Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Econometrics 121 (2004) 99–124
JOURNAL OF Econometrics
www.elsevier.com/locate/econbase

How robust is the evidence on the effects of college quality? Evidence from matching

Dan A. Black[a], Jeffrey A. Smith[b,c,*]

[a] *Department of Economics and Center for Policy Research, Syracuse University, 426 Eggers Hall, Syracuse, NY 13244-1020, USA*
[b] *Department of Economics, University of Maryland, 3105 Tydings Hall, College Park, MD 20742-7211, USA*
[c] *National Bureau of Economic Research (NBER) and Institute for the Study of Labor (IZA), USA*

Abstract

We estimate the effects of college quality using propensity score matching methods and the National Longitudinal Survey of Youth 1979 cohort. Matching allows us to relax the linear functional form assumption implicit in regression-based estimates. We also examine the support problem by determining whether there are individuals attending low-quality colleges similar to those attending high-quality colleges, and find that the support condition holds only weakly. Thus, the linear functional form plays an important role in regression-based estimates (and matching estimates have large standard errors). Point estimates from regression and matching are similar for men but not women.

JEL classification: C14; I21

Keywords: Returns to college quality; Propensity score matching models

1. Introduction

A recent literature attempts to estimate the labor market effects of college quality, focusing on the wage effects of attending a higher quality college. The literature measures quality either in terms of inputs, such as expenditures per student or faculty salaries, or in terms of peer quality (or selectivity), such as the average SAT score of the entering class. Recent papers in this literature include Black et al. (2003a, b), Brand

* Corresponding author. Department of Economics, University of Maryland, 3105 Tydings Hall, College Park, MD 20742-7211, USA. Tel.: +1-301-405-3532; fax: +1-301-405-3542.
E-mail addresses: danblack@maxwell.syr.edu (D.A. Black), smith@econ.umd.edu (J.A. Smith).

0304-4076/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2003.10.006

(2002), Brewer et al. (1999), Dale and Krueger (2002), Light and Strayer (2000) and Turner (1998). The basic finding is that college quality matters for later labor market outcomes. [1]

The key econometric difficulty in this literature results from the non-random selection of students into colleges of varying qualities. Better students sort into better quality colleges (see, e.g., Hoxby, 1997). With a few exceptions, the literature relies on an assumption of what Heckman and Robb (1985) call "selection on observables" to identify the effects of college quality in the presence of non-random selection. Under this assumption, bias resulting from the differential selection of more able, more motivated, and otherwise better students into better colleges is removed by conditioning on pre-determined observable characteristics of the students. Standard practice in this literature, as in many others, consists of entering each characteristic in levels (with perhaps a squared term for age) in a linear outcome model such as a wage equation.

In this paper, we address two related weaknesses of this identification strategy. The first weakness arises from the fact that the linearity assumption can hide the failure of the "common support" condition. To illustrate the problem, consider the case in which only high test score students attend high-quality colleges and only low test score students attend low-quality colleges. The counterfactual outcome—what high test score students experience when attending low-quality colleges—is not non-parametrically identified. Instead, the linear functional form assumption identifies the counterfactual outcome. Even if the support problem does not prevent estimation of college quality effects, the assumption that linearly conditioning on the observables suffices to take account of selection bias remains problematic. This is the second weakness of the standard approach in the literature on the labor market effects of college quality. Given that theory does not suggest specific functional forms for outcome equations, and given the evidence in, e.g., Tobias (2003) on the importance of non-linearities in ability in returns to schooling, reliance on the linear functional form seems heroic.

Matching methods allow us to address both the support issue and the linear conditioning issue in a convenient way. Similar concerns motivate the matching analyses in Dearden et al. (2002), who examine the effects of secondary school quality in Britain, and Brand (2000), who estimates the impacts of college selectivity using data on Wisconsin high school graduates from 1957. Matching methods represent, depending on the particular method employed, either a semi-parametric or non-parametric alternative to linear regression. [2] While matching does not solve the support problem, it does highlight the problem in a way that the linear regression model does not. In order to reduce the dimensionality of our matching problem, we employ propensity score matching methods, in which we match on the predicted probability of attending a high-quality university, which is a function of observed X, rather than matching directly on X. [3]

[1] Brewer and Ehrenberg (1996) ably survey the earlier literature.

[2] See the discussions in Heckman et al. (1997, 1998b), Heckman et al. (1998a), Heckman et al. (1999), Dehejia and Wahba (1999, 2002), and Smith and Todd (2004). Current matching methods require binary (or multinomial) treatments. Thus, in our context, the treatment consists of attendance at a high-quality university rather than a low-quality university, where high and low refer to quartiles of the quality distribution in our sample. See Imbens (2000) and Lechner (2001) for the generalization to multiple treatments.

[3] Section 5 discusses the matching methods we employ in detail.

Once we have the distributions of estimated propensity scores for sample members in high- and low-quality universities, we can compare the two densities to get a clear sense of the extent of the common support problem.

Matching directly addresses the issue of conditioning only linearly on the X. The semi-parametric propensity score matching methods we adopt combine a flexible parametric logit specification for the propensity scores with non-parametric matching on the estimated scores. Untreated observations similar to each treated observation in terms of the probability of participation, $P(X)$, serve as counterfactuals. By constructing an observation-specific counterfactual for each treated observation, matching methods avoid bias due to misspecification of the functional form in a linear model.[4]

Using the data from the 1979 cohort of the National Longitudinal Survey of Youth (NLSY), we examine how students of different abilities, as measured by the first principal component of the 10 tests that comprise the Armed Services Vocational Aptitude Battery (ASVAB), sort into colleges of different qualities. For reasons discussed in Hoxby (1997), this sorting is of interest in its own right, and also informs our analysis of the support condition. We then estimate the probability that a student attends a college in the top quartile of the quality distribution in our sample, conditional on attending a college in either the top quartile or the lowest quartile. The predicted probabilities from this choice equation form the propensity scores we use to produce our matching estimates; we examine the common support condition using these estimated scores. We select the particular matching estimator we employ and the associated bandwidth using cross-validation methods, and then compare the estimates from propensity score matching to estimates of the same parameter based on the standard linear regression specification in the literature.

We reach five important empirical conclusions. First, we quantify the extent of sorting of students by ability into colleges of different qualities. There is less sorting in a random sample than in the non-random sample of high-end schools examined in Bowen and Bok (1998), and less than suggested by Herrnstein and Murray (1994). In addition, we find that the sorting is asymmetric: there are more high-ability students in low-quality colleges than low-ability students in high-quality colleges. Second, unlike the findings of Heckman and Vytlacil (2001) in their study of the returns to years of schooling, we find that sorting on ability is not sufficiently strong to cause the support condition to fail in our sample. Using our estimated propensity scores, however, we find that the support condition only weakly holds for persons with a high probability of attending a high-quality college. A small number of high-ability individuals at low-quality schools provide the counterfactual for a much larger number of high-ability individuals at high-quality schools. As a result, our matching estimates have larger standard errors than the corresponding OLS estimates. Third, although they are imprecisely estimated, our matching estimates for women differ substantially from the corresponding linear regression estimates, but the OLS and matching estimates for

[4] Other alternatives to the standard practice of entering each conditioning variable in levels in a linear regression include non-parametric regression and more flexible parametric regression models containing many higher order terms. Our aim here is to contrast the usual practice in the literature, which does not adopt these alternatives, with propensity score matching methods.

men are quite similar. Fourth, we find larger estimated effects in the "thick support" region defined by $0.33 < \hat{P}(X) < 0.67$. Larger effects in this region can result from heterogeneous treatment effects. They can also result from either measurement error in college quality or residual selection on unobservables, both of which should matter less in this region than in the tails of the propensity score distribution. This finding suggests that standard estimates may understate the labor market effect of college quality. Finally, we match on an alternative set of propensity scores containing only a small number of variables selected via cross-validation. We obtain larger estimated effects with much smaller standard errors, indicating a potential trade-off in finite samples between the plausibility of the conditional independence assumption and the variance of the estimates.

The remainder of the paper proceeds as follows. Section 2 describes the NLSY data we use in our analysis. Section 3 describes our measures of college quality and the construction of our college quality index. Section 4 examines the strength of the relationship between our measure of ability and the quality of college a student attends. Section 5 defines our parameter of interest and lays out the identifying assumptions on which we rely. Section 6 describes our propensity score estimates and examines the support problem. Section 7 presents standard linear regression estimates of the effect of college quality on wages using the NLSY data. Section 8 outlines the matching methods we use and Section 9 presents the corresponding estimated wage effects. Section 10 presents estimates based on an alternative specification of the propensity score. Section 11 concludes.

