

HOW ROBUST IS THE EVIDENCE ON THE EFFECTS OF COLLEGE QUALITY?
EVIDENCE FROM MATCHING

Dan A. Black
Department of Economics
and Center for Policy Research
Syracuse University
danblack@maxwell.syr.edu

Jeffrey A. Smith
Department of Economics
University of Maryland
and NBER
smith@econ.umd.edu

Version of April 9, 2002

This research was supported in part by the Social Science and Humanities Research Council of Canada. We thank Alex Whalley for excellent research assistance, Shannon Seitz and Alex Whalley for helpful comments, and Barbara Sianesi for providing her matching program.

1. Introduction

The labor market effects of college quality hold great interest for students (and their tuition-paying parents) deciding where to go to college. Estimates of college quality effects could and should play a larger role in the design of government loan and grant programs. They also have relevance to the literature on secondary school quality effects (see, e.g., Hanushek, 2002), to which they present an interesting counterpoint.

A recent literature attempts to estimate the labor market effects of college quality. In particular, the literature focuses on the wage and earnings 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 Brewer, Eide and Ehrenberg (1999), Dale and Krueger (1999), Turner (1998) and Black, Daniel and Smith (2002a,b). The basic finding in this literature is that college quality matters for later labor market outcomes. Going to a higher quality college appears to increase later earnings and wages, by an amount that, on average and, in discounted present value terms, more or less equals its price.¹

The key econometric difficulty in this literature results from the non-random selection of students into colleges of varying qualities. Better students, in terms of test scores, high school quality, parental education, motivation, extracurricular activities in high school or athletic talent, sort into better quality colleges (see, e.g., Hoxby, 1997). With few exceptions (e.g., Dale and Krueger, 1999), the literature relies on a common set of assumptions to identify the college quality effects of interest in the presence of such

¹ Brewer and Ehrenberg (1996) ably survey the earlier literature. Black, Daniel and Smith (2002a,b) summarize the more recent literature for women and men, respectively.

non-random selection. Identification in this literature comes from an assumption of linear “selection on observables”, in the language of Heckman and Robb (1985). Under this assumption, bias resulting from the differential selection of more able, more motivated or otherwise better students into better colleges is (hopefully) taken care of by including pre-determined observable characteristics of the students in a linear outcome model such as an earnings or wage equation.

This paper addresses two (related) potential weaknesses of the linear selection on observables identification strategy employed in this literature. The first potential weakness arises from the fact that the linearity assumption can hide the failure of the so-called “common support” condition. The common support refers to the set of values of the conditioning variables for which there is positive density in all groups being compared. In this context, it means that, for any given set of values on the conditioning variables, we observe persons in each of the college quality levels we seek to compare. The counterfactual outcome is not non-parametrically identified for persons outside the common support. Instead, their counterfactual outcome derives solely from projections based on the linear functional form. Given that theory rarely suggests specific functional forms for outcome equations, failure of the common support condition represents an important limitation, should it occur.

Consider two simple examples. Suppose for the moment that ability is the only conditioning variable. Suppose further that there are only two abilities, high and low, and two levels of college quality, high and low. Figure 1 illustrates the case of perfect sorting by ability. In this extreme case, the common support condition fails for the entire population. There are no low ability students at high quality colleges to provide the

counterfactual for low ability students at low quality colleges. Similarly, there are no high ability students at low quality colleges to provide the counterfactual for the ones at the high quality colleges. The data provide no way to separate out the effects of college quality and ability.

Now consider the less extreme situation in Figure 2. We again assume two levels of college quality, but allow ability to be a continuous variable. The vertical axis represents the outcome Y and the horizontal axis represents ability A . The points represent individual observations in the data. In this example, the common support condition fails only for students with very high ability (region “C” in the figure) and students with very low ability (region “A” in the figure). The effect of college quality on Y is not identified non-parametrically for these students, but is identified for the students with moderate levels of ability (region “B” in the figure). Estimating the linear regression of Y on A using these data would yield coefficient estimates, but the implied counterfactual for persons outside the common support would be the product solely of the linear functional form assumption.

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 potential weakness of the standard approach in the literature on the labor market effects of college quality. Consider again ability as an example. Suppose that we estimate, as most papers in this literature do, an equation of the form

$$Y_i = \mathbf{b}_0 + \mathbf{b}_1 X_{1i} + \dots + \mathbf{b}_k X_{ki} + \mathbf{b}_A A_i + \mathbf{e}_i, \quad (1)$$

where Y_i is the outcome variable of interest (the wage in our case), A_i is a measure of ability (usually a test score), the X_{1i}, \dots, X_{ki} represent other conditioning variables also thought to be related to both the choice of college quality and outcomes, and e_i is a random error term.

Suppose at the same time that the true relationship is given by

$$Y_i = \mathbf{b}_0 + \mathbf{b}_1 X_{1i} + \dots + \mathbf{b}_k X_{ki} + \mathbf{b}_A A_i + \mathbf{b}_{A^2} A_i^2 + u_i. \quad (2)$$

Estimating (1) rather than (2) means that the error term contains the non-linear portion of the ability variable, so that

$$\mathbf{e}_i = u_i + \mathbf{b}_{A^2} [A_i^2 - E(A_i^2 | X_{1i}, \dots, X_{ki}, A_i)]$$

If there is selection on this component of the variation in ability, then estimation of (1) yields biased estimates. This holds even though conditioning on observable variables, as in (2), would solve the selection problem.

Matching methods allow us to address both the support issue and the linear conditioning issue in a convenient way. Matching methods constitute the non-parametric analog to linear regression.² Like linear regression, matching assumes selection on observables, so that, conditional on some set of observable characteristics, participation in the treatment is independent of the untreated outcome. Current matching methods require binary (or multinomial³) 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.

² See, e.g., the discussions in Heckman, Ichimura and Todd (1997,1998), Heckman, Ichimura, Smith and Todd (1998), Heckman, LaLonde and Smith (1999), Dehejia and Wahba (1999,2001) and Smith and Todd (2002). We provide additional econometric details in Sections 5 and 8.

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 s, rather than on the X s themselves.⁴ Once we have the distributions of estimated propensity scores for sample members in high and low quality universities, we can compare the two densities, either visually or more formally, to get a clear sense of the extent of the common support problem in our application.

Matching does directly solve the linear conditioning issue. Matching methods make no functional form assumptions – instead, untreated observations similar to each treated observation in terms of X , or the probability of participation, $P(X)$, serve as counterfactuals. By constructing an observation-specific counterfactual for each treated observation, matching methods avoid potential bias due to misspecification of the functional form in a linear model.

We apply matching methods to estimate the wage effects of college quality using data from the National Longitudinal Survey of Youth (NLSY). The NLSY data include information on college attended, along with a wealth of background variables including an “ability” test score measure, parental characteristics, high school characteristics, and characteristics of the home when the respondent was young. This wealth of covariate information makes the assumption of selection on observables plausible in our context. We combine the NLSY data with measures of college quality drawn from a variety of published sources.

