

DETECTING SELECTION BIAS, USING PROPENSITY SCORE MATCHING, AND ESTIMATING TREATMENT EFFECTS: An Application to the Private Returns to a Master's Degree

Marvin A. Titus^{*,**}

.....

Most research in the area of higher education is plagued by the problem of endogeneity or self-selection bias. Unlike ordinary least squares (OLS) regression, propensity score matching addresses the issue of self-selection bias and allows for a decomposition of treatment effects on outcomes. Using panel data from a national survey of bachelor's degree recipients, this approach is illustrated via an analysis of the effect of receiving a master's degree, in various program areas, on wage earning outcomes. The results of this study reveal that substantial self-selection bias is undetected when using OLS regression techniques. This article also shows that, unlike OLS regression, propensity score matching allows for estimates of the average treatment effect, average treatment on the treated effect, and the average treatment on the untreated effect on student outcomes such as wage earnings.

.....

KEY WORDS: selection bias; propensity score; matching; treatment effects.

INTRODUCTION

Like research in the social sciences, studies in the area of higher education are plagued by the problem of non-random assignment problems or selection bias. This is a serious problem that exists for at least two reasons. First, most research in the area of higher education does not address non-random assignment problems by employing the use of

*Department of Adult and Higher Education, North Carolina State University, Raleigh, NC USA.

**Address correspondence to: Marvin A. Titus, Department of Adult and Higher Education, North Carolina State University, 300M Poe Hall, Raleigh, NC 27695-7801, USA. E-mail: marvin_titus@ncsu.edu.

experimental designs or randomized trials, the “gold standard” of quantitative research methods. These methods are being encouraged by the newly established Institute of Education Sciences, the research arm of the U.S. Department of Education (Glenn, 2005; Whitehurst, 2002). Because of logistical, ethical, political, and economic reasons, randomized trials may not be feasible in most social science and educational studies, including higher education research.

Second, largely because of the first reason, most higher education researchers conduct studies which use “after the fact” or *ex post facto* data. Although they may encompass a wealth of variables, *ex post facto* data introduce serious challenges. These challenges include the need to carefully address possible threats to validity such as selection bias. Because students non-randomly select to receive financial aid, live in dorms on campus, or become involved campus activities, studies addressing the effect of financial aid on college enrollment, residential life on student development, student involvement on persistence, and college completion on wage earnings may be seriously flawed if selection bias is not appropriately taken into account.

Although selection bias has been addressed extensively in economics (e.g., Amemiya, 1985; Garen, 1984; Heckman, 1979; Maddala, 1983; McMillen, 1995; Olsen, 1980; Willis and Rosen, 1979), it has been not been adequately addressed in the higher education literature, according to several higher education researchers (e.g., DesJardins, McCall, Ahlburg, and Moye, 2002; Porter, 2006; Thomas and Perna, 2004). This study addresses the problem of selection bias in higher education, demonstrates the use of propensity score matching, and estimates treatment effects, which are applied to examining the private returns to receiving a master’s degree by program area.

The Problem of Selection Bias

DesJardins and associates (2002) contend that in order to achieve more precision with respect to the impact of college on students, research in higher education needs to address the issue of selection bias. Using a sample consisting only of students who applied for financial aid, a study of the effects of college loans on enrollment may suffer from imprecise results, due to sample-selection bias. In addition, if the same study does not take into account students’ selecting to take loans, the “treatment” effects of loans on enrollment may also be estimated with a substantial amount of imprecision, due to self-selection bias.

As illustrated in the example above, selection bias may originate from sample-selection or self-selection (endogeneity) problems. Sample-selection

bias involves the non-random selection of certain individuals, based on the availability of observable data, such as wage income and college completion. Using an analytic sample restricted to individuals who reported wage earnings, a study of the effect of college completion on wage earnings may produce biased rates of return to a college education. This is because such a sample is restricted to include only individuals who are employed. To properly address this type of bias, a researcher may have to conduct either of the two following procedures. Prior to estimating the earnings effect of completing a college degree, a researcher would estimate a model which predicts individual decisions to participate in the labor market (Maddala, 1983). The researcher would split the sample into sub-samples of college graduates and non-graduates and then derive separate estimates for each sub-sample (Idson and Feaster, 1990; Main and Reilly, 1993).

Endogeneity or self-selection bias occurs when predictors of an outcome are themselves associated with other unobserved or observed variables. Continuing with the example above, if whether or not an individual earned a bachelor's degree is used as primary predictor of wage income among employed individuals and the non-random nature of completing a bachelor's degree is ignored, biased estimates of the college wage premium may result. Typically, in past research (e.g., Heckman, 1974, 1979), this particular type of bias was corrected by employing a two-step method similar to that mentioned above. In first step, an analyst would estimate the individual's probability of completing a college degree. In the second step, a model is estimated to determine the college degree earnings premium, taking into account the probability of an individual completing a bachelor's degree (Amemiya, 1985).

Some studies may be plagued by both sample-selection and endogeneity bias. For example, an analysis of the wage earnings effect of completing a bachelor's degree among women may produce biased estimates of the returns to college completion for two reasons. First, the dependent variable, wage earnings of women, is only observed among those women who choose to enter the labor market, resulting in a non-random restricted sample. Consequently, the observed wages for women in the non-random restricted sample may be higher than the true wage in the full sample, thereby underestimating the college wage gap among women. Second, the independent variable of interest, whether or not a woman completed a bachelor's degree, is a *choice* variable, dependent on a variety of factors related to the completion of a bachelor's degree. Ignoring these factors may lead to imprecise coefficient estimates and more importantly, model misspecification.

To address the twin problem of sample-selection and endogeneity bias, econometricians typically employ the use of a double-selection method that would first involve estimating a selection model (e.g., probit regression model) for labor force participation and a second selection model explaining bachelor's degree completion status (Amemiya, 1985). The appropriate terms from the selection models are then included in two earnings models estimated via ordinary least squares (OLS) regression techniques. The first earnings model would use data from a sub-sample of bachelor's degree recipients while the second earnings model would utilize data from a sub-sample sample of non-bachelor's degree recipients. Statistical tests (i.e. Chow tests) would then be employed to test if the predictors of earnings differ by OLS regression model.

Because of their strong distributional assumptions (discussed in more detail below), the above techniques are rather limited in their approach to address the problem of sample-selection or self-selection (endogeneity) bias. Using propensity score matching, this study addresses the problem of self-selection bias in higher education research. The use of propensity score matching is used to estimate the treatment effects of receiving a master's degree on wage earnings. This study demonstrates the use of propensity score matching techniques to adjust for self-selection bias by applying these techniques to an examination of the private returns to receiving a master's degree by program area.