2. The NLSY data

Our primary data source is the National Longitudinal Survey of Youth (NLSY), a panel data set based on annual surveys of a sample of men and women who were 14–21 years old on January 1, 1979. Respondents were first interviewed in 1979 and an attempt has been made to re-interview them annually (biannually since 1994) since then. Of the five sub-samples that comprise the NLSY, we use only the representative cross-section and the minority over-samples. Table 1 presents basic descriptive statistics for our sample. The top panel gives (unweighted) statistics for the full sample, while the bottom panel gives statistics for the representative cross-section only. The sample includes only persons who had attended college at some point prior to the 1998 survey. [5]

The NLSY suits our purpose well for several reasons. First, the timing means that we have information on wages for a relatively recent cohort of college graduates that is old enough that the vast majority of those who will attend college have already done so. Furthermore, those who will attend graduate school have largely completed doing so

[5] For men (women) we start with 4100 (4299) observations in the 1998 cross-section. We drop 196 (156) observations with missing ASVAB scores, 512 (442) observations not completing high school, 1451 (1335) observations that complete high school but never attend college, 3 (1) observations with a missing value for race or ethnicity, 306 (516) observations with missing or zero wages and 922 (1157) observations with missing values for the variables in our college quality index.

Table 1
NLSY descriptive statistics, 1998

	Men	Women
Full sample		
Age	36.7	36.8
Black	0.239	0.280
Hispanic	0.166	0.167
Years of education	14.91	14.79
Associate degree	0.116	0.156
Bachelor's degree	0.411	0.363
Master's degree	0.148	0.157
N	1504	1695
Representative sample		
Age	36.7	36.8
Black	0.083	0.106
Hispanic	0.057	0.070
Years of education	15.15	14.92
Associate degree	0.101	0.149
Bachelor's degree	0.481	0.413
Master's degree	0.175	0.182
N	1012	1136

Note: Authors' calculations using unweighted NLSY data. The full sample includes all respondents while the representative sample excludes the minority and military over-samples. Both samples include only those respondents who attend college before the 1998 interview.

as well. Second, the NLSY confidential files provide information on individual colleges attended, which allows us to match up information on specific colleges from external sources. Third, the NLSY allows us to construct a compelling "ability" measure using the ASVAB, which was administered to over 90 percent of the sample.[6] Fourth, the NLSY is rich enough in other covariates to make the assumption that conditioning on observable characteristics alone solves the problem of non-random sorting into colleges of varying qualities plausible. These covariates include detailed information on family background, home environment and high school characteristics.

Table 2 describes the set of covariates included in the log wage regressions and in the propensity score models. For the region of birth, a dummy variable was created if the region could not be determined. For the family and high school variables if a particular measure could not be constructed because of missing data or invalid responses, we set the measure to zero and generated a dummy variable indicating that the data are missing. Our ability controls were created in two steps. First, we created age-adjusted ASVAB scores by regressing the scores from each of the ten ASVAB components on age dummy variables. The residuals from these regressions are the age-adjusted scores.

[6] Neal and Johnson (1996) describe the test in detail and discuss the issues of interpretation surrounding it.

Table 2
Variables for propensity score and wage equations

Log wage	Log of average real wage (1982 dollars) on all jobs held during the year
Basic characteristics	
Region of birth	A vector of 10 dummy variables indicating region in which respondent was born
Age	Respondent's age at the interview, quadratic in age is used
Years of education	Highest grade or year of school the respondent completed as of the 1998 interview. Only those who attended a college are in the sample
Black	Dummy variable indicating the respondent is black
Hispanic	Dummy variable indicating the respondent is Hispanic (black & Hispanic are mutually exclusive)
ASVAB test scores	Scores on the 10 components of the Armed Services Vocational Aptitude Battery, administered in 1980. We use the first two principal components of the age-adjusted scores.
Home characteristics	
Magazine	"When you were about 14 years old, did you or anyone else living with you get magazines regularly?"
Newspaper	"When you were about 14 years old, did you or anyone else living with you get a newspaper regularly?"
Library card	"When you were about 14 years old, did you or anyone else living with you have a library card?"
Mom education	Highest grade or year of school completed by respondent's mother.
Mom living	Was the respondent's mother living at the 1979 interview (when respondents were between 14 and 22 years old)?
Mom age	At the 1987 interview.
Dad education	Highest grade or year of school completed by respondent's father
Dad living	Was the respondent's father living at the 1979 interview?
Dad age	At the 1987 interview
Living together	Indicator for whether the respondent's mother and father lived in the same household at the 1979 interview
Mom occupation	Occupation of job held longest by mother or stepmother in 1978, represented by dummy variables for each Census 1-digit occupation
Dad occupation	Occupation of job held longest by father or stepfather in 1978, represented by dummy variables for each Census 1-digit occupation
High school characteristics	
Size of high school	Asked of respondents' high schools: "As of 10/1/79 [or nearest date] what was [your] total enrollment?"
Books	Asked of respondents' high schools: "What is the approximate number of catalogued volumes in the school library (enter 0 if your school has no library)." [in 1979]
Teacher salary	Asked of respondents' high schools: "What is the first step on an annual salary contract schedule for a beginning certified teacher with a bachelor's degree?" [in 1979]
Disadvantaged	Asked of respondents' high schools: "What percentage of the students in [the respondent's high school] are classified as disadvantaged according to ESEA [or other] guidelines?" [in 1979]

The first two principal components of the age-adjusted scores (and their squares) are the ability variables used throughout the paper. [7]

3. Measuring college quality

We matched data on a large number of variables related to college quality to the NLSY data using the information on college attended. [8] We only matched data on 4-year colleges; roughly one-half of the people in our sample attended a 4-year college, and many of the quality variables are not available for 2-year colleges. [9]

We make use of only three measures of college quality: average faculty salary in 1997, the average Scholastic Aptitude Test (SAT) score of the entering class in 1990 and the average freshman retention rate in 1990. The retention rate is the fraction of freshmen that return to the same school in their sophomore year. All three variables are presumptively positively related to college quality, but each reflects a different aspect of it. Faculty salaries represent a measure of inputs, the average SAT score represents a measure of selectivity (or, alternatively, of peer quality, which is a different sort of input), and the retention rate represents a "voting with your feet" measure of quality as perceived by students and their parents. [10] Descriptive statistics for the three variables included in our index appear in Table 3. The table documents substantial variation in all three measures among the colleges attended by both male and female NLSY respondents.

For reasons of parsimony, and also because we think that each of our individual quality variables represents an error-ridden measure of underlying quality, we combine the three variables into an index. In particular, we take the first principal component of our three variables as our index of college quality. We have examined the resulting ranking and find that it accords with a priori notions of quality. For example, the top five colleges in the data set according to this index are Stanford, MIT, Yale, Princeton, and the University of Pennsylvania.

[7] We were concerned about the fact that some NLSY respondents complete the ASVAB after starting college. To determine the empirical importance of this issue, we examined the differential in ASVAB scores between persons in the first and fourth quartiles of the college quality distribution as a function of their age at the time of the test and found no relationship.

[8] We obtained these variables from the Department of Education's Integrated Post-secondary Education Data System (IPEDS) for 1997 and the *US News and World Report's* (1991) Directory of Colleges and Universities. These variables change only very slowly, so utilizing values from a single point in time adds little measurement error.

[9] Our sample includes persons who went on to graduate study. The college quality variable refers to the most recent college attended as an undergraduate in all cases.

[10] We use only three quality measures in constructing the index because we do not observe each measure for all colleges, so that adding additional measures to the index reduces the sample size. We can construct the index utilized in this paper for 81.36 percent of the women who attended a 4-year college and 81.70 percent of the men. Black et al. (2003a, b) note that additional quality measures beyond the third (or, indeed, alternative sets of three quality measures) do little to change the index (or the findings, in a regression context).