³ See Imbens (2000) and Lechner (2001) for the generalization to multiple treatments.

Using the NLSY data, we examine how students of different abilities, as measured by scores on 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 participation equation form the propensity scores we use to produce our matching estimates. We examine the common support condition using our estimated propensity scores and compare our matching estimates to regression-based estimates of the same parameter.

We reach three important empirical conclusions. First, we quantify the extent of sorting of students by ability, measured by the ASVAB test score, into colleges of different qualities. We find substantial sorting in our representative national sample of students from the NLSY. However, 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 is 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 in Heckman and Vytlačil (2001) in their study of the returns to years of schooling, we find that the sorting is not sufficiently strong to cause the support condition to fail in our sample. However, using our estimated propensity scores, we find that the support condition barely holds for persons with a high probability

⁴ See Rosenbaum and Rubin (1983), who show that the same assumptions that justify matching on X also justify matching on the propensity score.

of attending a high quality (in the upper quartile of the distribution in our sample) college. As a result, our matching estimates (correctly) have large standard errors. The linear functional form assumption makes a difference here, and is making a difference in other analyses of this type in data sets of similar size.

Third, although they are imprecisely estimated, our matching estimates often differ substantially from the corresponding regression estimates. In particular, our matching estimates over the region of the data with reasonable density in both samples substantially exceed the OLS estimates. This suggests that the existing literature may understate the wage effects of college quality as a result of the econometric specification it employs.

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 the identifying assumptions we rely on. Section 6 describes our propensity score estimates and examines the support problem. Section 7 presents standard semi-parametric 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 estimated wage effects we obtain by applying them. Section 10 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 to 21 years old on January 1, 1979. Respondents were first interviewed in 1979 and an attempt has been made to re-interview them every year 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.

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 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 includes an important “ability” measure in the form of scores on the Armed Services Vocational Aptitude Battery (ASVAB), which was administered to over 90 percent of the sample.⁵ 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

⁵ Neal and Johnson (1996) describe the test in detail and discuss the issues of interpretation surrounding it.

high school characteristics. Appendix A presents more details on the construction of the sample and the precise variables used in the estimation.

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.⁶ We only matched data on four-year colleges; roughly one-half of the people in our sample attended a four-year college, and many of the quality variables are not available for two-year colleges.⁷

In this paper, 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 for 1987-1989 freshmen. 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. Descriptive statistics for these variables for the men and women in our sample appear in Table 2.⁸ The table documents substantial variation in all three measures among the colleges attended by NLSY respondents.

⁶ We obtained these variables from the Department of Education’s Integrated Post-secondary Education Data System (IPEDS) for 1990 and the *U.S. News and World Report’s* (1991) Directory of Colleges and Universities.

⁷ Our sample includes persons who went on to graduate study, but the college quality variable refers to the most recent college attended as an undergraduate in all cases.

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.⁹ In particular, we take the first principal component of our three variables as our index of college quality.¹⁰ 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.

4. The Relationship between Ability and College Quality

In this section we examine the relationship between ability, as proxied by the ASVAB test score in the NLSY data, 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; to the best of our knowledge, we present the first such analysis using a representative national sample. 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.

⁸ See Black, Daniel and Smith (2002a,b) for descriptions of the other college quality variables, evidence on the correlation among the various measures of college quality and details on the construction of the college quality index.

⁹ We consider the measurement error interpretation of these variables in detail in Black and Smith (2002).

Tables 3 and 4 present the joint distribution of student ability and college quality in two different ways. In each case, we present the distributions for men and women separately. Table 3 presents the joint density directly in terms of quintiles. In each panel of the table, rows represent quintiles of the college quality distribution and columns represent quintiles of the ability distribution. Each cell of the table contains three numbers, the row percentage, the column percentage and the cell percentage. Thus, in the upper left corner of Table 3 for men, we find that 6.48 percent of the sample is in both the first quintile of the ability distribution and first quintile of the college quality distribution. As there are 25 cells and we are using quintiles, random sorting would yield roughly four percent in each cell, so this cell is substantially over-represented in the data.

Two main findings emerge from Table 3. First, there is substantial sorting based on ability. For both men and women, the fraction of observations on the diagonals and the surrounding bands (persons with a difference of one between the two quintile rankings) exceed what would be expected from random sorting. 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.

Table 4 presents the mean, standard deviation and selected percentiles of the ability distribution for the quartiles of the college quality index. This table includes persons with a missing quality index as a separate group. Most of these persons attended

¹⁰ Similar combinations of other college quality measures are highly correlated with this one. See the discussions in Black, Daniel and Smith (2002a,b).

two-year colleges or very small four-year colleges. Note that we have adjusted our ability measure so that the mean equals zero (and the variance equals one) in the samples used in this table.

With two exceptions, Table 4 makes the same points as Table 3. First, it is clearer, however, in regard to the asymmetry of the sorting. Compare the 10th and 90th percentiles of ability for the lowest and highest quartiles of quality. There is an upper tail of high quality students at the low quality colleges, but very little in the way of a low quality tail at the high quality colleges. Coincident with this, the skew of the ability distribution increases in absolute value as college quality increases, particularly for men. Second, it includes persons with a missing value of the college quality index. Their ability distribution is most similar to that in the lowest quality quartile – a bit worse than that for men. This is not surprising given the types of schools attended by this group.

Taken together, Tables 3 and 4 indicate clear sorting by ability, with more able students attending higher quality colleges. This sorting is asymmetric, with more high ability students at low quality colleges than low ability students at high quality colleges. Finally, at the level of quartiles and quintiles, the sorting by ability alone does not suffice to threaten the validity of the support condition. However, 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 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 the jargon of the literature, we estimate the impact of “treatment on the treated.” In terms of our notation, the parameter of interest is:

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

In a world of heterogeneous impacts, 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 matching methods and standard regression analyses estimate the impact of a “treatment” under the assumption of selection on observables. The selection on

observables assumption is formalized in the literature 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 available set of X variables. Moreover, 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).

The difference between regression estimation and matching is that regression makes the additional assumption that the CIA holds when conditioning only linearly on X . Both assume that conditioning on an available set of covariates removes all systematic differences between high and low quality college attendees in outcomes conditional on attending a low quality college.

Matching, but not regression, requires the so-called “common support” assumption, which can be expressed as:

$$\Pr(D = 1 | X) < 1 \text{ for all } X.$$

The support condition states that, for each X satisfying the CIA, there must be some individuals in both states; in our context, for each X , there must be some individuals who attend both high and low quality colleges. If there are X for which everyone participates, then there is no way in a matching context to construct the counterfactual outcome for these observations.^{12 13}

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 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, $\Pr(D = 1 | X)$, 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 equal proportions. As a result, any differences between subgroups with different X but the same propensity score balance out when constructing the estimates.