THEORETICAL PERSPECTIVE

The concept of private returns to a college degree, including a master's degree, is drawn from human capital theory (Becker, 1993), which posits that individual earned income is largely a function of labor productivity, derived from individual investments in education and training. Drawing on concepts from human capital theory, most econometric models examine the private benefits of higher education by assuming that an individual maximizes her or his college-going behavior after comparing the monetary as well as non-monetary costs to expected benefits associated with completing college (e.g., Fuller, Manski, and Wise, 1982; Manski and Wise, 1983; Schwartz, 1985). In conventional econometric models, non-economic information about education does not play a role in individual higher education investment decisions (e.g., Mincer, 1974; Willis, 1986; Willis and Rosen, 1979). Manski (1993) and others (Paulsen and St. John, 2002; Perna, 2000; St. John and Asker, 2001) suggest that the explanatory power of econometric models for determining college attendance is improved when including

such non-monetary concepts as values about education and access to college-related information. In an effort to further understand the forces that influence college enrollment decisions, prior research (e.g., Perna, 2000, 2004) utilized expanded econometric models that incorporate concepts from cultural and social capital theories in addition to human capital theory.

With underpinnings in sociology, cultural capital theories (Bourdieu, 1986; Bourdieu and Passeron, 1977) posit that class-based preferences, tastes, values or “habitus” are derived from parents and others while social capital theories (Coleman, 1988; Lin, 2001; Portes, 1988) hypothesize that social networks and institutions provide access to information. Perna (2004) demonstrated that measures of social capital help to further explain individual graduate school-going behavior among bachelor’s degree recipients. Based on prior research, this study uses variables reflecting measures of cultural and social, financial capital, and human capital to help discern the relationship between wage income and educational attainment at the master’s degree level by area of study.

RESEARCH DESIGN

Panel data from a nationally representative sample is used to address sequentially the following research questions:

1. What variables explain the chance of receiving a master’s degree by area of study?
2. After taking into account the chance of receiving a master’s degree by area of study what are the private returns to a master’s degree by area of study?

Sample

This study draws on data from the second follow-up (1997) to the 1993 Baccalaureate and Beyond (B&B:93/97) longitudinal survey, a restricted national database sponsored by National Center for Education Statistics (NCES). The B&B:93/97 study is based on a stratified two-stage sample design with postsecondary institutions stratified by the Carnegie Foundation’s classification system type and control (i.e. private versus public) as the first-stage sample unit and students within schools as the second stage sample unit (National Center for Education Statistics, 1999). This sample design represents all postsecondary students in the United States who completed a bachelor’s degree in the

1992–93 academic year (AY 93) and was a sub-sample of the students selected from the 1993 National Postsecondary Student Aid Study (NPSAS:93) survey.

The analytic sample used in this study is limited to 3948 cases that have complete data for the dependent variable, the annual salary of individuals who were employed full-time in 1997 with incomes between \$1300 and \$100,000, and the independent variables. The decision to restrict the analytic sample to individuals who were employed full-time helps to control for variables which may influence labor market entry and level of participation. This research takes into account the stratified and clustered nature of the B&B:93/97 sample design through the use of specific procedures in the statistical software, Stata (Eltinge and Sribney, 1996; Levy and Lemeshow, 1999; McDowell and Pitblado, 2002), and the panel weight (BNBPANEL) provided by NCES.

Analytical Framework

Researchers, particularly economists, have made an effort to address endogeneity bias or self-selection bias by using two procedures, two-step regression techniques and instrument variable (IV) estimation. Introduced by Heckman (1974) and further developed by Heckman (1979) and Lee (1978), two-step regression techniques, also known as endogenous switching models, involve the use of the “Heckman’s lambda” or the Inverse Mill’s Ratio (IMR). The IMR is typically calculated from the residuals or unobservable variables of a probit selection model. The IMR, which allows for an assessment of possible bias from “selection on the unobservables” is then included in the substantive model and calibrated, using OLS regression techniques. When using the Heckman technique, due to correlation between errors in the substantive (OLS regression) model and errors in the selection (probit regression) model, the OLS regression model may produce biased and inconsistent estimates of parameters other than the IMR. Consequently, econometricians (e.g., Maddala, 1998) have also employed maximum likelihood (ML) regression techniques, which involve the simultaneous estimation of a selection and a substantive model to correct for correlated errors across equations (Greene, 2000; Pindyck and Rubinfeld, 1998). But according to Greene (2000), the use of simultaneous ML regression or two-step Heckman procedures for estimating the effect of bias from selection on “unobservables” are limited by the assumptions of the distribution of the errors and the linear relationship between earnings outcomes and predictors of earnings.¹ More specifically, estimates from

the ML regression or two-step Heckman procedures are completely dependent on the assumption that unobserved variables are normally distributed.

Like the two-step Heckman technique, IV estimation remains one of the most widely used methods for addressing problems with self-selection or endogeneity bias. The IV method is most appropriate when a predictor variable (e.g., distance from the closet college) in a discrete choice model can be identified and is related to treatment (e.g., attending college) but not to an outcome (e.g., wage earnings). Instrumental variable estimation techniques involve the use of a variable or instrument that is highly correlated with treatment but not correlated with unobservable factors. Like the Heckman two-step method, the IV method controls for self-selection on unobservables in substantive models, which generally are calibrated using OLS regression. A major limitation of the IV approach is that it requires at least one predictor of treatment that does not determine the outcome (Blundell and Costa Dias, 2000; Heckman, 1997). Some researchers (e.g., Carneiro and Heckman, 2002) contend that in the economics of education literature, most instruments in IV estimation models are correlated to unobservables and consequently are invalid, thereby producing inconsistent estimators of the return to education (Heckman and Li, 2004).

Using a counterfactual framework, following Rosenbaum and Rubin (1983), and drawing from recent advances in econometric (e.g., Conniffe, Gash and O'Connell, 2000; Heckman, Ichimura, and Todd, 1997) and biometric methods (Rosenbaum, 2002; Rubin and Thomas, 1996, Rubin and Thomas, 2000), this research employs the use of propensity score matching techniques, thereby addressing the problem of the limited distributional assumption of the errors inherent in the endogeneous switching and IV estimation models.

The counterfactual framework originated with the work of natural scientists (e.g., Cochran and Cox, 1950; Fisher, 1935; Kempthorne, 1952; Neyman, 1923, 1935), was introduced to the social sciences by economists Roy (1951) and Quandt (1972), and was further developed by Rubin (1974, 1977, 1981) in scholarly articles directed towards educational researchers. A counterfactual framework allows for a decomposition of the treatment effect on outcomes. According to Winship and Morgan (1999), a counterfactual framework enables the detection of "two important sources of bias in the estimation of treatment effects: (a) initial differences between the treatment and control groups in the absence of treatment, and (b) the difference between the two groups in the *potential* effect of the treatment" (p. 703). An example of the former is the difference in wage earnings between college graduates (the

treatment group) *if* they had not attended college and high school graduates who did not graduate from college (the control group). An example of the latter is the difference in wage earnings between high school graduates with a high *probability* of completing college and high school graduates with a low *probability* of completing college. A counterfactual is the wage earnings of high school graduates who did complete college *if* those same high school graduates had not completed college.