Table 3
College quality measures, NLSY 1998

	Mean	25th percentile	50th percentile	75th percentile
Panel A: Men				
Faculty salaries ($N = 1312$)	$51,996	$43,646	$50,989	$59,284
Freshman retention rate ($N = 757$)	0.742	0.660	0.750	0.830
Average SAT score ($N = 832$)	935	835	927	1030
Panel B: Women				
Faculty salaries ($N = 1488$)	$50,205	$42,305	$49,418	$57,683
Freshman retention rate ($N = 739$)	0.735	0.680	0.740	0.830
Average SAT score ($N = 714$)	921	835	900	1005

Note: Authors' calculations using NLSY data, *US News and World Report's Directory of Colleges and Universities* data, and IPEDS data. Means are for the last college attended as of the 1998 interview.

4. The relationship between ability and college quality

In this section we examine the relationship between ability and college quality. This analysis provides a first pass at the support condition, as we expect substantial sorting on ability into colleges of different qualities. That sorting may suffice to cause the support condition to fail, even before conditioning on other variables. An examination of the extent and nature of sorting on ability holds interest in its own right as well. Existing studies such as Bowen and Bok (1998), Herrnstein and Murray (1994), and Cook and Frank (1993) examine primarily elite colleges near the top of the quality distribution. Hoxby (1997, Table 3) presents estimates of variation in mean student test scores among universities of varying qualities over time and estimates of the within-university variation in test scores over time, but does not look at the full joint distribution. Light and Strayer (2000) present similar evidence from the NLSY but using the selectivity of the first college attended rather than the quality of the last college attended.

Table 4 presents the joint density of student ability and college quality separately for men and women. In each panel, rows represent quintiles of the college quality distribution and columns represent quintiles of the ability distribution, where ability consists of the first principal component of the ASVAB scores. Each cell contains three numbers, the row percentage, the column percentage, and the cell percentage. Thus, in the upper left corner of Table 4 for men, we find that 6.48 percent of the sample is in both the first quintile of the ability distribution and the first quintile of the college quality distribution. As there are 25 cells and we are using quintiles, random sorting would yield roughly 4 percent in each cell, so this cell is substantially over-represented in the data.

Table 4
Bivariate distribution of ability and college quality measures, NLSY 1998

Quality index quintiles	Ability quintiles					
	First quintile	Second quintile	Third quintile	Fourth quintile	Fifth quintile	Total
Panel A: Men						
First quintile	(32.38)	(21.90)	(16.19)	(14.29)	(15.24)	(100.0)
	[32.38]	[21.90]	[16.19]	[14.29]	[15.24]	($N = 105$)
	6.48	4.38	3.24	2.86	3.05	
Second quintile	(23.81)	(20.95)	(20.95)	(20.95)	(13.33)	(100.0)
	[23.81]	[20.95]	[20.95]	[20.95]	[13.33]	($N = 105$)
	4.76	4.19	4.19	4.19	2.67	
Third quintile	(24.76)	(15.24)	(21.90)	(17.14)	(20.95)	(100.0)
	[24.76]	[15.24]	[21.90]	[17.14]	[20.95]	($N = 105$)
	4.95	3.05	4.38	3.43	4.19	
Fourth quintile	(11.54)	(18.27)	(27.88)	(20.19)	(22.12)	(100.0)
	[11.43]	[18.10]	[27.62]	[20.00]	[21.90]	($N = 104$)
	2.29	3.62	5.52	4.00	4.38	
Fifth quintile	(7.55)	(23.58)	(13.21)	(27.36)	(28.30)	(100.0)
	[7.62]	[23.81]	[13.33]	[27.62]	[28.57]	($N = 106$)
	1.52	4.76	2.67	5.52	5.71	
Total	[100.0]	[100.0]	[100.0]	[100.0]	[100.0]	100.0
	[$N = 105$]	[$N = 105$]	[$N = 105$]	[$N = 105$]	[$N = 105$]	$N = 525$
Panel B: Women						
First quintile	(31.07)	(19.42)	(20.39)	(15.53)	(13.59)	(100.0)
	[31.07]	[19.42]	[20.39]	[15.53]	[13.59]	($N = 103$)
	6.21	3.88	4.08	3.11	2.72	
Second quintile	(22.22)	(25.25)	(26.26)	(10.10)	(16.16)	(100.0)
	[21.36]	[24.27]	[25.24]	[9.71]	[15.53]	($N = 99$)
	4.27	4.85	5.05	1.94	3.11	
Third quintile	(25.71)	(19.05)	(20.95)	(19.05)	(15.24)	(100.0)
	[26.21]	[19.42]	[21.36]	[19.42]	[15.53]	($N = 105$)
	5.24	3.88	4.27	3.88	3.11	
Fourth quintile	(14.85)	(21.78)	(17.82)	(24.75)	(20.79)	(100.0)
	[14.56]	[21.36]	[17.48]	[24.27]	[20.39]	($N = 101$)
	2.91	4.27	3.50	4.85	4.08	
Fifth quintile	(6.54)	(14.95)	(14.95)	(29.91)	(33.64)	(100.0)
	[6.80]	[15.53]	[15.53]	[31.07]	[34.95]	($N = 107$)
	1.36	3.11	3.11	6.21	6.99	
Total	[100.0]	[100.0]	[100.0]	[100.0]	[100.0]	100.0
	[$N = 103$]	[$N = 103$]	[$N = 103$]	[$N = 103$]	[$N = 103$]	$N = 515$

Note: Authors' calculations using unweighted NLSY data, *US News and World Report's Directory of Colleges and Universities* data, and IPEDS data. The college quality measure is for the last college attended as an undergraduate as of the 1998 interview. The ability measure is the first principal component of the age-adjusted ASVAB scores. Samples include only respondents who attend colleges for which we can construct our college quality index.

Three main findings emerge from Table 4. First, there is substantial sorting based on ability. For both men and women, the fraction of observations on the diagonals, and the fraction on either the diagonal or the surrounding bands (persons with a difference of one between the two quintile rankings) exceed what would be expected from random sorting at the 5 percent level or better. For example, the percentages of observations on the diagonal are 24.76 and 27.17 for men and women, respectively, compared to the 20 percent expected in the absence of sorting. This sorting appears slightly stronger for women than for men. Second, a comparison of the off-diagonal corner cells suggests an asymmetry to the sorting, with more high-quality students at low-quality schools than the reverse. Light and Strayer (2000) find a much stronger asymmetry than we do when using the first, rather than the last, college attended. This difference suggests that high-ability students who start their academic careers at low-quality colleges often move to higher quality colleges over the course of their academic careers. Third, at the level of quartiles and quintiles, the sorting by ability alone does not threaten the validity of the support condition. Yet there is sufficient sorting to suggest that when looking at the support condition more finely, and with additional conditioning variables, troubles could arise. We return to the support issue in Section 6; in the next section we make precise our parameter of interest and the identifying assumptions underlying our estimates.

5. The parameter of interest and our identifying assumptions

Let Y_1 be the outcome in the "treated" state and Y_0 be the outcome in the "untreated" state. In our application, both groups receive a treatment in the literal sense. Thus, Y_1 corresponds to the potential outcome associated with attending a high-quality college (one in the upper quartile of our sample) and Y_0 corresponds to the potential outcome associated with attending a low-quality college (one in the lower quartile of our sample). We call these potential outcomes because we observe only one of (Y_1, Y_0) for each person. Let $D = 1$ indicate that a person attended a high-quality college and $D = 0$ indicate that a person attended a low-quality college. Finally, let X be a vector of observed covariates affecting both the choice of college quality and economic outcomes.

Our parameter of interest—the impact of treatment on the treated—is the mean effect of attending a high-quality college rather than a low-quality college on the persons who chose to attend a high-quality college. In terms of our notation, the parameter of interest is:

$$\Delta^{TT} = \mathrm{E}(Y_1 - Y_0 \mid D = 1). \tag{1}$$

If impacts are heterogeneous, this parameter may differ from the mean impact of attending a high-quality college on those persons currently attending a low-quality college and from the mean impact of attending a high-quality college on a randomly selected person. Our parameter, combined with information on the differential costs incurred by persons attending a high-quality college, provides evidence on the extent of any economic returns to those additional costs.