¹¹ As noted in Heckman, Ichimura, Smith and Todd (1998), the CIA is stronger than we actually require. In fact, conditional mean independence $E(Y_0 | X, D = 1) = E(Y_0 | X, D = 0)$ suffices to identify our parameter of interest.

¹² When estimating the average treatment effect, the additional condition that $\Pr(D = 1 | X) > 0$ is required in order to estimate the high quality college counterfactual for persons in a low quality college.

6. Propensity Scores and the Common Support Condition

We estimate four sets of propensity scores, two for men and two for women, using the NLSY data. The first set for each group includes age, age squared, race/ethnicity, region of birth dummies and the first two principal components of the ten ASVAB test scores and their squares. The second set for each group includes these variables plus characteristics of the respondent's high school, the respondent's parents and the respondent's home environment as a child.

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. We consider two sets of covariates as a sensitivity analysis and to explore the importance of the extensive set of background characteristics available in the NLSY but not available in other widely used data sets. The correlation between the two sets of propensity scores is 0.926 for men and 0.872 for women.

The only potentially controversial variable included in the 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 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 attended and whether or not the student completes a degree. We adopt the more conservative approach

¹³ Both matching and regression analysis require the assumption of no general equilibrium effects. In the matching literature, this assumption has the name SUTVA, for "stable unit treatment value assumption."

of conditioning on years of schooling in our propensity scores. In the regression estimates presented in Section 7, we do it both ways and show that it makes a big difference to the estimates.

We examine the support condition in terms of the propensity scores in Table 5 and Figure 3. In both cases, we present only the richer specification that includes the parental, home and high school background variables. Similar results emerge with the sparser scores. Table 5 shows the cumulative distribution functions for the treatment (high quality college) and comparison (low quality college) groups for both men and women. Figure 3 presents the same information graphically in the form of the probability density functions for both men and women. In Figure 3, the top histogram in each graph corresponds to the $D = 1$ group while the bottom histogram corresponds to the $D = 0$ group.

Table 5 and Figure 3 illustrate 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 55 percent of the comparison group lies below the 5th percentile of the treatment group, and nearly 55 percent of the treatment group lies above the 95th percentile of the comparison group. This requires us to use the upper tail of the comparison group quite intensively in generating the matching estimates presented in Section 9. Any measurement error or other problems with these “outliers” will pose severe problems for the matching estimator applied to our data. 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.

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, we present evidence from 16 regression models, eight for men and eight for women. The first row of four models for each group does not include the years of schooling variable, the second row of four models does include it. Within each row, we progressively add additional covariates. The variables included in the two models in the lower right correspond to those in the two propensity score specifications defined in Section 6.

The dependent variable is the natural log 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 2nd, 3rd and 4th 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 1st quartile of the quality distribution is the omitted group and, therefore, the implicit counterfactual.

We have four major findings. First, as shown in Black, Daniel and Smith (2002a,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

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 is equal to total

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. 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 per cent for women.

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

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

for the individual counterfactual for treated observation i . In this equation, $j = 1, \dots, J$ indexes the untreated comparison group observations. All matching estimators construct an estimate of the 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.

We use three alternative matching estimators in our empirical work: the nearest neighbor estimator and two variants of the kernel matching estimator. The weighting function of the nearest neighbor matching estimator is given by:

wage income divided by total hours worked. The resulting variable was then logged. Persons with no jobs in 1998 are excluded from the sample.

$$w(\hat{P}(X_i), \hat{P}(X_j)) = \begin{cases} 1 & \text{if } j = \operatorname{argmin}_{k \in \{D=0\}} \{|P(X_i) - P(X_k)|\}; \\ 0 & \text{otherwise.} \end{cases}$$

In words, the nearest neighbor matching estimator assigns a weight of one to the comparison (low quality college) observation with the closest propensity score to each treated observation, and zero to all other comparison observations.¹⁵

In contrast, kernel matching potentially assigns a non-zero weight to several, or even all, comparison observations in constructing the counterfactual for each treated observation. The standard form for the weighting function is given by:

$$w(\hat{P}(X_i), \hat{P}(X_j)) = \frac{K \left[\frac{\hat{P}(X_i) - \hat{P}(X_k)}{a_n} \right]}{\sum_{k \in \{D=0\}} K \left[\frac{\hat{P}(X_i) - \hat{P}(X_k)}{a_n} \right]},$$

where $K(\cdot)$ is a kernel function and a_n is a bandwidth. We employ two different kernels, the normal or Gaussian, where $K(\cdot) = \mathbf{f}(\cdot)$, the standard normal probability density function, and the Epanechnikov kernel, where

$$K(\mathbf{y}) = \begin{cases} \frac{3}{4}(1-\mathbf{y}^2) & \text{if } |\mathbf{y}| < 1; \\ 0 & \text{otherwise.} \end{cases}$$

Asymptotically, all the different matching estimators produce the same estimate, because in an arbitrarily large sample, they all compare only exact matches. In finite

¹⁵ This form of the nearest neighbor estimator assumes matching with replacement, in which each comparison group observation can form the counterfactual for more than one treated observation. In the statistics literature, it is common to match without replacement, so each comparison observation can form the counterfactual for at most one treated observation. The findings in Section 9 make it clear why we choose to match with replacement. For further discussion, see Dehejia and Wahba (1999).

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.

Consider first the tradeoff between bias and variance. The single nearest neighbor estimator minimizes bias, as it chooses only the closest comparison group observation and assigns all the weight to it in constructing the counterfactual. In contrast, kernel estimators will generally assign positive weight to several comparison group observations. This increases the average distance (in propensity score terms) between the treated observation and the observations used to construct the counterfactual, which implies greater bias. At the same time, by using information from additional comparison observations to construct the estimated counterfactual, the variance of the estimate is reduced. The kernel estimator will be preferred in cases where many treated observations have multiple comparison group observations with similar propensity scores. The same tradeoff between bias and variance holds in choosing the bandwidth for a particular kernel estimator.

A second source of differences in finite samples concerns how the support condition is implicitly or explicitly dealt with. The single nearest neighbor estimator and the Gaussian kernel estimator do not impose the support condition at all. Both will construct a counterfactual for every treated observation, no matter how large the distance, in propensity score terms, to the nearest comparison group observation.

To overcome this problem, the nearest neighbor estimator is often combined with a caliper. The caliper defines an interval around each treated unit, say plus or minus 0.05. If no comparison group observation has a propensity score within the interval, the

corresponding treated observation is dropped from the estimation. This, of course, changes the parameter being estimated, as it now becomes the impact of treatment on the treated for treated persons for which there are comparison observations within their caliper. This parameter may differ substantially from the original parameter of interest if large numbers of treated observations fail the caliper condition.