Recent extensions by Rubin (1986, 1990, 1991) of the counterfactual framework to observational data suggest that causality can be inferred by positing that even though individuals can only be observed in one state, individuals have *potential* outcomes in both the treatment and untreated state. Despite the development of this analytical framework for conceptualizing statistical inference for treatment effects, most higher education researchers use OLS regression models to identify the average treatment effect (ATE) for an entire sample or population, which includes the treated and the untreated. This approach implicitly assumes that individuals are randomly chosen for treatment, such as receiving a college degree. OLS regression also assumes that treatment affects all individuals in a similar fashion. But according to Heckman and associates (1997), using OLS regression analysis to examine the ATE among all individuals in a sample may be “comparing the incomparables” (p. 647). Several researchers (e.g., Heckman and Robb, 1985; Heckman et al., 1997) contend that rather than an ATE for an entire sample, the effect of average treatment on individuals who select treatment, what is known as the average treatment on the treated (ATT) effect, may be more of an interest to researchers and program evaluators who would like to discern the effectiveness of a program. A conceptual example of the ATT effect would be the average difference in wage earnings between individuals who completed a job training program at a community college and the same individuals who did not complete the program. Because the same individual cannot be observed both completing the job training program and not completing the job training program, the actual basis of the comparison is the difference in wage income between individuals who did complete the program and individuals who did not complete the program but had a similar *high probability* of completing the program.

Using a counterfactual framework, a policy-relevant approach to evaluating outcomes may also include examining the effect of treatment on individuals who do not select treatment or the average effect of treatment on the untreated (ATU). Using the example above, the ATU effect would be, among those who did not complete a job training program, the wage outcome *if* they had completed the job training program.

This study uses a counterfactual framework that involves the utilization of selection model generated propensity scores and matching on propensity scores to produce estimates of the average treatment effect, treated on the treated effect, and treated on the untreated effect of various treatment options. These options include receiving a master's degree in education, or business/management, or in another ("all other") area, or no master's degree.

Further discussion on using selection models to generate propensity scores and matching on propensity scores is necessary. The propensity score is typically generated from a selection model (logit or probit regression) using predictors of treatment, which suggests that treatment is determined via "selection on observables". Because of the categorical nature of the dependent variable, initially a multinomial logit model was used in this research. Based on several statistical tests (Hausman and Small-Hsiao tests), the estimated multinomial logit model did not meet the Irrelevance of Independent Alternatives (IIA) assumption of multinomial models (i.e. odds-ratio between two given alternatives does not change due to a change in the total number of alternatives). Consequently, this study uses a multinomial probit model, which does not suffer from the IIA problem (Daganzo, 1979; Hausman and Wise, 1978, Louviere, Hensher, and Swait, 2000; Maddala, 1983).²

Following the notation by Davidson and MacKinnon (2004), in the multinomial probit model, $J + 1$ possible outcomes are derived from the following latent model,

$$y_{ij}^{\circ} = W_{ij}\beta^j + u_{ij}, \quad u_t \sim N(0, \Omega), \quad (1)$$

where y_{ij}° is an unobserved discrete variable, W is a matrix of covariates, β is a vector of coefficients, u_t is a vector of errors, Ω is a positive definite matrix (e.g., the column and row vectors are linearly independent).³ The subscript t is an observation of a sample, j is a specific outcome, and i denotes the probability assigned to each discrete variable. The observed variables are y_{ij} and are defined as:

$$\begin{aligned} y_{ij} &= 1 \text{ if } y_{ij}^{\circ} - y_{ii}^{\circ} \geq 0 \quad \text{for all } i = 0, \dots, J + 1, \\ y_{ij} &= 0 \text{ otherwise.} \end{aligned} \quad (2)$$

When J is small, the multinomial probit model can be estimated without too much difficulty. In this study, because J is constrained to 4 (a master's degree in education, a master's degree in business/management, a master's degree in "all others", and no master's degree) computational problems associated with a large J are avoided (Davidson and MacKinnon, 2004).

The multinomial probit selection model estimates the probability receiving a master's degree in one of three categories, a master's degree in education, a master's degree in business/management, a master's degree in "all others". Each of these categories is relative to the baseline category, not receiving a master's degree.

The multinomial probit model is utilized to produce estimated probabilities or propensity scores, which are used to match individuals in the analytic sample. For example, if an individual who did not receive a master's degree in business/management (no treatment) has a propensity score of 0.30 (30%), that individual is "matched" with an individual who did receive a master's degree in business/management (treatment) and has the same propensity score. All "matched" individuals who did not receive a master's degree are included in a control group and all "matched" individuals who did receive a master's degree are included in the treatment group. Matching by propensity scores assumes that all differences between individuals affecting treatment and outcome can be captured by observable pre-treatment characteristics. This is known as the conditional independence (CI) assumption. Theoretically, if the CI assumption holds, when choosing between treatment and non-treatment, individuals will base their decision on *expected* rather than the *actual* outcome of the treatment option (Rubin, 1977; Rosenbaum and Rubin, 1983).⁴ In other words, the CI assumption implies that individuals expect their outcome to equal the *average outcome* for individuals with similar observed pre-treatment characteristics.

Rosenbaum and Rubin (1983) demonstrate that if the CI assumption holds for an individual's pre-treatment characteristics it also holds for a *function* of an individual's predicted probability to receive treatment. In more general terms, when a data set does not contain all variables influencing both treatment and the outcome, the CI assumption does not hold, since the effects of treatment are partially accounted for variables not available or unobservable to the researcher. For the purposes of this initial study, it is assumed that the CI assumption holds and the probability of receiving treatment, known as the propensity score, is defined as follows:

$$p(X) = Pr\{T = 1|X\} = E\{T|X\}, \quad (3)$$

where $T = 0, 1$ is the indicator of treatment, X is a vector of pre-treatment characteristics, and E is the mathematical expectations operator which refers to expectations in the overall population of individuals, conditional on X .

In this study, propensity score matching is employed for several reasons. First, according to other researchers (Heckman et al., 1997), when

comparing outcomes among individuals who are not comparable, the use of propensity score matching addresses the bias that may result in the estimate of an effect of treatment on those outcomes.

Second, propensity score matching assumes that, conditional on observable pre-treatment characteristics, outcomes are independent of selecting treatment. In other words, matching, given observable pre-treatment characteristics, *approximates randomization* by “balancing” the “observables” and determines an appropriate control group (Becker and Ichino, 2002).

Third, unlike parametric techniques, propensity score matching requires no assumption about the functional form of the relationship between outcomes and predictors of outcome. When using parametric techniques such as multiple regression, most often a linear (e.g. OLS) and sometimes nonlinear (e.g., logistic) functional form of the relationship between the outcome and covariates is assumed. This may be particularly problematic if in regression models, the covariate distributions differ substantially between groups. Such differences may result in biased estimates of the outcome. Consequently, in an effort to address problems with selection bias, in the past, propensity score matching has been widely utilized in epidemiology and related fields and recently employed in labor economics (e.g., Bellio and Gori, 2003; Doiron, 2004; Hagen, 2002; Lechner, 2002) but also growing in such areas as program evaluation (e.g., Agodini and Dynarski, 2004). Additionally, using data from the B&B:93/97 survey and propensity score matching, Saiz and Zoido (2005) recently investigated, among U.S. college graduates, the relationship between bilingualism and earnings.