Both cross-sectional matching methods and standard linear regression analyses estimate the impact of a "treatment" under the assumption of selection on observables. That is, both approaches assume that conditioning on an available set of covariates removes all systematic differences in outcomes in the "untreated" state between high- and low-quality college attendees. The literature formalizes the selection on observables assumption that justifies matching as the Conditional Independence Assumption (CIA), given by

$$(Y_0 \perp D)|X. \tag{CIA}$$

This assumption states that the outcome in the base state (your wage if you attend a low-quality college) is independent of the treatment (attending a high-quality college), conditional on some set of observed covariates X. Put differently, within subgroups defined by X, attendance at a high-quality college is unrelated to what your outcome would be if you attended a low-quality college.[11] It is important to emphasize that the CIA is just that, an assumption. In any given context, it need not hold for any particular set of X available in the data, and it may not hold for any set of X variables in the available data. Moreover, cross-sectional matching does nothing to account for selection on unobservables, and can even act to increase the bias relative to not matching for certain configurations of the unobservables; see Heckman and Siegelman (1993).[12]

The key difference between matching and linear regression is that regression makes the additional assumption that simply conditioning linearly on X suffices to eliminate selection bias. Of course, with sufficient higher order terms, the linear model can approximate a given non-linear function of the X arbitrarily well. Most of the linear regression models in the college quality literature, however, include no higher order terms other than perhaps squared terms in age or experience. For such models, the linearity assumption potentially has real empirical bite.

Moreover, matching methods, but not linear regression, rely on the "common support" assumption, which can be expressed as

$$\Pr(D = 1 \mid X) < 1 \quad \text{for all } X. \tag{2}$$

The support condition states that, for each X satisfying the CIA, there must be some individuals who do not get treated; in our context, for each X, there must be some individuals who attend a low-quality college. If there are X for which everyone attends a high-quality college, then there is no way in a matching context to construct the counterfactual outcome for these observations.

Matching on X when X is of high dimension, as in our application, raises the problem of empty cells—the so-called curse of dimensionality. With high-dimensional X, the number of distinct vector values becomes very large, and many (even all in some contexts) of the treated persons will have no corresponding untreated person with

[11] As noted in Heckman et al. (1998a), this version of the CIA is stronger than we actually require. Mean independence conditional on the propensity scores $E(Y_0 \mid P(X), D=1) = E(Y_0 \mid P(X), D=0)$ suffices to identify our parameter of interest, and even this condition need only hold for respondents attending colleges in the two quartiles of the college quality distribution being compared in a given set of estimates.

[12] Difference-in-differences matching methods allow for selection on time-invariant unobservables. See the discussions in Heckman et al. (1998a) and Smith and Todd (2004).

exactly the same values of X. One response to this is to reduce the dimension of X by reducing the number of matching variables, but this will reduce the plausibility of the CIA. Instead, Rosenbaum and Rubin (1983) show that the assumptions that justify matching on X also justify matching on the probability of treatment, $\Pr(D = 1 \mid X)$, which the literature calls the "propensity score." The intuition behind propensity score matching is that subgroups with values of X that imply the same probability of treatment can be combined because they will always appear in the treatment and (matched) comparison groups in the same proportion. As a result, any differences between subgroups with different X but the same propensity score balance out when constructing the estimates. [13]

Our estimates constitute partial equilibrium estimates, and can be thought of as indicating the effect of changing college quality for one student at the margin. In formal terms, we make the Stable Unit Treatment Value Assumption (SUTVA), which states that (Y_1, Y_0) does not depend on who attends what college or on how many attend each type of college. A general equilibrium analysis of this question would allow for the endogeneity of the college quality measures in response to large changes in student choices.

6. Propensity scores and the common support condition

We estimate propensity scores for men and women using a logit model and the NLSY data. The propensity score specification for each group includes age, age squared, race/ethnicity, region of birth dummies, the first two principal components of the 10 ASVAB test scores and their squares, and characteristics of the respondent's high school, the respondent's parents, and the respondent's home environment as a child. They pass (men) or almost pass (women) standard balancing tests such as those described in Smith and Todd (2004) for $\widehat{P(X)} < 0.75$. We had trouble obtaining balance for high propensity scores due to the small number of $D = 0$ observations available.

We select our X variables to include factors expected to affect both the college quality a respondent selects as well as outcomes in the baseline, low college quality state. The only potentially controversial variable included in some of our estimated scores is years of schooling. This variable poses conceptual problems in this literature, as years of schooling depend in part on college quality, yet they also have a separate, exogenous effect on outcomes. Including years of schooling, whether in a linear regression context or in a matching context, understates the effect of college quality, as that part of the college quality effect that works through increasing years attended gets netted out in the conditioning. On the other hand, not including years of schooling risks assigning to college quality the effects of other factors that affect years of college

[13] The curse of dimensionality reappears when estimating the propensity scores unless a parametric model, such as a logit or probit, is employed to do so. The literature suggests gains from being non-parametric on outcomes (typically continuous) but parametric on participation (a binary variable), compared to utilizing a parametric model for outcomes, particularly when a flexible specification is employed in estimating the propensity scores.

attended and whether or not the student completes a degree. In the linear regression estimates presented in Section 7 we follow Dearden et al. (2002) and do it both ways; as in their paper, we find that it makes a difference to the estimates. As a result, we also report the matching estimates with and without including years of education in the propensity score specification.

In Fig. 1, we examine the support condition using the propensity scores that include years of education by plotting histograms of the estimated scores for both men and women. The graphs for the scores excluding years of education show a similar pattern. For each group, the top histogram corresponds to respondents who attended high-quality colleges (the $D=1$ group), while the bottom histogram corresponds to respondents who attended low-quality colleges (the $D=0$ group). The horizontal axis defines intervals of the propensity score and the height (or depth) of each bar on the vertical axis indicates the fraction of the relevant sample with scores in the corresponding interval.

Fig. 1 illustrates that, when looking more finely than the quartiles and quintiles examined in Section 4, and when considering propensity scores that incorporate additional covariates beyond ability, the support condition gets stretched even thinner. For both men and women, nearly 42 percent of the comparison group lies below the 5th percentile of the treatment group, and nearly 52 percent of the treatment group lies above the 95th percentile of the comparison group. Indeed, the mean propensity score given $D=1$ is about 0.70 while the mean for $D=0$ is about 0.30.

In Table 5, we provide an alternative way of examining the intensity with which the upper tail of the comparison group gets used in constructing the estimated counterfactual mean by presenting the deciles of the distribution of the relative weights from matching estimates presented in Section 9 based on an Epanechnikov kernel. The weights are normalized by the mean of the distribution so that a number less than one lies below the mean and a number greater than one lies above the mean. We see that the comparison group observations around the 80th percentile get used over 4.5 times more heavily than those around the 20th percentile, with the mean of the data at about the 65th percentile. Thus, while the support condition does not fail in our data, we are skating on thin ice in terms of identification for high values of the probability of participation. Comparing the comparable in these data means using only a small number of comparison observations to construct the counterfactual for a large number of treated observations.

To examine how much difference the thin support in the upper tail makes to the estimates, we present separate estimates for the "thick support" region, defined as $0.33 < \hat{P}(X) < 0.67$. Two concerns motivate these estimates. First, following Hausman et al. (1998), we worry that high $\hat{P}(X)$ respondents observed at low-quality schools may actually represent high $\hat{P}(X)$ respondents with measurement error in their college attended variable.

Second, we remain concerned about any lingering selection on unobservables, which will have its largest effects on bias for values of the propensity score in the tails of the distribution. To see this, suppose that you believe that earnings are given by

$$Y = g_0(X) + D(g_1(X) - g_0(X)) + \varepsilon_0 + D(\varepsilon_1 - \varepsilon_0), \tag{3}$$