Similarly, the Epanechnikov kernel, and other kernels such as the triangle kernel, assigns zero weight to observations more than a certain distance away from the treated observation whose counterfactual is being estimated. This distance is increasing in the chosen bandwidth. If no comparison observations exist in the region of positive weight, the corresponding treated observation is again omitted from the estimation.

Rather than attempting to resolve all these issues *a priori* by selecting a particular matching estimator with a particular caliper width or bandwidth and presenting estimates only for that estimator, we present estimates based on nearest neighbor matching, and kernel matching using the Gaussian kernel and the Epanechnikov kernel. In each case, we present estimates corresponding to two calipers or bandwidths. The calipers and bandwidths were chosen by inspecting the data, as suggested in Pagan and Ullah (1999, pg. 120), but are similar to those that would be chosen by the rules of thumb in Silverman (1986).¹⁶

¹⁶ Heckman, Ichimura and Todd (1997) and Smith and Todd (2002) discuss the issue of which matching estimator to choose in more detail. Frölich (2000) presents Monte Carlo evidence comparing alternative matching estimators.

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 estimates based on nearest neighbor matching, kernel matching using a Gaussian kernel and kernel matching using the Epanechnikov kernel. In each case, we present two alternative estimates that vary either the caliper width (for the nearest neighbor estimator) or the bandwidth (for the kernel estimator). We also indicate, for each estimate, the number of treated observations for which an estimated counterfactual could be constructed using that particular estimator. As noted in Section 8, dropping observations due to failure of the support condition given a particular choice of weighting function and bandwidth or caliper width changes the nature of the parameter being estimated. Corresponding to each matching estimator, we also present an estimated impact for the region of “thick” common support, which we define 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. Bootstrap standard errors based on 500 replications appear in parentheses below each estimate.

The final row of the table presents 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 1st or 4th quartile of the quality distribution. This sample corresponds is the same one 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.

The table has four columns. The first two columns present estimates for men and the second two present estimates for women. For each group, we present estimates based on propensity scores that do and do not include the conditioning variables representing home, high school and parental characteristics. Comparing the two sets of estimates gives a sense of the importance of these variables in controlling for selection on observables. If the estimates differ, this suggests that they play an important role in making the CIA credible in our application. If the estimates are about the same, then either both sets of variables satisfy the CIA, or there remain unobserved factors, unrelated to the home, parental and high school characteristics, that bias both sets of estimates.

For men, both OLS estimates indicate about a 13 percent increase in wages as the effect of moving from a college in the 1st quartile of the quality distribution to one in the 4th quartile. Both OLS estimates are statistically significant at the five percent level. Now consider the matching estimates based on the full sample. These estimates range from 0.20 to 0.91, suggesting a smaller impact. The estimates are a bit smaller with the full set of conditioning variables in most cases. In general, changing the bandwidth or caliper to make it narrower either reduces the estimates slightly or leaves them unchanged. None of the matching estimates is statistically significant at conventional levels; being non-parametric has a price.

In contrast to the full sample estimates, the estimates that consider only the “thick support” region are almost all higher than the OLS estimate, sometimes substantially so. These estimates range from 0.118 to 0.220. In the dense region, it makes sense to pay more attention to the kernel estimates, as these use more information. With the full set of covariates, these estimates range from 0.161 to 0.208, which suggests either higher

impacts for this subgroup than the full population, or that the regression estimates understate the wage effect of college quality for men.

The estimates tell a similar story for women. Here the OLS estimates indicate a wage effect of 10 to 11 percent associated with attending a high quality college. Both are (just) significant at the five percent level. The full sample matching estimates range from 0.045 to 0.142, and bracket the OLS estimates. Adding additional matching variables increases the nearest neighbor estimates and decreases the kernel estimates. Even more so than for men, changing the bandwidth within the range shown in the table matters little to the estimates. The “thick support” estimates are uniformly higher than the estimates for the full sample, just as we find for men. Adding additional variables in the thick support case increases the kernel estimates and decreases the nearest neighbor estimates. For both men and women, the “thick support” estimates suggest either higher impacts for middle values of the propensity score (something somewhat difficult to reconcile with an economic model of participation) or that the OLS estimates understate the wage impact of attending a higher quality college.

While the estimates, and their associated large standard errors, do not tell a strong story about 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 estimates end up having a high variance. To see why, consider Table 8, which indicates the number of times each comparison group observation was used in the matching estimates for men. The table considers the rich specification of the propensity scores for calipers of width 0.1 and 0.01. Results are similar for women and for other caliper widths.

In both cases shown in Table 8, we can match 175 of 176 treatment group observations. Amazingly, though, just seven observations account for over half the matches when the caliper is set to 0.1. Of course, we can impose a narrower caliper, but this reduces the number of treated observations matched to 131, while lowering the fraction of matches accounted for by the top seven observations by only a few percent, to 45.8 percent. The tighter caliper has only a marginal impact on the relative contribution made by a small number of comparison group observations.

Taken together, the estimates in Table 7, and the frequency of use analysis for the comparison group observations in Table 8, 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 age and other 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 non-parametric inference regarding the wage effects of college quality.

10. 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 three 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. However, sorting on ability alone 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. The difficulty comes for respondents with a high probability of attending a high quality college, as our data include only a handful of comparable individuals who attend low quality colleges. 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 regressions hides the problem by implicitly borrowing strength from comparison observations with lower probabilities of attending a high quality college.

Third, in substantive terms, we feel 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 often differ substantially from the OLS estimates. In the region of thick support, they are higher than the OLS estimates, which

is consistent with underestimation of the wage effects of attending a high quality university.

Bibliography

Black, Dan, Kermit Daniel, and Jeffrey Smith. 2002a. "College Quality and the Wages of Young Men." Unpublished manuscript, University of Maryland.

Black, Dan, Kermit Daniel, and Jeffrey Smith. 2002b. "College Quality and the Wages of Young Women." Unpublished manuscript, University of Maryland.

Black, Dan and Jeffrey Smith. 2002. "What is the Parameter of Interest in the Literature on College Quality?" Unpublished manuscript, University of Maryland.

Bowen, William and Derek Bok. 1998. *The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions*. Princeton, NJ: Princeton University Press.

Brewer, Dominic and Ronald Ehrenberg. 1996. "Does it Pay to Attend an Elite Private College? Evidence from the Senior Class of 1980." In *Research in Labor Economics, Volume 15*, ed. Solomon Polachek, 239-271. Greenwich, CT: JAI Press.

Brewer, Dominic, Eric Eide, and Ronald Ehrenberg. 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(1): 104-123.

Cook, Philip and Robert Frank. 1993. "The Growing Concentration of Top Students at Elite Schools." In *Studies of Supply and Demand in Higher Education*, ed. Charles Clotfelter and Michael Rothschild, 121-140. Chicago: University of Chicago Press for NBER.

