeconomically advantaged and more extrinsically motivated to take the SAT then uncoached counterparts. It does not appear 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 both report that they do about the same amount of homework per week.

Coaching Effects and the Heckman Model

I start by specifying all covariates with a theoretical relationship to coaching status and SAT performance in the underlying behavioral model described by equations 1 and 2. There are a total of 21 covariates in the vector $\boldsymbol{X_i}$.

- Pre-coaching SAT scores: (*PSAT-V* and *PSAT-M*).

- Demographic characteristics[7]: *AGE*, *SES*, *FEMALE*, *ASIAN*, *BLACK*, *HISPANIC*, *AM_INDIAN*, *PRIVATE*, *SCH_URB*, and *SCH_RUR*.
- Academic background[8]: *AP* , *RE_MATH*, *RE_ENG*; *RIGHSP*, *F1MATH*, *F1READ*, *MTHCRD*, and *MTHGRD*.
- Intrinsic student motivation: *F1ESTEEM*, *F1LOCUS* and *HOMEWORK*.

Note that in specifying the set of covariates to include in $\boldsymbol{X_i}$, a commitment is made to the sort of causal relationship shown in Equations 1 and 2. That is, I assume there is no confounding of the relationship between SAT performance and coaching status beyond that captured by the covariates in $\boldsymbol{X_i}$. The only other theoretical source of bias in the estimate of the coaching effect comes from the correlation of $COACH_i$ with ε_i. The Heckman Model is employed to control for this source of bias

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

[7] The reference categories are *WHITE* and *SCH_SUB* for the racial/ethnic and school location dummy variables respectively. 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 the NELS subsample considered here, the SES index has a mean of .44, a standard deviation of .73, and a range from –2.4 to 2.5.

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

i will be coached or not. The specification decision hinges upon what covariates are included in $\mathbf{Z}_i$. Ideally, students in the NELS survey would have been assigned to coaching programs by some known process, or at least asked questions about why they did or did not enroll in coaching programs, but as NELS was not designed to answer this sort of question, such data is not available. In many empirical applications of the Heckman Model, the decision of what covariates to include in $\mathbf{Z}_i$ appears to be largely a matter of ensuring that the model is well identified.

Figure 1. Five Selection Function Specification

SF1	$\boldsymbol{Z_i} = \{\boldsymbol{X_i}\}$
SF2	$\boldsymbol{Z_i} = \{\boldsymbol{X_i}, PARENT_i\}$
SF3	$\boldsymbol{Z_i} = \{PARENT_i, PPRESS_i, HWTUTOR_i, HI\_MOT_i\}$
SF4	$\boldsymbol{Z_i} = \{SES_i, SCH\_RUR_i, REMATH_i, MTHCRD_i, PPRESS_i, HWTUTOR_i, HI\_MOT_i\}$
SF5	$\boldsymbol{Z_i} = \{\text{AGEi}, SES_i, SCH\_RUR_i, MTHGRD_i, PARENT_i, PPRESS_i, HWTUTOR_i, HI\_MOT_i\}$

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 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 to be expected 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*, 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 response schedule (*PPRESS*, *HWTUTOR* and *HI_MOT*) are included as predictors in the selection function[9]. Under SF3, 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[10] 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 an approximation of a theory-based specification approach. Here the data analyst has taken

[9] Values for the predictors *PARENT*, *PPRESS* and *HWTUTOR* were missing for anywhere from 2 to 10% of the NELS subsample of 3,144 students used in the linear regression model. 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.

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

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 3 presents the parameter estimates generated from a probit model for each of the five SF 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. 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 3. 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.

Insert Table 3 about here

Another possible criterion to 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. SF1 and SF2 are particularly weak relative to this criterion, with only 13% and 20% of estimated coefficients statistically significant.

For each of the five SF specifications, one can examine the predicted probabilities of being coached as a function of selection function covariates. For SF4 and SF5 the highest estimated probability is about .2 higher than that estimated under SF1, SF2 and SF3. In terms of the actual and predicted number of coached students for each specification, 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.

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.