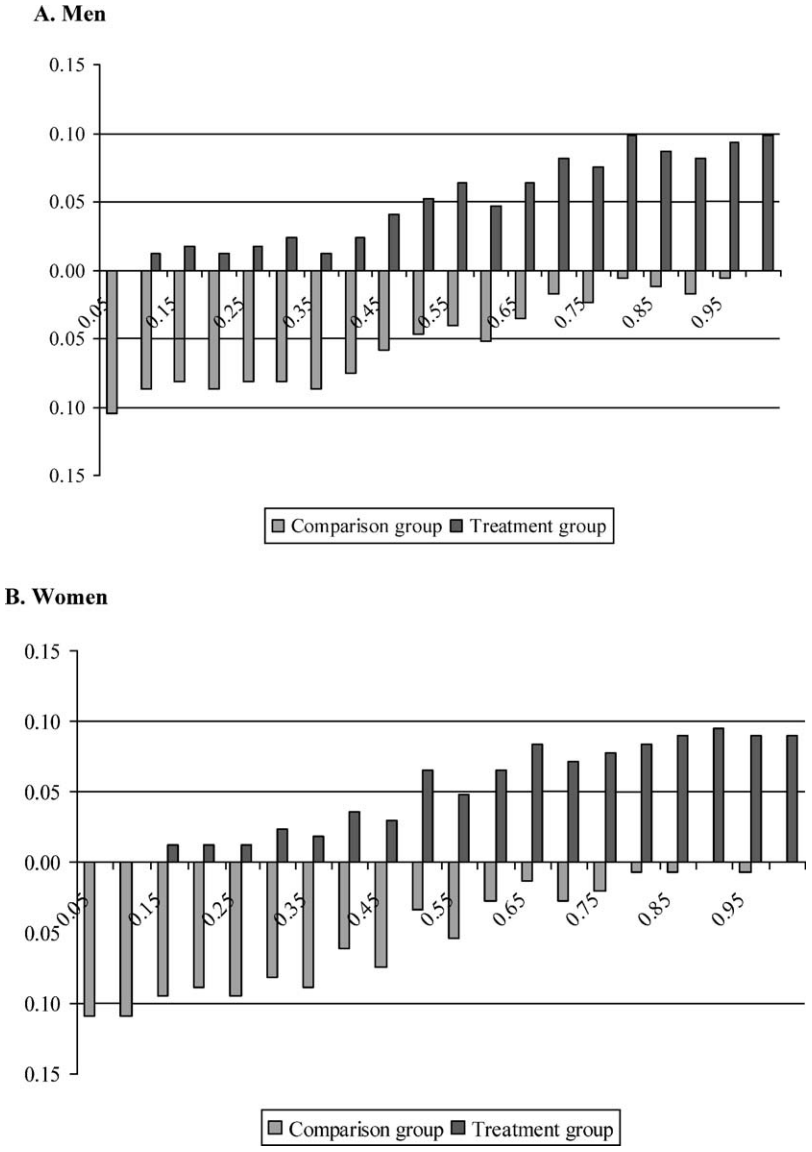

Fig. 1. The distributions of the propensity scores.

where $g_0(X)$ is the deterministic portion of the wage when in the comparison group, $g_1(X)$ is the deterministic portion of the wage when in the treatment group, and $(\varepsilon_0, \varepsilon_1)$ are the corresponding error terms. Treatment is determined by the latent variable $D^* = h(X) - u$ with $D = 1$ if $D^* > 0$ and $D = 0$ otherwise, and $\Pr(D^* > 0 \mid X) \equiv P(X)$. With the propensity score matching estimator, we want $\mathrm{E}(Y_0 \mid D = 1, P(X)) = \mathrm{E}(Y_0 \mid D = 0, P(X))$, but if there is any residual selection bias after conditioning on $P(X)$ and we

Table 5
Distribution of weights from Epanechnikov kernel of the comparison group, NLSY 1998

Deciles of the distributions of weights	Weights relative to mean weight
10 percentile	0.33
20 percentile	0.39
30 percentile	0.46
40 percentile	0.57
50 percentile	0.70
60 percentile	0.88
70 percentile	1.21
80 percentile	1.76
90 percentile	2.32
N	172

Note: Authors' calculations using unweighted NLSY data, *US News and World Report's Directory of Colleges and Universities* data, and IPEDS data. College quality is for the last college attended as an undergraduate as of the 1998 interview. The propensity scores are estimated using a logit model and the specification includes quadratics in the first two principal components of the age-adjusted ASVAB score, a black indicator, a Hispanic indicator, age, age squared, region of birth indicators, and high school, parental, and home characteristics. All matching weights are divided by the mean matching weight.

assume that (u, ε_0) are jointly normal, we will have the bias

$$\begin{aligned} \mathrm{B}(P(X)) &= \mathrm{E}(\varepsilon_0 \mid D^* > 0, h(X)) - \mathrm{E}(\varepsilon_0 \mid D^* \leqslant 0, h(X)) \\ &= \rho\sigma_{\varepsilon_0} \frac{\phi(h(X))}{\Phi(h(X))[1 - \Phi(h(X))]}. \end{aligned} \tag{4}$$

This bias is minimized about $h(X)=0$ or $P(X)=1/2$. Indeed, if we take the distribution of estimated propensity scores as the true $P(X)$, assume the error terms are normal, and assume that we can exactly match each individual in the treatment group with individuals having exactly the same $P(X)$, the selection bias from using the entire treatment group sample is 7.77 times larger than the bias if we limit our sample to $P(X) \in (0.33, 0.67)$. While this analytic result obviously relies on the joint normality, the intuition underlying the result does not depend on distributions: when the probability of being in the treatment group is high, unobservable factors on average play a larger role than for probabilities near 0.5. Thus, when matching estimators must rely on the right tail of the distribution of propensity scores in the comparison group, the selection bias may be considerable even when u and ε_0 are only weakly correlated.

7. Regression estimates of the impact of college quality

In this section we present standard regression-based estimates of the impact of college quality on wages. In particular, Table 6 presents evidence from 16 linear regression models, eight for men and eight for women. The dependent variable is the natural log

Table 6
Regression estimates of the wages effects of college quality, NLSY 1998

	College quality index			
Men				
Without years of education				
Ability measures	No	Yes	Yes	Yes
Individual characteristics	No	No	Yes	Yes
Home, high school, and parental characteristics	No	No	No	Yes
Second quartile	0.080	0.054	0.031	0.026
	(0.0501)	(0.0490)	(0.0491)	(0.0499)
Third quartile	0.170	0.132	0.095	0.082
	(0.0472)	(0.0457)	(0.0459)	(0.0473)
Fourth quartile	0.280	0.220	0.177	0.158
	(0.0480)	(0.0475)	(0.0490)	(0.0492)
With years of education				
Second quartile	0.044	0.033	0.007	0.005
	(0.0491)	(0.0486)	(0.0487)	(0.0497)
Third quartile	0.107	0.094	0.055	0.050
	(0.0464)	(0.0457)	(0.456)	(0.0469)
Fourth quartile	0.192	0.167	0.123	0.116
	(0.0469)	(0.0468)	(0.0481)	(0.0492)
Years of education	0.048	0.038	0.038	0.032
	(0.0059)	(0.0063)	(0.0064)	(0.0062)
Women				
Without years of education				
Ability measures	No	Yes	Yes	Yes
Individual characteristics	No	No	Yes	Yes
Home, high school, and parental characteristics	No	No	No	Yes
Second quartile	0.144	0.127	0.105	0.102
	(0.0385)	(0.0377)	(0.0377)	(0.0387)
Third quartile	0.135	0.105	0.075	0.065
	(0.0401)	(0.0394)	(0.0402)	(0.0406)
Fourth quartile	0.205	0.149	0.124	0.112
	(0.0418)	(0.0416)	(0.0418)	(0.0422)
With years of education				
Second quartile	0.115	0.105	0.083	0.082
	(0.0370)	(0.0366)	(0.0368)	(0.0378)
Third quartile	0.090	0.074	0.043	0.039
	(0.0390)	(0.0386)	(0.0394)	(0.0398)
Fourth quartile	0.136	0.107	0.078	0.074
	(0.0410)	(0.0412)	(0.0416)	(0.0421)
Years of education	0.054	0.048	0.047	0.042
	(0.0050)	(0.0050)	(0.0050)	(0.0051)

Note: Authors' calculations using unweighted NLSY data, *US News and World Report's Directory of Colleges and Universities* data, and IPEDS data. College quality is for the last college attended as an undergraduate. There are 1632 observations for men and 1849 for women. Each model includes indicator variables for the 2nd, 3rd and 4th quartiles of the quality distribution and an indicator for missing college quality. The ability measures consist of quadratics in the first two principal components of the age-adjusted ASVAB scores. The individual characteristics include a black indicator, an Hispanic indicator, age, age squared, and Census region of birth indicators. The high school, parental, and home characteristics comprise the remaining variables listed in Table 2 other than years of education, which is included in the model when indicated. Huber–White standard errors are reported in parentheses.

of the respondent's real wage in 1998.[14] In addition to the conditioning variables, we include indicator variables for having attended a college in the second, third, and fourth quartile of the quality distribution in our sample, as measured by the college quality index described in Section 3, and for having a missing value of the quality index. The first quartile of the quality distribution is the omitted group and, therefore, the implicit counterfactual. The columns vary the set of conditioning variables included in the model. Table 2 describes the individual conditioning variables and the notes to Table 6 define the blocks of variables included in each column. The top panel for each group omits the years of education variable, while the bottom panel includes it.