Dale, Stacy Berg and Alan Krueger. 1999. "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables." NBER Working Paper No. 7322.

Dehejia, Rajeev and Sadek Wahba. 1999. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluations of Training Programs." *Journal of the American Statistical Association* 94(448): 1053-1062.

Dehejia, Rajeev and Sadek Wahba. 2001. "Propensity Score Matching Methods for Non-Experimental Causal Studies." *Review of Economics and Statistics* 84(1):151-161.

Frölich, Markus. 2000. "Nonparametric Covariate Adjustment: Pair-matching versus Local Polynomial Matching." Discussion paper 2000-17, Department of Economics, Universität St.Gallen.

Hanushek, Eric. 2002. "Publicly Provided Education." In *Handbook of Public Finance*, ed. Alan Auerbach and Martin Feldstein, forthcoming. Amsterdam: North-Holland.

Heckman, James and Peter Siegelman. 1993. "The Urban Institute Audit Studies: Their Methods and Findings" In *Clear and Convincing Evidence: Measurement of Discrimination in America*, ed. Michael Fix and Raymond Struyk, 187-258. Washington, DC: Urban Institute Press.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66(5):1017-1098.

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64(4): 605-654.

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65(2): 261-294.

Heckman, James, Robert LaLonde, and Jeffrey Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs" In *Handbook of Labor Economics, Volume 3A*, ed. Orley Ashenfelter and David Card, 1865-2097. Amsterdam: North-Holland.

Heckman, James and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data*, ed. James Heckman and Burton Singer, 156-246. Cambridge: Cambridge University Press.

Heckman, James, and Edward Vytlacil. 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): 1-12.

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

Hoxby, Caroline. 1997. "How the Changing Structure of U.S. Higher Education Explains College Tuition." NBER Working Paper No. 6323.

Imbens, Guido. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87(3): 706-710.

Lechner, Michael. 2001. "Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumption." In *Econometric Evaluation of Labour Market Policies*, ed. Michael Lechner and Friedhelm Pfeiffer, 43-58. Heidelberg: Physica.

Neal, Derek and William Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences." *Journal of Political Economy* 104(5): 869-895.

Pagan, Adrian and Aman Ullah. 1999. *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Rosenbaum, Paul and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41-55.

Silverman, Robert. 1986. *Density Estimation*. London: Chapman and Hall.

Smith, Jeffrey and Petra Todd. 2002. "Does Matching Overcome LaLonde's Critique of Nonexperimental Methods?" *Journal of Econometrics*, forthcoming.

Turner, Sarah. 1998. "Changes in the Returns to College Quality." Unpublished manuscript, University of Virginia.

U.S. News and World Report. 1991. "Directory of Colleges and Universities." In *1992 College Guide*.

Table 1: NLSY Descriptive Statistics, 1998

Full sample	Men	Women
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	Men	Women
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

Notes: Authors' calculations using NLSY data. The full sample includes all respondents while the representative sample excludes the minority and military over-samples.

Table 2: College Quality Measures, NLSY 1998

Panel A: Men	Mean	25 th percentile	50 th percentile	75 th percentile
Faculty salaries (n=1,312)	\$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	1,030
Panel B: Women	Mean	25 th percentile	50 th percentile	75 th percentile
Faculty salaries (n=1,488)	\$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	1,005

Notes: 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.

Table 3: Bivariate Distribution of Ability and College Quality Measures, NLSY 1998

Panel A: Men

Ability quintiles

Quality index quintiles	First quintile	Second quintile	Third quintile	Fourth quintile	Fifth quintile	Total
First quintile	(32.38) [32.38] 6.48	(21.90) [21.90] 4.38	(16.19) [16.19] 3.24	(14.29) [14.29] 2.86	(15.24) [15.24] 3.05	(100.0) (N=105)
Second quintile	(23.81) [23.81] 4.76	(20.95) [20.95] 4.19	(20.95) [20.95] 4.19	(20.95) [20.95] 4.19	(13.33) [13.33] 2.67	(100.0) (N=105)
Third quintile	(24.76) [24.76] 4.95	(15.24) [15.24] 3.05	(21.90) [21.90] 4.38	(17.14) [17.14] 3.43	(20.95) [20.95] 4.19	(100.0) (N=105)
Fourth quintile	(11.54) [11.43] 2.29	(18.27) [18.10] 3.62	(27.88) [27.62] 5.52	(20.19) [20.00] 4.00	(22.12) [21.90] 4.38	(100.0) (N=104)
Fifth quintile	(7.55) [7.62] 1.52	(23.58) [23.81] 4.76	(13.21) [13.33] 2.67	(27.36) [27.62] 5.52	(28.30) [28.57] 5.71	(100.0) (N=106)
Total	[100.0] [N = 105]	[100.0] [N = 105]	[100.0] [N = 105]	[100.0] [N = 105]	[100.0] [N = 105]	100.0 N = 525

Panel B: Women

Ability quintiles

Quality index quintiles	First quintile	Second quintile	Third quintile	Fourth quintile	Fifth quintile	Total
First quintile	(31.07) [31.07] 6.21	(19.42) [19.42] 3.88	(20.39) [20.39] 4.08	(15.53) [15.53] 3.11	(13.59) [13.59] 2.72	(100.0) (N=103)
Second quintile	(22.22) [21.36] 4.27	(25.25) [24.27] 4.85	(26.26) [25.24] 5.05	(10.10) [9.71] 1.94	(16.16) [15.53] 3.11	(100.0) (N=99)
Third quintile	(25.71) [26.21] 5.24	(19.05) [19.42] 3.88	(20.95) [21.36] 4.27	(19.05) [19.42] 3.88	(15.24) [15.53] 3.11	(100.0) (N=105)
Fourth quintile	(14.85) [14.56] 2.91	(21.78) [21.36] 4.27	(17.82) [17.48] 3.50	(24.75) [24.27] 4.85	(20.790) [20.39] 4.08	(100.0) (N=101)
Fifth quintile	(6.54) [6.80] 1.36	(14.95) [15.53] 3.11	(14.95) [15.53] 3.11	(29.91) [31.07] 6.21	(33.64) [34.95] 6.99	(100.0) (N=107)
Total	[100.0] [N = 103]	[100.0] [N = 103]	[100.0] [N = 103]	[100.0] [N = 103]	[100.0] [N = 103]	100.0 N = 515

Notes: Authors' calculations using 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. The ability measure is the first principal component of the age-adjusted ASVAB scores.