Fourth, propensity score matching ensures that only data on variables associated with an individual in both the treatment and control groups are used in the estimation of the treatment effect. By eliminating outliers, more precise estimates of the treatment effects can be obtained. According to other researchers (Dehejia and Wahba, 1999; Hill, Reiter, and Zanutto, 2004) propensity score matching estimates are more consistent with estimates derived from experimental design.

Finally, compared to matching on multidimensional pre-treatment variables, matching on propensity scores results in greater efficiency and is less demanding computationally when determining a comparable control group (Becker and Ichino, 2002). According to Rosenbaum and Rubin (1983), increasing the number of variables related to both the treatment and the control group decreases the chances of an exact match. Thus, rather than relying on a large number of variables, an analyst can make use of a single propensity score to match an individual who was not treated to an individual who was treated.

Because there is no consensus in the literature on the best matching method to adopt, this study utilizes several techniques. Using the Stata module, PSMATCH2, which is provided by Leuven and Sianesi (2004), this research matches cases on propensity scores generated by a previously identified selection model and estimates the ATE, ATT effect, and ATU effect via nearest neighbor, kernel, and local linear regression matching techniques. Nearest neighbor matching involves matching individuals without treatment with individuals who are treated and who have the same estimated propensity score.⁵ More explicitly, in a randomly ordered dataset, a treated case is selected and matched with the “closest” untreated case, based on the same or similar propensity score. The advantage of using nearest neighbor matching is that it is straightforward while the disadvantage is that it may result in few matches and matched cases with large differences in propensity scores.

The kernel algorithm involves matching cases with the propensity scores of cases in an untreated group to a weighted average of cases in a treated group, with weights that are inversely proportional to the distance between the propensity scores of treated individuals and untreated individuals.⁶ Or in other words, cases with comparable propensity scores receive large weights while cases with differing propensity scores receive small weights, resulting in “smoothed” weighted matching estimators, also known as bandwidths.⁷ Unlike nearest neighbor matching, kernel matching uses the propensity scores of all cases in order to identify the best counterfactual and matches several untreated cases with treated cases. According to Frölich (2004), among the different matching algorithms, kernel matching produces the most precise estimates. Because of its features, kernel matching is used to compare estimates of treatment effects to parameter estimates generated by the OLS regression models in the analyses.

Similar to kernel matching, local linear regression matching involves the matching of untreated cases, based on weighted averages of treated cases. Local linear matching utilizes the weighted average of all treatment outcomes, with weights derived from the treatment outcomes within a certain range of the propensity scores.⁸ The precision of local linear regression matching depends on the number of cases in the treatment group compared to the sample size of the untreated comparison group. [For a detailed discussion of the distinction between kernel and local linear regression methods, see Fan and Gijbels (1996).]

All of the matching techniques described above help to ensure that the distribution of the pre-treatment characteristics of individuals in the treatment and the control groups overlap or there is *common support*, thereby making the groups more comparable and casual inferences more

valid. The average difference in outcomes between the treated group and control group in the *common support* area yields the average treatment effect. Unlike regression model parameters estimates of the average treatment effect, propensity score matching estimates of the average treatment effect take into account observations within the area of *common support*.

Using the PSMATCH module in Stata and each matching technique described above, this study will estimate the average treatment effect (ATE), average treatment on the treated (ATT) effect, and the average treatment on the untreated (ATU) wage earnings effect of receiving a master's degree by the major program areas previously mentioned.

Each estimate of the ATE effect will be compared to the respective estimate of the ATT and ATU effect. The estimates of propensity score matching estimates of ATE, ATT effect, and the ATU effect will also be compared to OLS regression parameter estimates. The difference between the OLS regression parameter estimates and the estimates of the ATE generated by the kernel matching technique will be used to help uncover any OLS estimation bias. The OLS regression parameter estimates of the ATE and the estimates of ATT effect, generated from the kernel matching technique, will be compared to help detect self-selection bias.

Variables

This study uses two substantive OLS regression models. In both models, the dependent variable is continuous and defined as the natural log of annual wage earnings. The use of the natural log of annual earnings in the substantive earnings models allows for the interpretation of the unstandardized regression coefficient as a percentage change in earnings that is related to one-unit change in each independent variable. In the first OLS regression model, the independent variables include master's degree by area of study plus the independent variables that are used in the multinomial probit model mentioned below. The second OLS regression model includes the same variables that were included in the first OLS regression model plus variables reflecting labor market experiences such as whether an individual is employed in a job that requires a college degree and in a job that has potential for career advancement, as well as job tenure, labor force experience, and the average hours worked per week. Job tenure and labor force experience reflect variables in a Mincerian (Mincer, 1974) wage equation. Job tenure is measured by number of months employed at the most recent job and the number of months employed at the most recent job squared. Labor force experience includes the number of months employed and the number of

months employed squared (Mincer, 1974). Table 1 presents the variables used in this study.

In the multinomial probit model, the dependent variable has four categories: received a master's degree in education; received a master's degree in business/management; received a master's degree in another category (i.e., "all other"); and did not receive a master's degree, the reference category. In this study, the "all other" category is used for two reasons. First, as of 1997, few individuals in the analytic sample received masters' degrees in each of the areas (e.g., arts and humanities, social behavioral sciences, physical and life sciences, etc.) other than education or business/management. Second, when using multinomial probit regression techniques, which is described above, increasing the number of categories beyond four creates computation problems.

Because problems in finding a match may arise when using continuous variables, all independent variables are categorical (Heckman et al., 1997). The independent variables include measures of the following: gender; race/ethnicity; academic capital, financial capital, and cultural and social capital. Four racial/ethnic groups are included in the analyses: African American, Hispanic, Asian, and White. White is the reference group.

Academic capital is reflected by undergraduate grade point average (GPA) and undergraduate major. Undergraduate GPA is categorized into four quartiles, with the lowest quartile as the reference group. Undergraduate major is represented by a series of dichotomous variables reflecting the following areas: education, business; engineering; public administration or social services; math, physical or life science; social science or psychology; history or humanities, other, and health science. Health science is the reference group.

Financial capital includes parents' income and amount borrowed for undergraduate education. Parents' income is organized into four quartiles, with the highest quartile as the reference group. The amount borrowed for undergraduate education is organized into the following categories: borrowed less than \$5,000; borrowed \$5,000 to \$10,000; borrowed \$10,000 to \$20,000; borrowed more than \$20,000; and did not borrow. Did not borrow is the reference category.