The regression results embody five patterns of interest. First, as shown in Black et al. (2003a, b), conditioning on ability makes a big difference to the estimates; in particular, it reduces the estimated effect by about one-quarter. This is consistent with the sorting shown in Section 4. Second, including additional individual characteristics other than ability, such as race, age, and region of birth, again reduces the estimated effects. Third, including the years of schooling variable reduces the estimated effects by about a third relative to the corresponding specification without years of schooling. Fourth, for men, but not for women, the estimated effects increase monotonically as college quality increases regardless of the set of included covariates. Finally, if we look at the two specifications that correspond to our propensity scores, we find that attending a high-quality college rather than a low-quality college increases wages by 11 or 12 percent for men and by about 7.5 percent for women. These findings are broadly consistent with the regression-based literature.

8. Matching methods

A variety of different methods exist for implementing matching. These methods differ in the specific weights assigned to each comparison group observation. All matching estimators have the generic form

$$\hat{\mathrm{E}}(Y_0 \mid \hat{P}(X_i)) = \sum_{j=1}^{J} w(\hat{P}(X_i), \hat{P}(X_j)) Y_{0j} \tag{5}$$

for the individual counterfactual for treated observation i. In this equation, $j = 1, \ldots, J$ indexes the untreated comparison group observations. All matching estimators construct an estimate of the expected unobserved counterfactual for each treated observation by taking a weighted average of the outcomes of the untreated observations. What differs among the various matching estimators is the specific form of the weights.

[14] In particular, the dependent variable is the log of the average real wage (in 1982 dollars) over all jobs held in 1998. Two variables are used to construct wages: total income from wages and salary in the past calendar year and number of hours worked in the past calendar year. The wage variable equals the log of total wage income divided by total hours worked. Persons with no jobs in 1998 are excluded from the sample. We exclude 9.86 percent of the male respondents and 16.0 percent of the female respondents due to zero earnings. To determine the sensitivity of our results to this exclusion, we constructed estimates using earnings levels as the dependent variable both including and excluding the zeros and found that it made little substantive difference to the results.

We consider three alternative matching estimators in our empirical work: the nearest neighbor estimator, the Gaussian kernel estimator, and the Epanechnikov kernel estimator. Asymptotically, all the different matching estimators produce the same estimate, because in an arbitrarily large sample, they all compare only exact matches. In finite samples, different matching estimators produce different estimates because of systematic differences between them in which observations they assign positive weight, how much weight they assign them, and how they handle (implicitly) the support problem.

Given the large number of competing estimators, we immediately face a problem of which estimator to use. We use a least squares leave-one-out validation mechanism to choose among the nearest neighbor, Gaussian kernel, and Epanechnikov kernel estimators and to pick the bandwidth for the two kernel estimators; see Racine and Li (2004) and Pagan and Ullah (1999) for discussions of leave-one-out validation. Recall that the estimation problem we face is to estimate the missing counterfactual Y_0 for those who attend colleges in the highest quality quartile. Unfortunately, we have no observations of Y_0 for the treatment group, but we do, of course, have observations on Y_0 in the comparison group. Leave-one-out validation uses these observations to determine which of the competing models best fit the data.

As the name implies, leave-one-out validation drops the jth observation in the comparison group and uses the remaining $N-1$ observations in the comparison group to form an estimate of Y_{0j}, which may be denoted $\hat{Y}_{0j,-j}$. The associated forecast error is given by $e_{j,-j} = Y_{0j} - \hat{Y}_{0j,-j}$. As the estimation does not include the jth observation, it represents an "out-of-sample" forecast, and, because the estimation sample is of size $N-1$ (rather than N), it presumably does a good job of replicating the essential features of the estimation problem. Repeating the process for the remaining $N-1$ observations allows comparisons of the mean squared error or root mean squared error of the forecasts associated with different matching estimators (or bandwidths, when selecting a bandwidth) to guide the choice of estimator (or bandwidth).

The use of the leave-one-out validation mechanism yields three interesting results. First, the nearest neighbor estimator performs worse than either the Gaussian kernel or the Epanechnikov kernel estimator. This is consistent with the findings in Fröhlich's (2004) Monte Carlo analysis of the performance of alternative matching methods. Second, the Epanechnikov kernel estimator performs modestly better than the Gaussian kernel almost independent of the bandwidth selected. Third, the performance of the estimators is relatively insensitive to the bandwidth selected until one gets to very small bandwidths. Given these results, we rely on the Epanechnikov kernel to construct the matching estimates presented in the remaining sections. In addition to its superior performance in our cross-validation exercise, the Epanechnikov kernel converges faster than the Gaussian kernel because it has only a limited support and, for the same reason, it implicitly imposes the support condition through the choice of the bandwidth.

9. Matching estimates of the impact of college quality

Our matching estimates of the impact on wages of attending a high-quality rather than a low-quality college appear in Table 7. As described in Section 8, we present matching

Table 7
Propensity score estimates of the effects of college quality: fourth and first quartiles, NLSY 1998

$\Delta_{41} = Y_{i4} - Y_{i1}$	Men		Women	
	Using years of education in propensity score estimation	Not using years of education in propensity score estimation	Using years of education in propensity score estimation	Not using years of education in propensity score estimation
Epanechnikov kernel, bandwidth 0.40 for men and 0.30 for women	0.120 (0.0867) [$n = 158$]	0.139 (0.0767) [$n = 152$]	0.067 (0.0862) [$n = 145$]	0.078 (0.0830) [$n = 155$]
OLS estimates	0.122 (0.0584)	0.159 (0.0584)	0.112 (0.0557)	0.155 (0.0552)
Thick support region	0.199 (0.1357) [$n = 44$]	0.250 (0.1181) [$n = 44$]	0.124 (0.1407) [$n = 39$]	0.157 (0.1418) [$n = 39$]
OLS estimates, thick support region	0.121 (0.0639)	0.156 (0.0653)	0.144 (0.0724)	0.184 (0.0720)

Note: Authors' calculations using unweighted NLSY data, *US News and World Report's Directory of Colleges and Universities* data, and IPEDS data. College quality is for the last college attended. There are 177 observations in comparison group and 176 in the treatment group for men and 173 in both the treatment group and the comparison group for women. The propensity scores are estimated using a logit model and the specification includes years of schooling (in columns 1 and 3 only), quadratics in the first two principal components of the age-adjusted ASVAB scores, a black indicator, an Hispanic indicator, age, age squared, region of birth indicators, and high school, parental, and home characteristics. The OLS estimates use only the observations with college quality in the first or fourth quartile. For the OLS estimates, Huber–White standard errors are reported in parentheses. Bandwidths are selected using a minimum root mean squared error criterion from leave-one-out cross-validations. Bootstrap standard errors for the matching estimates are based on 2000 replications.

estimates using the Epanechnikov kernel with leave-one-out cross-validated bandwidths. In each case, we present two alternative estimates: one that uses education in the estimation of the propensity score and one that excludes education from the propensity score. We also indicate, for each estimate, the number of treated observations for which an estimated counterfactual could be constructed using that particular estimator. Bootstrap standard errors based on 2000 replications appear in parentheses below each estimate. Each bootstrap includes re-estimation of the propensity scores used in the matching on the bootstrap sample. In the second row, we present the OLS estimates of the parameter of interest. These estimates differ from those in Table 6 because the sample includes only persons who attended a college in the first or fourth quartile of the quality distribution. This sample corresponds to that used for the matching estimates. The differences in the estimates that result from changing the sample signal the potential importance of relaxing the linear functional form assumption through matching.

In the third row, we present estimated impacts for the "thick support" region, defined in Section 6 as the region with $0.33 < \hat{P}(X) < 0.67$. In this region, there are substantial numbers of observations in both the treatment and comparison groups. The final row of the table presents OLS estimates of the impact of treatment for observations in the "thick support" region.[15]

For men, the OLS estimates indicate a 13 to 17 percent increase in wages as the effect of moving from a college in the first quartile of the quality distribution to one in the fourth quartile. Both OLS estimates are statistically significant at the 5 percent level. In contrast, the matching estimates range from 0.120 to 0.139, suggesting modestly smaller impacts. The estimates are a bit smaller when we condition on education than when we do not. Neither of the matching estimates is statistically significant at conventional levels; relaxing the linear functional form assumption has a price.