**Table 4: Distribution of Standardized ASVAB Scores by
Quartiles of College Quality Index, NLSY 1998**

Panel A: Men	Missing quality index	1 st quartile of quality index	2 nd quartile of quality index	3 rd quartile of quality index	4 th quartile of quality index
Mean	-0.231	-0.059	0.211	0.327	0.556
10 th percentile	-1.686	-1.287	-1.111	-0.983	-0.374
25 th percentile	-0.930	-0.667	-0.314	-0.105	0.090
50 th percentile	-0.120	0.066	0.360	0.477	0.635
75 th percentile	0.548	0.632	0.838	0.991	1.075
90 th percentile	0.990	1.162	1.148	1.409	1.466
Standard deviation	1.019	0.936	0.864	0.946	0.793
Skewness	-0.294	-0.271	-0.517	-0.815	-0.782
N	794	177	178	179	176
Panel B: Women	Missing quality index	1 st quartile of quality index	2 nd quartile of quality index	3 rd quartile of quality index	4 th quartile of quality index
Mean	-0.196	-0.087	0.118	0.337	0.779
10 th percentile	-1.496	-1.587	-0.965	-0.965	-0.169
25 th percentile	-0.868	-0.789	-0.410	-0.189	0.360
50 th percentile	-0.174	0.069	0.242	0.465	0.853
75 th percentile	0.504	0.634	0.738	0.947	1.192
90 th percentile	1.078	1.075	1.153	1.392	1.788
Standard deviation	0.980	1.011	0.910	0.941	0.779
Skewness	-0.093	-0.487	-0.509	-0.331	-0.422
N	1003	173	175	171	173

Notes: Authors' calculations using NLSY data, *US News and World Report's Directory of Colleges and Universities* data, and IPEDS data. College quality is for last college attended. There are 1,504 observations for men and 1,695 observations for women. The ASVAB scores are adjusted for the age at which the respondent took the test. We further adjust the scores so that the mean in the sample is zero and the variance is one.

Table 5: Distribution of Propensity Score for First and Fourth Quartiles, NLSY 1998

	Men		Women	
	Propensity score of comparison group	Propensity score of treatment group	Propensity score of comparison group	Propensity score of treatment group
5 th percentile	0.012	0.222	0.002	0.251
10 th percentile	0.018	0.320	0.006	0.369
15 th percentile	0.033	0.421	0.015	0.445
20 th percentile	0.044	0.554	0.026	0.590
25 th percentile	0.070	0.620	0.041	0.650
30 th percentile	0.087	0.653	0.054	0.702
35 th percentile	0.106	0.702	0.078	0.740
40 th percentile	0.131	0.733	0.128	0.782
45 th percentile	0.156	0.775	0.144	0.806
50 th percentile	0.193	0.811	0.195	0.864
55 th percentile	0.225	0.845	0.216	0.890
60 th percentile	0.245	0.872	0.265	0.916
65 th percentile	0.289	0.898	0.305	0.934
70 th percentile	0.319	0.921	0.346	0.942
75 th percentile	0.374	0.943	0.400	0.968
80 th percentile	0.475	0.957	0.521	0.976
85 th percentile	0.556	0.975	0.625	0.983
90 th percentile	0.687	0.989	0.671	0.992
95 th percentile	0.768	0.995	0.811	0.996

Notes: Authors' calculations using 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 the both the treatment group and the comparison group for women. The estimated propensity score includes years of schooling, quadratics in the first two principal components of the age-adjusted ASVAB score, race, age, age squared, region of birth indicators and high school, parental, and home characteristics.

Table 6: Returns to College Quality Measures, NLSY 1998

Men

Without years of education

College quality index --Men

	No	Yes	Yes	Yes
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 (0.0501)	0.054 (0.0490)	0.031 (0.0491)	0.026 (0.0499)
Third quartile	0.170 (0.0472)	0.132 (0.0457)	0.095 (0.0459)	0.082 (0.0473)
Fourth quartile	0.280 (0.0480)	0.220 (0.0475)	0.177 (0.0490)	0.158 (0.0492)

With years of education

Second quartile	0.044 (0.0491)	0.033 (0.0486)	0.007 (0.0487)	0.005 (0.0497)
Third quartile	0.107 (0.0464)	0.094 (0.0457)	0.055 (0.456)	0.050 (0.0469)
Fourth quartile	0.192 (0.0469)	0.167 (0.0468)	0.123 (0.0481)	0.116 (0.0492)
Years of education	0.048 (0.0059)	0.038 (0.0063)	0.038 (0.0064)	0.032 (0.0062)

Table 6: Continued
Women

Without years of education	College quality index --Women			
	No	Yes	Yes	Yes
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 (0.0385)	0.127 (0.0377)	0.105 (0.0377)	0.102 (0.0387)
Third quartile	0.135 (0.0401)	0.105 (0.0394)	0.075 (0.0402)	0.065 (0.0406)
Fourth quartile	0.205 (0.0418)	0.149 (0.0416)	0.124 (0.0418)	0.112 (0.0422)
With years of education				
Second quartile	0.115 (0.0370)	0.105 (0.0366)	0.083 (0.0368)	0.082 (0.0378)
Third quartile	0.090 (0.0390)	0.074 (0.0386)	0.043 (0.0394)	0.039 (0.0398)
Fourth quartile	0.136 (0.0410)	0.107 (0.0412)	0.078 (0.0416)	0.074 (0.0421)
Years of education	0.054 (0.0050)	0.048 (0.0050)	0.047 (0.0050)	0.042 (0.0051)

Notes: Authors' calculations using 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 1,632 observations for men and 1,849 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 models also include quadratics in the first two principal components of the age-adjusted ASVAB scores, race, age, age squared, region of birth indicators and, when indicated, years of schooling and/or the high school, parental, and home characteristics.

**Table 7: Returns to College Quality Propensity Score Matching
Estimates: Nearest Neighbor Estimates, Gaussian Kernel Estimates,
and Epanechnikov Kernel Estimates, NLSY 1998**

$\Delta_{41} = Y_{i4} - Y_{i1}$	Men		Women	
	Without home, high school, and parental characteristics	With home, high school, and parental characteristics	Without home, high school, and parental characteristics	With home, high school, and parental characteristics
Nearest neighbor, caliper 0.1	0.041 (0.0804) [n=175]	0.020 (0.1320) [n=172]	0.116 (0.0885) [n=173]	0.142 (0.1432) [n=168]
Thick support region	0.118 [n=35]	0.220 [n=35]	0.246 [n=34]	0.173 [n=34]
Nearest neighbor, caliper 0.05	0.041 (0.0796) [n=175]	0.020 (0.1320) [n=172]	0.116 (0.0799) [n=173]	0.142 (0.1400) [n=168]
Thick support region	0.118 [n=35]	0.220 [n=35]	0.246 [n=35]	0.173 [n=34]
Gaussian kernel, bandwidth 0.1	0.091 (0.0585) [n=157]	0.072 (0.0960) [n=158]	0.111 (0.0651) [n=158]	0.055 (0.0912) [n=145]
Thick support region	0.172 [n=35]	0.167 [n=35]	0.138 [n=34]	0.153 [n=34]
Gaussian kernel, bandwidth 0.05	0.066 (0.0611) [n=157]	0.057 (0.1045) [n=158]	0.103 (0.0657) [n=158]	0.047 (0.0928) [n=145]
Thick support region	0.158 [n=35]	0.161 [n=35]	0.140 [n=34]	0.154 [n=34]
Epanechnikov kernel, bandwidth 0.1	0.063 (0.0604) [n=157]	0.055 (0.1067) [n=158]	0.103 (0.0646) [n=158]	0.045 (0.0968) [n=145]
Thick support region	0.156 [n=35]	0.154 [n=35]	0.136 [n=34]	0.152 [n=34]
Epanechnikov kernel, bandwidth 0.05	0.046 (0.0641) [n=157]	0.051 (0.1052) [n=158]	0.104 (0.0687) [n=158]	0.061 (0.1141) [n=145]
Thick support region	0.149 [n=35]	0.208 [n=35]	0.146 [n=34]	0.156 [n=34]
OLS estimates	0.130 (0.0496)	0.122 (0.0584)	0.100 (0.0496)	0.112 (0.0557)