Heckman Model Estimates

The Inverse Mills Ratio, $\lambda_{ik}(COACH_i, \hat{s}_{ik})$, 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 $\mathbf{X}_i$. Each regression is weighted by the variable *DESWGT* to account for the NELS population weights, as well as the design effects caused by the stratification and clustering of students in the NELS sample. The clustering of students in the NELS subsample amounts to a mean of 4 and a median of 6 students per school—relative to a mean and median of 14 for the full F1-F2 panel sample. In the NELS subsample there is on average just one coached student per sampled school. Given this, the design effect correction of 3 used here will probably overestimate standard errors, and should be viewed as a conservative upper bound. Finally, 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. Table 4 reports the results of these regressions for SAT-V and SAT-M test scores.

Insert Table 4 about here

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.

Depending upon the selection function that is specified, the Heckman Model tells 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 are to be believed, it would indicate that any coaching effects estimated by the linear regression model will 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 variables $\lambda_{ik}(COACH_i, \hat{s}_{ik})$ 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.

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 the underlying behavioral model has now been violated—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 paper, 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.

Comparing Effect Estimates

Figures 2 and 3 compare the SAT-V and SAT-M coaching effects estimated by 1) taking the unadjusted difference in average scores between coached and uncoached students, 2) using the five Heckman Model specifications and 3) using just linear regression with the covariates $\mathbf{X}_i$ (i.e., assuming that δ_i and ε_i are independent). I include around each point estimate the corresponding 95% confidence interval. All parameters are estimated with the same design effect correction.

Insert Figures 2 and 3 about here

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 tend to be 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.

Unlike the Lalonde study, there is no absolute criterion against which to compare the coaching effects estimated by the Heckman Model. 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.[11] 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. Yet Powers & Rock's SAT-V coaching effect estimate (12 points) produced using the Heckman Model was similar only to those produced under SF4 and SF5 with the NELS data; for the SAT-M their effect estimate (13 points) 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.

Discussion

The question of when causal inferences can be drawn in observational studies is a subject of perpetual debate. Much attention has been directed at the suitability of the

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

linear regression model for making claims of cause and effect (c.f. Berk, 2003; Freedman, 1995, 2002; Holland, 2001). With respect to the behavioral model presented here, a key difference between linear regression and the Heckman Model is the relaxation of the independence assumption between the error terms in Equations 1 and 2. Another key assumption of the Heckman Model is bivariate normality for those error terms. If normality does not hold, then the Heckman Model as described here falls apart as a correction for the selection bias problem. Note that normality is a necessary condition for consistent estimation under the Heckman Model, but not for linear regression. If the linear regression error term is iid, if the error terms in Equations 1 and 2 are independent, and if confounding covariates are included in the model, then linear regression will produce unbiased causal effect estimates even when the distribution of the error term (ε_i) is non-normal.

As a correction for the problem of selection bias, the Heckman Model is an intuitively appealing tool for estimating an unbiased effect of commercial coaching on SAT performance. That being said, I have shown that using a selection function specified just with the objective of identifying the model (e.g. SF1 and SF2), results in effect estimates from the Heckman Model that are substantially different than those estimated from a selection function specified on a slightly more theoretical basis (e.g. SF5). I have also shown that once a selection function has been specified, estimated, and used to calculate the Inverse Mills Ratio, a large degree of multicollinearity may serve to inflate standard error estimates. With access to the right software (e.g. STATA, LIMDEP), the Heckman Model is easily implemented with seemingly obvious causal conclusions. This

paper suggests successfully application of the Heckman Model in social science research must hinge upon a compelling theoretical rationale and a careful scrutiny of the data.

The results of the analysis of the NELS data here are consistent with the established notion that much caution must be exercised before applying the Heckman Model as a means of drawing causal inferences about a treatment effect. In the social sciences, bias in the estimated effects from any given study is very difficult to rule out, no matter how intuitively appealing the methodology. There is, unfortunately, no statistical silver bullet.

References

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

Berk, R. A. (2003) *Regression analysis: a constructive critique*. Thousand Oaks, CA: Sage.