The estimates tell a similar story for women. Here the OLS estimates indicate a wage effect of 12–17 percent associated with attending a high-quality college. Both are significant at the five percent level. The full sample matching estimates range from 0.067 to 0.078, but are less than their corresponding standard errors.

In contrast to the full sample estimates, the matching estimates that consider only the "thick support" region are almost all higher than the corresponding OLS estimates, sometimes substantially so. For men, these estimates equal 0.199–0.250 with and without conditioning on years of education, respectively. The corresponding estimates for women equal 0.124 and 0.157. Following the discussion in Section 6, higher estimates in the common support can have one of three sources. First, there may be heterogeneous treatment effects, with higher impacts for middle values of the propensity score (something somewhat difficult to reconcile with an economic model of college quality choice). Second, they may result from measurement error in college quality in the tails of the distribution. Third, they may result from lingering selection on unobservables, which has a larger effect for values of $P(X)$ outside the thick support region.

Table 8 presents matching and OLS estimates of the effect of "treatment on the treated" for attending a college in the third quality quartile rather than the first quartile and of attending a college in the second quality quartile rather than the first quartile. To ease the comparison, we also replicate the corresponding estimates from Table 6. Several patterns emerge. First, for men, the matching estimates are always smaller than the corresponding OLS estimates. Second, for both men and women, the OLS estimates are monotonically increasing as one moves to higher quartiles of college quality, and for men, the same is true of the matching estimator. For women, however, the matching estimator monotonically *declines* with increases in the quartile of college quality, although we cannot reject the null hypothesis that all of the estimates are the same. Fourth, the estimated standard errors for the matching estimates always exceed those for the OLS estimates. Thus, while the estimates do not tell a strong story about

[15] We also formed estimators using the trimming mechanism suggested by Heckman et al. (1998a, b). Their scheme defines the region of common support in terms of estimates of the densities of the propensity score in the $D = 1$ and 0 samples. When we apply a low cutoff value for the densities, we obtain estimates similar to the estimates for the full sample in Table 7. When we apply a high cutoff value for the densities, and drop around 40 percent of the sample, we obtain estimates similar to our thick support estimates in Table 7.

Table 8
Propensity score estimates of the effects of college quality, NLSY 1998

	Not using years of education in propensity score estimation	
	Men	Women
$\Delta_{41} = Y_{i4} - Y_{i1}$		
Epanechnikov kernel, bandwidth 0.40 for men and 0.30 for women	0.139 (0.0767) [$n = 152$]	0.078 (0.0830) [$n = 155$]
OLS estimates	0.159 (0.0584)	0.155 (0.0552)
$\Delta_{31} = Y_{i3} - Y_{i1}$		
Epanechnikov kernel, bandwidth 0.30 men and 0.50 women	0.056 (0.0695) [$n = 166$]	0.118 (0.0561) [$n = 133$]
OLS estimates	0.082 (0.0541)	0.104 (0.0498)
$\Delta_{21} = Y_{i3} - Y_{i1}$		
Epanechnikov kernel, bandwidth 0.20 for men and 0.50 for women	0.006 (0.0863) [$n = 147$]	0.123 (0.506) [$n = 159$]
OLS estimates	0.072 (0.0584)	0.094 (0.0458)

Note: Authors' calculations using unweighted NLSY data, *US News and World Report's Directory of Colleges and Universities* data, and IPEDS data. College quality is for the last college attended. There are 177 observations in the comparison group for men and 173 in the comparison group for women. In the fourth quartile, there are 176 in the treatment group for men and 173 in the treatment group for women. In the third quartile, there are 179 men and 171 women in the treatment group. In the second quartile there are 178 men and 175 women in the treatment group. The propensity scores are estimated using a logit model and the specification includes quadratics in the first two principal components of the age-adjusted ASVAB scores, a black indicator, an Hispanic indicator, age, age squared, region of birth indicators, and high school, parental, and home characteristics. OLS models are estimated separately for each quartile. For the OLS estimates, Huber–White standard errors are reported in parentheses. Bandwidths are selected using a minimum root mean squared error criterion from leave-one-out cross-validations. Bootstrap standard errors for the matching estimates are based on 2000 replications.

bias in the OLS estimates, they do tell an important story about the support problem in this context. The support condition does not fail here, but it holds so weakly that the matching estimates end up having high variances.

Taken together, the estimates in Tables 7 and 8, along with the frequency of use analysis for the comparison group observations in Table 5, teach two related lessons. First, substantively, our point estimates provide some reason for concern about college quality effect estimates based on OLS regressions that control only linearly for covariates. Second, in commonly used data sets similar in sample size to the NLSY, and

covering persons who have attended college during years where there is substantial sorting based on ability and other characteristics, it is likely that the support condition barely holds, with the result that the data lack sufficient information for strong inference regarding the wage effects of college quality without the imposition of functional form assumptions.

10. A minimal specification

The matching estimates presented in Section 9 relied on a matching estimator and associated bandwidth selected via leave-one-out cross-validation. In this section, we also rely on leave-one-out cross-validation to choose the set of variables included in the propensity score. In conceptual terms, this amounts to choosing the propensity score model based on goodness-of-fit considerations rather than based on theory and evidence about the set of variables related to both college quality choice and labor market outcomes. We implemented the model selection procedure by starting with a model containing only the first two principal components of the ASVAB scores (and their squares) and indicators for whether the respondent was black or Hispanic. We then successively added blocks of additional variables, such as the individual characteristics, the home environment variables, the parental characteristics variables and the high school characteristics variables, and compared the resulting mean squared errors. To our considerable surprise, the minimal specification we started with out-performed the full specification and every other competing specification we examined. Not surprisingly given the small number of conditioning variables, satisfying the common support condition poses much less of a problem with the minimal specification, as the conditional mean of the estimated propensity scores equals 0.60 for $D = 1$ and 0.40 for $D = 0$. Perhaps as a result, the cross-validation selects a narrower bandwidth for women and a much narrower bandwidth for men.

We present estimates for the parameter $\Delta_{41} = Y_{i4} - Y_{i1}$ based on the minimal specification (and without conditioning on years of education) in Table 9. For both men and women, the minimal specification yields larger point estimates than the full specification. In addition, the standard errors get much smaller, so that both estimates are statistically significant at conventional levels. We also present OLS estimates using the same minimal set of conditioning variables; these prove quite similar to the corresponding matching estimates.

While it is tempting to claim that this minimal specification is indeed the correct specification, we do not accept this interpretation. Too many studies have documented the important role of family background for both labor market outcomes and college quality choices to simply dismiss them from the analysis. Why then does the fit of the model get worse when we include such variables in the propensity score estimation? We believe the answer lies in the common support problem. With the full specification, the treatment and comparison groups have much more distinct distributions of propensity scores than with the minimal specification. As a result, the cross-validation selects somewhat larger bandwidths for women and much larger bandwidths for men with the full specification. As is well known, increases in the bandwidth induce more

Table 9
Propensity score estimates of the effects of college quality, cross-validation specification, NLSY 1998

	Not using years of education in propensity score estimation	
	Men	Women
$\Delta_{41} = Y_{i4} - Y_{i1}$		
Epanechnikov kernel,	0.189	0.159
bandwidth 0.20 for men	(0.0488)	(0.0490)
and 0.25 for women	[$n = 169$]	[$n = 171$]
OLS estimates	0.204	0.151
	(0.0477)	(0.0435)

Note: Authors' calculations using unweighted NLSY data, *US News and World Report's Directory of Colleges and Universities* data, and IPEDS data. College quality is for the last college attended. There are 177 observations in comparison group and 176 in the treatment group for men and 173 in both the treatment group and the comparison group for women. The propensity score specification includes only the first two principal components of the age-adjusted ASVAB scores and their squares, a black indicator, and an Hispanic indicator. The OLS estimates use only the observations with college quality in the first or fourth quartile. For the OLS estimates, Huber–White standard errors are reported in parentheses. Bandwidths are selected using a minimum root mean squared error criterion from leave-one-out cross-validations. The root mean squared error criterion indicates that the propensity score specification underlying the estimates in this table outperforms the corresponding specification in Table 7. Bootstrap standard errors for the matching estimates are based on 2000 replications.

bias as observations less similar to each treated observation receive greater weight in constructing the estimated counterfactuals. Thus, we suspect that our minimal specification reduces the bias resulting from the relatively wide bandwidths employed in the full specification. The cost of this reduction in the bias from smoothing appears to consist of additional selection bias resulting from the failure of the minimal specification to condition on all of the covariates required for the CIA to plausibly hold in the data.