Following the example of Perna (2004), cultural and social capital is reflected in parents' educational background, whether or not English is the first language spoken at home, and the institutional mission of the undergraduate institution attended. Parental educational background is measured by highest level of educational attainment and is reflected in the following categories: less than high school; no more than high school; some college; a bachelor's degree; and a master's degree or higher. A master's degree or higher is the reference category. The institutional

TABLE 1. Summary Statistics of Variables Using Weighted 1992/93 Bachelor's Degree Recipients Data

	Mean	SD	Minimum	Maximum
Annual Salary in 1997 (log)	10.34	0.009	7.17	11.51
Education	0.02	0.003	0	1
Business	0.01	0.002	0	1
“All Others”	0.08	0.005	0	1
Female	0.52	0.010	0	1
Male	0.48	0.010	0	1
African American	0.06	0.005	0	1
Asian	0.04	0.005	0	1
Hispanic	0.05	0.004	0	1
White	0.85	0.008	0	1
Undergraduate GPA—1st quartile	0.25	0.009	0	1
Undergraduate GPA—2nd quartile	0.25	0.008	0	1
Undergraduate GPA—3rd quartile	0.25	0.008	0	1
Undergraduate GPA—4th quartile	0.25	0.009	0	1
Undergraduate Major				
Education	0.11	0.005	0	1
Business	0.15	0.007	0	1
Engineering	0.08	0.005	0	1
Public Admin./Soc. Services	0.04	0.003	0	1
Math/Science	0.08	0.005	0	1
Sociology/Psychology	0.12	0.006	0	1
History/Humanities	0.09	0.006	0	1
Other	0.28	0.006	0	1
Health	0.06	0.004	0	1
Parental Income—1st quartile	0.25	0.008	0	1
Parental Income—2nd quartile	0.25	0.009	0	1
Parental Income—3rd quartile	0.25	0.008	0	1
Parental Income—4th quartile	0.25	0.009	0	1
Did not Borrow	0.47	0.010	0	1
Borrowed Less than \$5,000	0.15	0.007	0	1
Borrowed \$5,000–\$10,000	0.18	0.008	0	1
Borrowed \$10,000–\$20,000	0.15	0.007	0	1
Borrowed More than \$20,000	0.04	0.004	0	1
Parental Educational Background				
Less than High School	0.10	0.006	0	1
No More than High School	0.40	0.010	0	1
Some College	0.14	0.006	0	1
A Bachelor's Degree	0.22	0.009	0	1
A Master's Degree or Higher	0.14	0.008	0	1
English First Language Spoken at Home	0.91	0.006	0	1

TABLE 1. (Continued)

	Mean	SD	Minimum	Maximum
Undergraduate Research University	0.30	0.009	0	1
Undergraduate Doctoral University	0.14	0.007	0	1
Undergraduate Comprehensive University	0.36	0.010	0	1
Undergraduate Liberal Arts I College	0.04	0.003	0	1
Undergraduate Other Institution	0.16	0.007	0	1
Private Undergraduate College or University	0.33	0.009	0	1
Job Requires a College Degree	0.63	0.010	0	1
Job has Potential for Career Advancement	0.86	0.007	0	1
Tenure (months)	301.86	7.110	1	9,301
Tenure ² /100	209,815.50	19,515	1	865,000
Experience (months)	47.79	0.085	0	50
Experience ²	2,302.22	6.902	0	2,500
Hours Worked Per Week in 1997	45.22	0.178	2	95

Source: Analyses of the second follow-up (1997) to the NCES Baccalaureate and Beyond (B&B:93/97) longitudinal survey, using B&B:93/97 panel weight (BNBPANEL).

mission of the undergraduate institution attended is reflected by an aggregate of Carnegie classification and control (private versus public). The aggregated Carnegie categories include research, doctoral, comprehensive, liberal arts I, and other 4-year institutions. Other is the reference category.

Limitations

This research is limited in at least four ways. First, in this study, because the analytic sample is restricted to individuals who were employed full-time, the possibility of sample-selection bias is present. Although it may be important, investigating the issue of sample-selection bias is beyond the scope of this study.

Second, propensity score matching is limited in that it does not account for unobserved variables that may affect both the choice to complete a master's degree by program area (type of treatment) and wage earnings outcome. Additionally, propensity score matching does not guarantee that all individuals in the non-treatment group will be matched with individuals in the treatment. This is known as the

'support' problem. In this study, few individuals were poorly matched and consequently, at most, only 13% were not included in the subsequent analyses involving the use of propensity score matching to estimate treatment effects.

Third, the first stage of this study analyzes the attainment of a master's degree within 4–5 years of receiving a bachelor's degree. According to an analysis of 1996 NPSAS data by Choy and Moskovitz (1998), of those students who were enrolled in a master's degree program in 1995–96, less half (41%) enrolled within 2 years, while 71% enrolled within 6 years of attaining their bachelor's degree. Consequently, in this study, the limited number of individuals who received a master's degree during that time frame undermines the precision of the estimates of private returns to a master's degree by program area.

Fourth, the analysis is limited by missing data. The following variables had missing values: income (1.8%), student debt (2.3%), undergraduate GPA (2.7%), and B&B:93/97 panel weight (3.9%). Missing panel weights indicates that some individuals did not participate in the NPSAS:93, B&B:93/94, and the B&B:93/97 surveys. In this study, those individuals are excluded. The deletion of cases with at least one missing response (e.g., listwise deletion) resulted in a slight reduction in the initial analytic sample from 4113 to 3948 cases.

TABLE 2. Estimates of the Average Treatment Effect of Receiving a Master's Degree (by Area of Study) on Wage Earnings (natural log) in 1997, for 1992/93 Bachelor's Degree Recipients, Using OLS Regression

	(1)		(2)	
	OLS Regression Model Without Labor Market Variables		OLS Regression Model With Labor Market Variables	
Master's Degree by Area	Beta	S.E. [†]	Beta	S.E. [†]
Education	0.001	0.0323	0.046	0.0286
Business/Management	0.111	0.0494*	0.213	0.0496***
All Others	-0.081	0.0537	-0.016	0.0525
R ²	0.158		0.252	

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, [†]robust standard errors which take into account the clustered (by undergraduate institution) nature of the NCES Baccalaureate and Beyond sample design.

Source: Analyses of the second follow-up (1997) to the NCES Baccalaureate and Beyond (B&B:93/97) longitudinal survey, using B&B:93/97 panel weight (BNBPANEL).