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

Briggs, Derek C. (2004) Evaluating SAT coaching: gains, effects and self-selection. In: Rethinking the SAT: Perspectives Based on the November 2001 Conference at the University of California, Santa Barbara, R. Zwick, ed., RoutledgeFalmer.

Briggs, Derek C. (2002) SAT coaching, bias and causal inference. Ph.D. dissertation, University of California, Berkeley.

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

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.

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

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

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.

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

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

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.

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.

Johnson, N., and S. Kotz. (1971). Distributions in Statistics—Continuous Univariate Distributions, Vol. 2. New York: Wiley.

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

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: Wiley.

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

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

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.

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

Rosenbaum, P. R. (2002). Observational Studies. New York, Springer-Verlag, 2nd Edition.

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.

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

Table 1. Covariates by Coaching Status

Name	Description of Covariate	Uncoached	Coached	Stat Sig (p value)
PSAT-V	PSAT-Verbal	422 (3.1)	427 (7.5)	.31
PSAT-M	PSAT-Math	465 (3.6)	475 (8.2)	.49
AGE	Age at time of NELS F2 survey	18 (.5)	18 (.4)	<.01
FEMALE	Female %	56	58	.92
	Race/Ethnicity %			.51
AM_INDIAN	American Indian	<1	<1	
ASIAN	Asian	6	9	
BLACK	Black	9	8	
HISPANIC	Hispanic	8	9	
WHITE	White	77	74	
	Type of School %			.24
PUBLIC	Public	80	75	
PRIVATE	Catholic	13.2	16.7	
	Other Private	7	9	
	Location of School %			<.01
SCH_URB	Urban	36	47	
SCH_SUB	Suburban	43	44	
SCH_RUR	Rural	21	9	
SES	SES Index	.4 (.7)	.7 (.8)	<.01
	SES Quartile %			<.01
	Top Quartile	46	72	
	Second Quartile	29	17	
	Third Quartile	18	6	
	Bottom Quartile	8	5	
F1MATH	F1 Math Test Std Score	57 (8.2)	58 (7.8)	.54
F1READ	F1 Reading Test Std Score	57 (8.6)	56 (8.2)	.78
MTHCRD	Units of Math taken in high school	4 (.8)	4 (.6)	.53
MTHGRD	Weighted GPA in Math Courses	3 (.8)	3 (.8)	.01
	High School Program %			.18
RIGHSP	Rigorous Academic	40	41	
	General	54	56	
	Other %	7	4	
RE_ENG	Taken a remedial English course	6	9	.10
RE_MATH	Taken a remedial Math course	9	10	.58
AP	Taken an AP class	58	62	.27
F1ESTEEM	Self-Esteem Index	.12 (.7)	.18 (.6)	.25
F1LOCUS	Locus of Control Index	.20 (.6)	.20 (.6)	.99
	Homework done outside school (hrs per week) %			.20
HOMEWORK	16 or more hours	11	16	
	10-15 hours	24	27	
	1-9 hours	58	53	
	<1 hour	7	5	

standard errors in parenthesis

Table 2. Extrinsic Motivation by Coaching Status

Name	Description of Covariate	Uncoached	Coached	Stat Sig (p value)
HWTUTOR	Private tutor helped w/ homework in high school %	11	17	.02
	Student discussed plan to prepare for SAT w parents%			<.01
PPRESS	Often	20	45	
	Sometimes	55	36	
	Never	18	9	
	....Missing Response	7	10	
PARENT	Parents strongly encouraged student to prepare for SAT %	87	98	<.01
HI_MOT	Student scored below 1010 on PSAT, but has GPA > 3.25	11	22	<.01

Table 3. Selection Function Parameters Estimated using 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
N = 3,144
[a] These covariates are excluded from the regression equation

Table 4. SAT Coaching Effects using the Heckman Model

	SAT-V			SAT-M		
	$COACH_i$	$\lambda_i(COACH_i, \hat{s}_i)$	$\hat{\rho}$ of $(\delta_i, \varepsilon_i)$	$COACH_i$	$\lambda_i(COACH_i, \hat{s}_i)$	$\hat{\rho}$ of $(\delta_i, \varepsilon_i)$
SF1	69* (30)	-32* (16)	-.60	79* (30)	-33* (17)	-.64
SF2	58* (26)	-26 (14)	-.42	59* (28)	-22 (15)	-.36
SF3	0 (15)	7 (8)	.15	30 (16)	-6 (9)	-.10
SF4	17 (15)	-3 (9)	-.05	46* (16)	-16 (9)	-.25
SF5	12 (15)	-1 (8)	-.01	42* (15)	-13 (9)	-.20