Notes: Authors' calculations using 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 the both the treatment group and the comparison group for women. The estimated propensity score includes years of schooling, quadratics in the first two principal components of the age-adjusted ASVAB score, race, age, age squared, region of birth indicators and, in the second and fourth columns, high school, parental, and home characteristics. The OLS estimates use only the observations with college quality in the first or fourth quartile.

Table 8: The Use of Observations in Comparison Group

Men's nearest neighbor estimates with caliper 0.1	Frequency
Observations not used	120
Observations used once	29
Observations used twice	10
Observations used three times	5
Observations used four times	1
Observations used five times	3
Observations used nine times	4
Observations used twelve times	1
Observations used thirteen times	1
Observations used thirty one times	1
Fraction of matches accounted for by 7 observations	52.6%

Men's nearest neighbor estimates with caliper 0.01	Frequency
Observations not used	120
Observations used once	29
Observations used twice	10
Observations used three times	6
Observations used four times	1
Observations used five times	2
Observations used six times	1
Observations used seven times	1
Observations used nine times	1
Observations used thirteen times	1
Observations used fifteen times	1
Fraction of matches accounted for by 7 observations	45.8%

Notes: Authors' calculations using 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 and 176 in the treatment group. The estimated propensity score includes years of schooling, quadratics in the first two principal components of the age-adjusted ASVAB score, race, age, age squared, region of birth indicators and high school, parental, and home characteristics.

Figure 1

Ability

		Low	High
College Quality	Low		
	High		

Figure 2

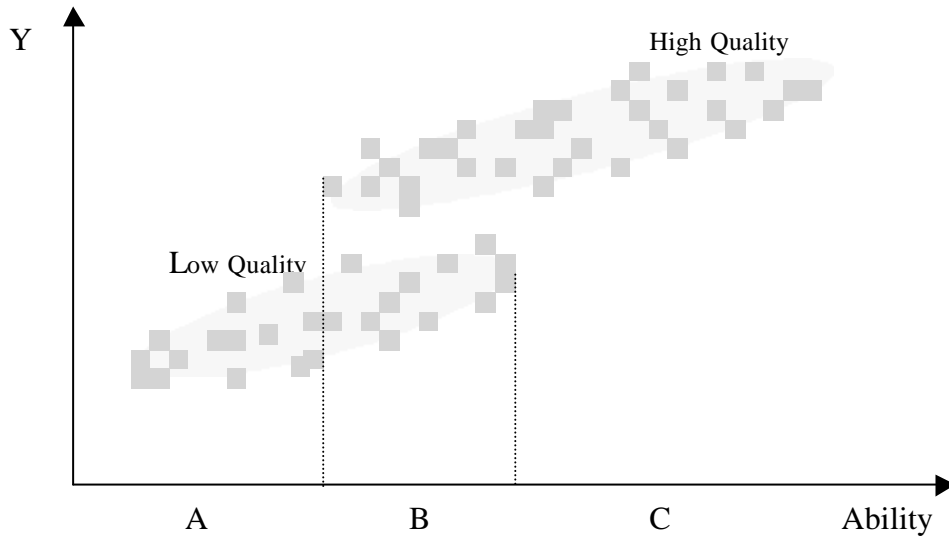
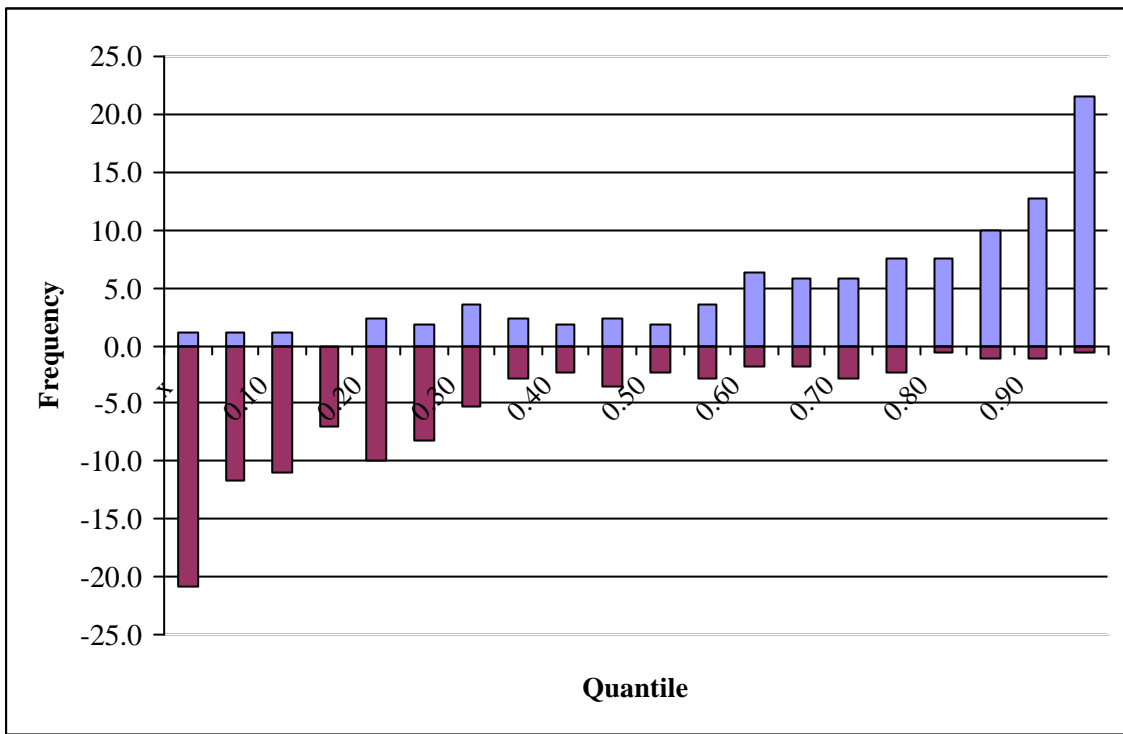
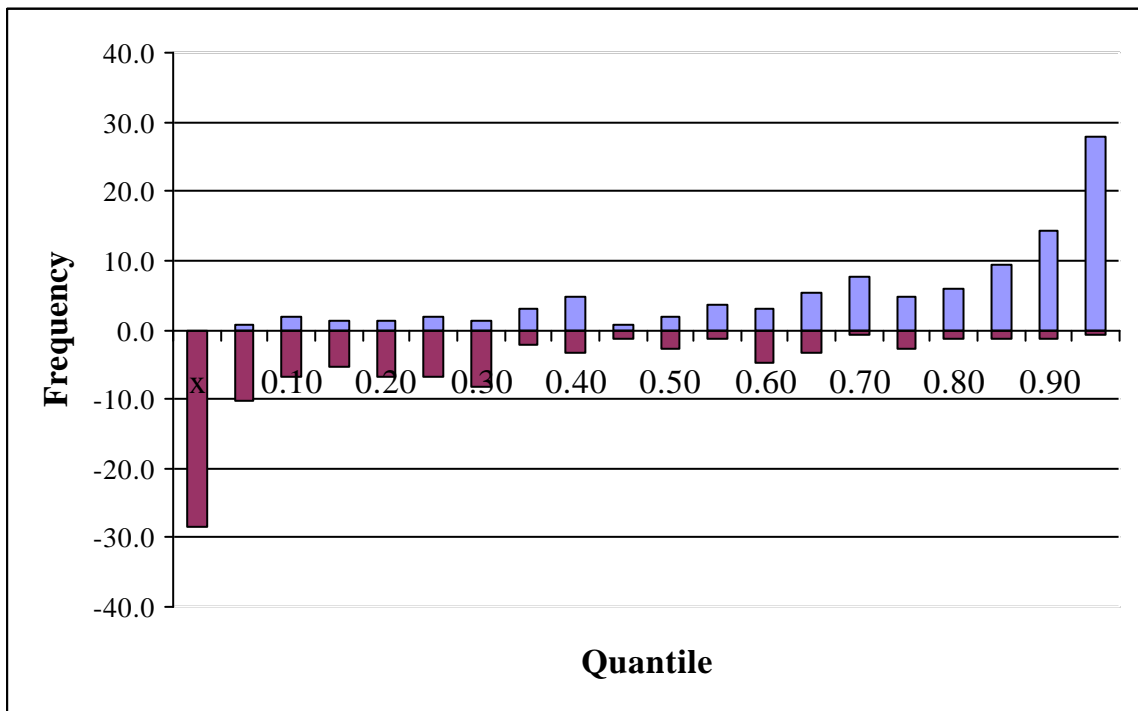