TABLE 3. Predictors of Receiving a Master's Degree (by Area of Study) by 1997 for 1992/93 Bachelor's Degree Recipients, Using A Multinomial Probit Model

	Education		Business/Management		All Others	
	Marginal Effect	S.E.†	Marginal Effect	S.E.†	Marginal Effect	S.E.†
Student Characteristics						
Female	0.011	0.137**	-0.004	0.154	0.113	0.007
Male (reference)						
African American	-0.007	0.276	0.022	0.259*	0.034	0.205*
Asian	-0.009	0.490	0.009	0.422	0.044	0.226**
Hispanic	-0.005	0.350	0.007	0.337	0.041	0.258
White (reference)						
1 st quartile GPA (reference)						
2 nd quartile GPA	0.001	0.199	0.044	0.263***	0.045	0.163***
3 rd quartile GPA	0.007	0.193*	0.051	0.263***	0.052	0.167***
4 th quartile GPA	0.002	0.213	0.075	0.256***	0.082	0.157***
Undergraduate Major						
Education	0.126	0.265***	-0.014	0.275**	0.004	0.188
Business	-0.003	0.197	0.013	0.181*	-0.004	0.136
Engineering	0.000	0.489	-0.012	0.324	0.101	0.214***
Public Administration or Social Services	0.008	0.433	-0.001	0.340	0.082	0.270**
Math, Physical or Life Science	0.040	0.320*	-0.012	0.263*	0.057	0.192***
Social Science or Psychology	0.064	0.280***	-0.006	0.213	0.040	0.179**
History or Humanities	0.064	0.303***	-0.014	0.354*	0.027	0.179*
Other	0.026	0.283*	-0.012	0.247*	0.043	0.178**

<i>Health (reference)</i>						
1 st quartile parents' income (reference)	0.000	0.182	0.002	0.217	0.149	0.010
2 nd quartile Parents' Income	0.010	0.185*	0.004	0.235	0.144	0.011
3 rd quartile Parents' Income	0.007	0.178	0.014	0.223*	0.145	0.012*
<i>Did Not Borrow (reference)</i>						
Borrowed Less than \$5,000	0.004	0.204	-0.001	0.227	0.001	0.140
Borrowed \$5,000 - \$10,000	-0.002	0.162	0.006	0.223	-0.011	0.133
Borrowed \$10,000 - \$20,000	-0.003	0.198	0.001	0.201	-0.013	0.161
Borrowed \$20,000 or More	0.005	0.305	-0.007	0.339	-0.002	0.315
Parents' Less than High School	-0.005	0.219	-0.003	0.242	0.010	0.186
Parents' No More than High School	-0.007	0.156	0.005	0.156	-0.011	0.111
Parents' Some College	0.000	0.271	0.008	0.238	-0.021	0.165*
Parents' Bachelor's Degree	-0.003	0.144	-0.004	0.173	0.001	0.119
<i>Parents' Master's Degree or More (reference)</i>						
English Spoken at Home	0.005	0.258	-0.005	0.306	0.007	0.189
<i>English Not Spoken (reference)</i>						
Undergraduate Research University	0.003	0.234	0.014	0.270*	0.030	0.170**
Undergraduate Doctoral university	0.005	0.238	0.007	0.299*	0.029	0.180*
Undergraduate Comprehensive University	0.000	0.202	0.018	0.252	0.000	0.156
Undergraduate Liberal Arts 1 College	0.000	0.277	0.017	0.337	0.001	0.230
<i>Undergraduate Other Colleges (reference)</i>						
Private	0.000	0.154	0.003	0.180	0.135	0.010
<i>Public (reference)</i>						
Wald χ^2	408.02					

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † robust standard errors of the beta coefficients (not shown) which take into account the clustered (by undergraduate institution) nature of the NCES Baccalaureate and Beyond sample design.

Source: Analyses of the second follow-up (1997) to the NCES Baccalaureate and Beyond (B&B:93/97) longitudinal survey, using B&B:93/97 panel weight (BNBPANEL).

RESULTS

Table 2 presents the results of the OLS regression models for master's degree by major program area. Table 3 shows the results of the selection model, which is estimated via the multinomial probit regression model. The results of the multinomial probit regression model are reported as marginal effects, the probability of receiving a master's degree in each area due to a discrete change in the value of each independent variable from 0 to 1, holding all other variables constant. Table 4 presents the estimates, generated from propensity score matching techniques, of the average treatment effect (ATE), average treatment on the treated (ATT), and average treatment on the untreated (ATU) effect of receiving a master's degree by major area on wage income.

OLS Regression Results

In Table 2, the first OLS regression model includes, although not shown, the pre-treatment variables that are described above. In addition to the pre-treatment variables in the first OLS regression model, the labor market variables described above are included in the second OLS regression model.

The statistically significant parameter estimate shown in the first OLS regression model indicates that receiving a master's degree in business/management results in an 11% ($\beta = 0.111$, $p < 0.05$) return, after taking into account the pre-treatment variables described above. The second OLS regression model, which takes into account the pre-treatment variables and variables reflecting labor market experience, shows that receiving a master's degree in business/management results in an 21% return ($\beta = 0.213$, $p < 0.001$). As shown in Table 2, the OLS regression models do not indicate that receiving a master's degree in education or other areas result in a gain in wage earnings. A Durbin-Wu-Hausman test ($\chi^2 = 342.607$, $p < 0.001$) reveals that the OLS estimates are inconsistent due to the endogeneity of the variables reflecting receiving a master's degree in education, business/management, and all other areas. In an effort to check the OLS regression model estimates of treatment effects for endogeneity (self-selection) bias, it is necessary to generate predicted propensity scores from a multinomial probit regression, use propensity score matching estimates of treatment effects, and compare those estimates to the OLS parameter estimates.

Multinomial Probit Regression Results

Table 3 reveals that being female increases the probability (marginal effect = 0.011, $p < 0.01$) of receiving a master's degree in education. Having an undergraduate major in education increases the likelihood (marginal effect = 0.126, $p < 0.001$) of receiving a master's degree in education. The chance of earning a master's degree in education is also positively influenced by having an undergraduate major in the social sciences or psychology (marginal effect = 0.064, $p < 0.001$), and history or humanities (marginal effect = 0.064, $p < 0.001$).

Table 3 also shows that the likelihood of an individual receiving a master's degree in business/management increases with undergraduate GPA, particularly with an undergraduate GPA in the highest quartile (marginal effect = 0.075, $p < 0.001$). Having an undergraduate degree in education (marginal effect = -0.014, $p < 0.01$), math, physical, and life sciences (marginal effect = -0.012, $p < 0.05$), history or humanities (marginal effect = -0.014, $p < 0.05$), or other fields (marginal effect = -0.012, $p < 0.01$), reduces the chances of receiving a master's degree in business/management.

As indicated in Table 3, the probability of an individual receiving a master's degree in the "All Other" category increases with undergraduate GPA. Compared to those whose undergraduate GPA was in the first quartile, individuals who have an undergraduate GPA in the highest quartile are more likely (marginal effect = 0.082, $p < 0.001$) to receive a master's degree in an area other than education or business/management. An individual with an undergraduate major in areas other than education or business, has a higher chance of receiving a master's degree in an area other than education or business/management. Receiving a master's degree in the "All Other" category is also positively associated with attending a research (marginal effect = 0.030, $p < 0.01$) or a doctoral university (marginal effect = 0.029, $p < 0.05$).

Compared to those from families where parents have a master's degree or higher, individuals from families where parents have some college have a lower chance (marginal effect = -0.021, $p < 0.05$) of receiving a master's degree in an area other than business/management or education.

Propensity Scores

The multinomial probit selection model in Table 3 was used to generate predicted probabilities of earning a master's degree by the program

areas above. These predicted probabilities were used as propensity scores. In the appendix, Tables A.1 through A.2 present the cumulative distributions of propensity scores for each major program of master's degree. The overall cumulative distributions are consistent with what is generally found for propensity scores (Rosenbaum, 1984). The tables reveal that for each program area, there are a substantial number of non-recipients of master's degrees with similar propensity scores.