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

SF1 : $\boldsymbol{Z_i}$ = all covariates in $\boldsymbol{X_i}$
SF2 : $\boldsymbol{Z_i}$ = all covariates in $\boldsymbol{X_i}$ + 1 covariate (*PARENT*) not used in $\boldsymbol{X_i}$
SF3 : $\boldsymbol{Z_i}$ = only covariates not in $\boldsymbol{X_i}$ (*HWTUTOR*, *PARENT*, *PPRESS*, *HI_MOT*)
SF4 : $\boldsymbol{Z_i}$ = covariates chosen by stepwise selection (*SCH_RUR*, *PPRESS*, *HWTUTOR*, *REMATH*, *HI_MOT*, *SES*, *MTHCRD*
SF5 : $\boldsymbol{Z_i}$ = covariates that were stat sig in coaching crosstabs (*AGE*, *SES*, *MTHGRD*, *SCH_RUR*, *HWTUTOR*, *PARENT*, *PPRESS*, *HI_MOT*)

Figure 2. Comparison of SAT-V Coaching Effect Estimates

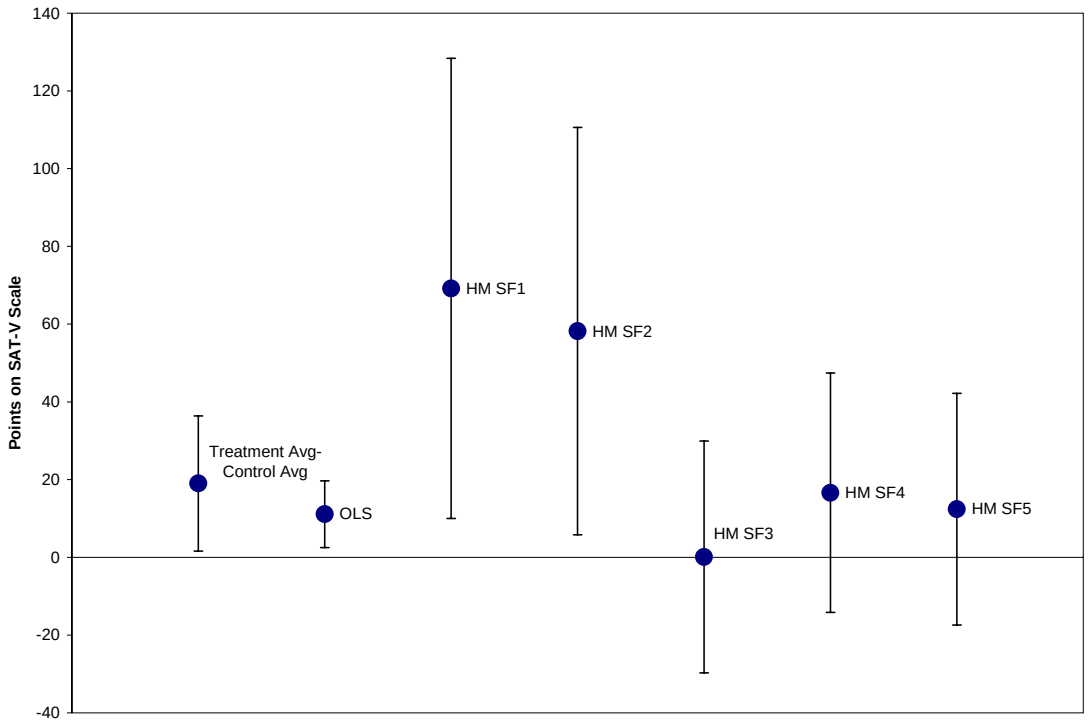

Figure 3. Comparison of SAT-M Coaching Effect Estimates