Figure 3

A. Men



B. Women



Source: Authors' calculations using NLSY data.

Appendix A: Detailed Data Description

This appendix describes our data in more detail. To be included in the sample, a respondent must have indicated that they attended a college. Table A1 describes the basic 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. Table A2 reports the results of the principal components analysis used to create the ability controls. Our ability controls were created in two steps. First, we created age-adjusted ASVAB scores by regressing each of the ten ASVAB scores for each individual on age dummy variables and an indicator of whether or not the respondent had completed high school when he or she took the ASVAB. The residuals from these regressions are the age-adjusted scores. These are the data for the principal components analysis. The first two principal components of the age-adjusted scores (and their squares) are the ability variables used throughout the paper.

Table A3 describes the family and other background variables included in some of the regressions and propensity score models. 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.

Table A1: Regressors for Log Wage Regressions

log wage	Log of average real wage (1982 dollars) on all jobs held during the year
region of birth	a vector of 10 dummy variables indicating region in which respondent was born
age	respondent's age at the interview, quartic in age is used
highest grade completed	highest grade or year of school the respondent completed as of the 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)
AFQT score	Armed Forces Qualification Test, based on Armed Services Vocational Aptitude Battery, administered in 1980.

Table A2: Construction of Age-Adjusted Ability Measure

Our ability measures are the first two principal components of ASVAB residuals

Component	Eigenvalue	Difference	Explained Proportion	Cumulative Explained
1	6.276	5.057	0.628	0.628
2	1.219	0.655	0.122	0.750
3	0.564	0.045	0.056	0.806
4	0.519	0.205	0.052	0.858
5	0.315	0.033	0.032	0.889
6	0.281	0.044	0.028	0.917
7	0.237	0.006	0.024	0.941
8	0.231	0.041	0.023	0.964
9	0.190	0.022	0.019	0.983
10	0.167	---	0.017	1.000

Eigenvectors, 1st and 2nd Principal Components

	First Principal Component	Second Principal Component
general science residuals	0.348	-0.140
arithmetic reasoning residuals	0.342	0.034
word knowledge residuals	0.349	0.051
paragraph comprehension residuals	0.325	0.171
numerical operations residuals	0.277	0.482
coding speed residuals	0.248	0.536
auto and shop knowledge residuals	0.291	-0.445
mathematics knowledge residuals	0.322	0.121
mechanical comprehension residuals	0.318	-0.334
electrical information residuals	0.328	-0.322

Note: ASVAB scores are adjusted for age by regressing each test score on age dummy variables and a variable indicating whether the respondent had completed high school when the ASVAB was administered. Principal components analysis is performed on the OLS residuals from these regressions.

Table A3: Family Background & High School Controls

Home: magazine	“When you were about 14 years old, did you or anyone else living with you get magazines regularly?”
Home: newspaper	“When you were about 14 years old, did you or anyone else living with you get a newspaper regularly?”
Home: library card	“When you were about 14 years old, did you or anyone else living with you have a library card?”
Parents: mom education	Highest grade or year of school completed by respondent’s mother.
Parents: mom living	Was the respondent’s mother living at the 1979 interview (when respondents were between 14 and 22 years old)?
Parents: mom age	At the 1987 interview.
Parents: dad education	Highest grade or year of school completed by respondent’s father.
Parents: dad living	Was the respondent’s father living at the 1979 interview?
Parents: dad age	At the 1987 interview.
Parents: living together	Indicator for whether the respondent’s mother and father lived in the same household at the 1979 interview.
Parents: mom occupation	Occupation of job held longest by mother or stepmother in 1978, represented by dummy variables for each Census 1-digit occupation.
Parents: dad occupation	Occupation of job held longest by father or stepfather in 1978, represented by dummy variables for each Census 1-digit occupation.
HS: Size	Asked of respondents’ high schools: “As of 10/1/79 [or nearest date] what was [your] total enrollment?”
HS: 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]

HS: 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]

HS: 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]