Although not shown, box plots of the propensity scores based on whether an individual received a master's degree in education, or business/management, or in another area reveal that there is an overlap in propensity scores between the treated and the untreated. This indicates that conditioning on the observable independent pre-treatment variables above, the difference in earnings outcomes between the individuals who received a master's degree and individuals who did not can be compared within each of the relevant common-support regions. This further warrants the use of the propensity scores to match individuals within each program area. Among matched individuals, the difference in earnings between the treatment group (received a master's degree) and those in the untreated group (did not receive a master's degree) can be interpreted as the treatment effect by program area.

Propensity Score Matching Results—Treatment Effects

Derived from each of the matching techniques, the average treatment effect (ATE), treatment on the treated (ATT) effect, and treatment on the untreated (ATU) effect of receiving a master's degree by area of study are shown in Table 4. The standard errors that are presented were generated from a repeated re-sampling procedure or bootstrap repetitions. Table 4 also includes the percent of cases that were matched (common support region) by each matching technique for each of the treatment types (master's degree by area of program).

With the exception of the estimates of the treatment effects of receiving a master's degree in education derived from the nearest neighbor matching technique, the estimates from propensity score matching techniques are consistent. All three matching techniques produced consistent estimates of the treatment effects of receiving a master's in business/management and "all other" areas.

In contrast to the results from the OLS regression models in Table 2, the results in Table 4 reveal that the kernel matching estimates of the ATE and the ATU effect of receiving a master's degree in education are statistically significant and negative. These findings indicate that individuals with a low probability of receiving a master's degree in education

TABLE 4. Estimates of the Average Treatment Effect (ATE), Average Treated on the Treated (ATT) and Average Treated on the Untreated (ATU) Effect of Receiving a Master's Level (by Area of Study) on 1992/93 Bachelor's Degree Recipient's Wage Earnings (natural log) in 1997, Based on Matching of Propensity Scores Derived from the Multinomial Probit Model in Table 3

	Education			Business/Management			All Other		
	ATE	ATT	ATU	ATE	ATT	ATU	ATE	ATT	ATU
Nearest Neighbor	-0.052	0.038	-0.056	0.198**	0.153*	0.199**	-0.002	-0.045	0.001
S.E.	0.0382	0.0384	0.0391	0.0660	0.0701	0.0721	0.0426	0.0423	0.0407
Percent of Cases Matched	87.1			94.4			98.1		
Kernel	-0.095**	0.031	-0.101**	0.192**	0.163**	0.193**	-0.024	-0.040	-0.023
S.E.	0.0359	0.0299	0.0332	0.0586	0.0433	0.0653	0.0454	0.0329	0.0382
Percent of Cases Matched	98.1			94.0			98.1		
Local Linear Regression	-0.086*	0.025	-0.090*	0.192**	0.147**	0.194**	-0.026	-0.040	-0.025
S.E.	0.0346	0.0260	0.0379	0.0640	0.0444	0.0643	0.0358	0.0323	0.0388
Percent of Cases Matched	87.1			94.4			98.1		

Source: Analyses of the second follow-up (1997) to the NCES Baccalaureate and Beyond (B&B:93/97) longitudinal survey, using propensity scores generated from a multinomial probit model and panel weighted (BNBPANEL) data. Standard errors were generated using bootstrapping replications.

realized a negative private return. As shown in Table A.1, two-thirds of the individuals who received a master's degree in education had less than a 10% chance of receiving such a degree. Because, in this study, the period of observation is limited to 4 years, it is possible that the negative private return to receiving a master's degree in education may be the result of substantially slower wage income growth among individuals in the field of education compared to other fields. In the future, more research, using data with a longer observation period, is needed to shed additional light on this particular finding. It is worth noting that the ATT effect of receiving a master's degree is not statistically significant. This indicates that *if* individuals who had received a master's degree had not received a master's degree, they would have not experienced either a loss or a gain in wage income.

In comparison to the OLS regression estimates of the treatment effects of receiving a master's degree in business/management, the estimates of the treatment effects from the kernel matching of propensity scores are more revealing. Compared to the first OLS regression model parameter estimate of the ATE of receiving a master's degree in business/management [11% (0.111, $p < 0.05$)] reported in Table 2, the kernel matching technique estimate of the ATE of 19% (0.192, $p < 0.01$), ATT effect of 16% (0.163, $p < 0.01$), and ATU effect of 19% (0.193, $p < 0.01$) are higher. Using the kernel matching estimates of treatment effects as reference points, the first OLS regression model parameter estimate of receiving a master's degree in business/management has an overall bias of -0.08 (0.111–0.192) and reveals a self-selection bias of -0.05 (0.111–0.163). The kernel matching estimate of the ATE of receiving a master's degree in business/management is consistent with the estimate derived from the second OLS regression model in Table 2. The estimates of the ATE and ATT effect derived from the kernel matching of propensity scores is only slightly below the OLS regression model parameter estimate of the ATE of receiving a master's degree in business/management ($\beta = 0.213$, $p < 0.001$) (See Table 2), which takes labor market variables into account. This suggests that using observable variables and kernel matching of propensity scores to estimate the ATE for receiving a master's degree in business/management largely adjusts for self-selection bias associated with to the effect of unobservable variables, which may include variables related to labor market experience.

Because Rosenbaum (1984) contends that a counterfactual framework that is inclusive of both pre-treatment and post-treatment variables, which are not influenced by the treatment, may actually help to produce less biased estimates of treatment effects, labor market variables were added to a subsequent kernel propensity score matching procedure. The

results of that procedure indicate that the estimate of the ATE of receiving a master's degree in business/management (0.231, $p < 0.001$) is slightly above the second OLS regression model's estimated coefficient for business/management (See Table 2). This suggests that a more fully-specified OLS regression model tends to slightly underestimate the private returns to receiving a master's degree in business/management.⁹

Table 4 shows that the kernel matching estimate of the ATT effect of receiving a master's degree in business/management is slightly lower than the estimates for the ATE and ATU effect. This demonstrates that *if* individuals who received a master's degree in business/management had not received a master's degree, their return would have been slightly lower than the average return for individuals in the entire sample. The results of kernel matching estimates of the ATT effect imply that individuals who have a higher propensity of receiving a master's degree in business/management do not realize private returns over and above individuals who are less inclined to receive a master's degree in business/management. The kernel matching estimates of the ATU effect indicate that *if* individuals who did not receive a master's degree in business/management had actually received the degree, their return would be 19% or the average for the entire sample. This suggests that the private returns to receiving a master's degree in business/management is slightly higher for individuals with slightly lower chances of receiving such a degree.

Consistent with the OLS regression models in Table 2, the propensity score matching did not produce statistically significant estimates of the ATE, ATT effect or the ATU effect of receiving a master's degree in an area other than education and business/management. Observation over a longer period and more data disaggregated are needed to further our understanding of how receiving of master's degree in areas other than education or business/management influences wage earnings.