11. Conclusions

In this paper, we have investigated two potential weaknesses in the most commonly used econometric approach in the literature that estimates the labor market effects of college quality. These weaknesses are failure to attend to the support condition, which may be problematic in this context due to the sorting of highly qualified students into higher quality colleges, and the failure to condition non-linearly on important covariates such as ability. We have five main findings.

First, there is substantial sorting based on ability into colleges of differing qualities for both men and women in the NLSY. Higher ability students disproportionately attend higher quality colleges. We find some evidence of an asymmetry in this sorting, with more high-ability students at low-quality colleges than low-ability students at high-quality colleges. Sorting on ability alone, however, does not break the support condition.

Second, using our estimated propensity scores, which include ability as well as numerous other background variables, we show that the support condition, while it does

not fail, holds only weakly in our data. In particular, our data include only a handful of individuals who attend low-quality colleges but have characteristics that give them a high probability of attending a high-quality college. As a result, we end up with large standard errors. This is not a problem with the matching estimator; rather, it is a problem with the data. Running linear regressions hides the problem by implicitly borrowing strength from comparison observations with lower probabilities of attending a high-quality college.

Third, our estimates raise some concerns regarding the conventional practice of using linear selection on observables models to investigate the labor market effects of college quality. Although the point estimates from our matching estimators are imprecise, they sometimes differ substantially from the corresponding OLS estimates. In all cases, however, the matching estimates support the overall finding of the regression-based literature that college quality matters for labor market outcomes.

Fourth, our estimates based only on the "thick support" region of propensity scores around 0.5 consistently turn out larger than those constructed using the full sample. This difference could arise from genuinely larger impacts in this region, though we think this unlikely on theoretical grounds. More likely, it results from either measurement error in college quality or lingering selection on unobservables, both of which play a bigger role outside the thick support region than within it.

Fifth, a comparison of our full propensity score specification, which includes a rich set of covariates affecting both college quality choice and labor market outcomes, with the minimal specification selected by cross-validation on the basis of goodness-of-fit reveals an interesting and empirically important trade-off. In our full specification bias arises from selecting a wide bandwidth in response to the weakness of the common support. In the minimal specification, selection bias arises from leaving out many of the variables whose presence makes the CIA plausible and hence justifies matching (and regression) in our data. This trade-off also affects the estimated standard errors, which are much smaller for the minimal specification wherein the support condition poses much less of a problem.

Acknowledgements

This research was supported in part by the Social Science and Humanities Research Council of Canada. We thank Alex Whalley for excellent research assistance, Markus Frölich, Shannon Seitz, Barbara Sianesi, Alex Whalley, four anonymous referees and seminar participants at the Board of Governors of the U.S. Federal Reserve, George Mason, Ohio State, St. Gallen, Sydney, UCD, UCL, UC-Irvine, UCLA and Uppsala for helpful comments. We also thank Barbara Sianesi for comments and for providing her matching program.

References

Black, D., Daniel, K., Smith, J., 2003a. College quality and the wages of young men. University of Maryland, unpublished manuscript.

Black, D., Daniel, K., Smith, J., 2003b. College quality and the wages of young women. University of Maryland, unpublished manuscript.

Bowen, W., Bok, D., 1998. The Shape of the River: Long-term Consequences of Considering Race in College and University Admissions. Princeton University Press, New Jersey.

Brand, J., 2000. Matching on propensity scores to estimate the effects of graduating from an elite college on early- and mid-career outcomes. MS Thesis, University of Wisconsin-Madison, unpublished.

Brewer, D., Ehrenberg, R., 1996. Does it pay to attend an elite private college? Evidence from the senior class of 1980. In: Polachek, S. (Ed.), Research in Labor Economics, Vol. 15. JAI Press, Greenwich, CT, pp. 239–271.

Brewer, D., Eide, E., Ehrenberg, R., 1999. Does it pay to attend an elite college? Cross cohort evidence on the effects of college type on earnings. Journal of Human Resources 34, 104–123.

Cook, P., Frank, R., 1993. The growing concentration of top students at elite schools. In: Clotfelter, C., Rothschild, M. (Eds.), Studies of Supply and Demand in Higher Education. University of Chicago Press for NBER, Chicago, pp. 121–140.

Dale, S.B., Krueger, A., 2002. Estimating the payoff to attending a more selective college: an application of selection on observables and unobservables. Quarterly Journal of Economics 117, 1491–1528.

Dearden, L., Ferri, J., Meghir, C., 2002. The effect of school quality on educational attainment and wages. Review of Economics and Statistics 84, 1–20.

Dehejia, R., Wahba, S., 1999. Causal effects in non-experimental studies: re-evaluating the evaluations of training programs. Journal of the American Statistical Association 94, 1053–1062.

Dehejia, R., Wahba, S., 2002. Propensity score matching methods for non-experimental causal studies. Review of Economics and Statistics 84, 151–161.

Fröhlich, M., 2004. Finite sample properties of propensity-score matching and weighting estimators. Journal of Econometrics, forthcoming.

Hausman, J., Abrevaya, J., Scott-Morton, F., 1998. Misclassification of a dependent variable in a discrete choice setting. Journal of Econometrics 87, 239–269.

Heckman, J., Robb, R., 1985. Alternative methods for evaluating the impact of interventions. In: Heckman, J., Singer, B. (Eds.), Longitudinal Analysis of Labor Market Data. Cambridge University Press, Cambridge, pp. 156–246.

Heckman, J., Siegelman, P., 1993. The urban institute audit studies: their methods and findings. In: Fix, M., Struyk, R. (Eds.), Clear and Convincing Evidence: Measurement of Discrimination in America. Urban Institute Press, Washington, DC, pp. 187–258.

Heckman, J., Vytlacil, E., 2001. Identifying the role of cognitive ability in explaining the level of and change in the return to schooling. Review of Economics and Statistics 83, 1–12.

Heckman, J., Ichimura, H., Todd, P., 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Review of Economic Studies 64, 605–654.

Heckman, J., Ichimura, H., Smith, J., Todd, P., 1998a. Characterizing selection bias using experimental data. Econometrica 66, 1017–1098.

Heckman, J., Ichimura, H., Todd, P., 1998b. Matching as an econometric evaluation estimator. Review of Economic Studies 65, 261–294.

Heckman, J., LaLonde, R., Smith, J., 1999. The economics and econometrics of active labor market programs. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, Vol. 3A. North-Holland, Amsterdam, pp. 1865–2097.

Herrnstein, R., Murray, C., 1994. The Bell Curve: Intelligence and Class Structure in American Life. The Free Press, New York.

Hoxby, C., 1997. How the changing structure of U.S. higher education explains college tuition. National Bureau of Economic Research Working Paper No. 6323, NBER, Cambridge.

Imbens, G., 2000. The role of the propensity score in estimating dose-response functions. Biometrika 87, 706–710.

Lechner, M., 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: Lechner, M., Pfeiffer, F. (Eds.), Econometric Evaluation of Labour Market Policies. Physica, Heidelberg, pp. 43–58.

Light, A., Strayer, W., 2000. Determinants of college completion: school quality or student ability? Journal of Human Resources 35, 299–332.

Neal, D., Johnson, W., 1996. The role of premarket factors in black-white wage differences. Journal of Political Economy 104, 869–895.

Pagan, A., Ullah, A., 1999. Nonparametric Econometrics. Cambridge University Press, Cambridge.

Racine, J.S., Li, Q., 2004. Nonparametric estimation of regression functions with both categorical and continuous data. Journal of Econometrics, forthcoming.

Rosenbaum, P., Rubin, D., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Smith, J., Todd, P., 2004. Does matching overcome LaLonde's critique of nonexperimental methods? Journal of Econometrics, forthcoming.

Tobias, J., 2003. Are returns to schooling concentrated among the most able? A semiparametric analysis of the ability-earnings relationship. Oxford Bulletin of Economics and Statistics 65, 1–30.

Turner, S., 1998. Changes in the returns to college quality. University of Virginia, unpublished manuscript.

US News and World Report, 1991. Directory of colleges and universities. In: 1992 College Guide.