CONCLUSIONS

Based on the results of this research, several conclusions can be made. First, this line of inquiry also provided an alternative approach to estimating self-selection bias, which is typically found in higher education research. This study demonstrated the usefulness of a counterfactual analytical framework, which allows for the estimation of potential outcomes in a treated and untreated state, enabling the use of studies utilizing matched comparison groups to make stronger inferences about causality. Using a counterfactual analytical framework, this investigation revealed how the use of propensity score matching techniques can

be employed by institutional and other researchers to help address growing concerns about the lack of “evidenced-based” policies and practices in education, including higher education.

Second, this study demonstrated how unlike OLS regression analysis, propensity score matching, can be utilized to decompose treatment effects. In addition to average treatment effects (ATE), average treatment on the treated (ATT) effects and average treated on the untreated (ATU) effects can be estimated using propensity score matching. In this research, it was demonstrated that, in the absence of experimental design or randomized trials, the “randomization” of *ex post facto* data, via propensity score matching techniques, helps us to decompose the wage effects of receiving a master’s degree.

Third, this study showed how propensity score matching can avoid the problems of distributional assumptions inherent in other methods which have been used to estimate the ATE. This study demonstrated how such methods as OLS regression may produce inaccurate estimates of the ATE, due to self-selection bias.

Fourth, using national survey data and propensity score matching, this study demonstrated how researchers can estimate the ATE, ATT effect and ATU of receiving a master’s degree in certain program areas on wage earnings. Using this example, the study also detected the differences between OLS regression model parameter estimates of the ATE and estimates of the ATE generated by propensity score matching. In some instances, propensity score matching produced statistically significant estimates of the ATE (for the master’s degree in education) while the OLS regression model did not. In other instances, propensity score matching produced noticeably higher estimates of the ATE (for master’s degree in business/management) than the parameter estimate from the OLS regression model which included the same variables that were used to generate the relevant propensity score from the selection model. This latter finding is consistent with research in other areas (Heckman and Li, 2004; Morgan 2001) that compared propensity score matching estimates with OLS regression parameter estimates of the ATE. Therefore, this study adds to a growing body of recent evidence that OLS regression models do not adequately adjust for self-selection bias, cannot appropriately decompose treatment effects, and tend to underestimate the private returns to education.

IMPLICATIONS FOR FUTURE RESEARCH

The results of this study have implications for future research in at least three areas. First, in the future, when making inferences about

causality, institutional and other higher education researchers should consider the use of a counterfactual analytical framework and propensity score matching to estimate treatment outcomes. In the future, research which utilizes *ex post facto data*, at the very least, should employ propensity score matching to check the robustness of OLS regression-derived parameter estimates of the average treatment effect and for self-selection bias.

Second, although this article used national survey data on bachelor's degree recipients to demonstrate the use propensity score matching techniques, other data sets should be used to investigate the usefulness of applying these particular techniques to other important areas of study in higher education. Such areas could include the various treatment effects of financial aid on the student access to, involvement in, and completion of college.

Third, in the future, additional studies should also employ the use of propensity score matching to examine the average treatment, average treatment on the treated, and average treatment on the untreated effect of different levels of educational attainment on earnings and other outcomes over the long term. The combined use of propensity score matching and temporal difference-in-differences analyses (Blundell and Costa, 2000) will allow for an investigation of the effects of treatment on changes in outcomes. More importantly, by combining propensity score matching with other techniques, institutional and other higher researchers will have a wider array of analytical tools at their disposal to help them address the problem of selection bias when conducting research using *ex post facto* observational data and non-experimental design.

ACKNOWLEDGMENTS

A research award from the North Carolina State University Faculty Research and Professional Development Fund (FR&PD) supported this research. Opinions reflect those of the author and do not necessarily reflect those of the granting institution. This paper is based on a revision of a paper presented at the Annual Forum of the Association for Institutional Research Boston, MA, May 30–June 2, 2004. The author thanks Stephen L. DesJardins for his helpful comments on the earlier version of this paper.

END NOTES

1. The error terms in these equations are assumed to be jointly normally distributed. More specifically, the errors in the selection (probit or logit regression) models and the outcome

- (OLS regression) models equations are assumed to follow a bivariate normal distribution.
2. Unlike the logit model in which the errors are assumed to follow the logistic distribution, the errors of the probit model are assumed to follow the standard normal distribution.
 3. Compared to the multinomial logit model, the multinomial probit model allows for correlations among the errors (u_i), which relaxes the IIA assumption.
 4. Additional assumptions include partial equilibrium and stable unit treatment value assumption (SUTVA) exists. Partial equilibrium exists to the extent to which an individual's treatment decision does not depend on the treatment decisions of others. The SUTVA requires that the effect of treatment one individual is not dependent on the treatment of other individuals or on how many individuals receive treatment. These assumptions are violated if peer effects influenced an individual's treatment decision and the selection criteria that were used by treatment program administrators.
 5. In this study, when using the nearest neighbor method, matching based is on one-to-one neighbor.
 6. With the kernel method of the propensity score matching, each individual in the analytic sample receives a weight which is calculated as follows:

$$w^{KM}(i,j) = \frac{K\left(\frac{P_j - P_i}{h}\right)}{\sum_{\lambda \in \{D=0\}} K\left(\frac{P_k - P_i}{h}\right)}$$

where $w^{KM}(i,j)$ is the weight with which individual j without a master's degree is assigned to individual i with a master's degree, $K(\cdot)$ is a kernel function (e.g., normal distribution), P_i is the propensity score of individual i without a master's degree, P_j is the propensity score of individual j with a master's degree, and h is a bandwidth parameter. This method of propensity score matching assures that individuals with comparable propensity scores receive large weights and individuals with differing scores receive small weights.

7. In this study, when kernel propensity score matching was used, a narrow bandwidth (0.1) was chosen so as to minimize the extent to which the scores are smoothed and thus enable more fine grained estimation of the treatment effects.
8. With the local linear regression method of the propensity score matching, each individual in the analytic sample receives a weight which is calculated as follows:

$$W(i,j) = \frac{K_{ij} \sum_{k \in I_0} K_{ik}(P_k - P_i)^2 - [K_{ij}(P_j - P_i)][\sum_{k \in I_0} K_{ik}(P_k - P_i)]}{\sum_{j \in I_0} K_{ij} \sum_{k \in I_0} K_{ik}(P_k - P_i)^2 - (\sum_{k \in I_0} K_{ik}(P_k - P_i))^2},$$

where I_0 is a set of controls. In this study, when using the local linear regression method of the propensity score matching, so as to compare to the kernel matching, the bandwidth was set to 0.1.

9. Although not shown, compared to kernel propensity score matching estimates of the ATE, an instrumental variable (IV) model (using propensity scores as the instruments) tends to overestimate the positive (negative) ATE of receiving a master's degree in business/management (education).

APPENDIX

TABLE A.1. Cumulative Distribution of Estimated Propensity Scores for Receiving a Master's of Degree in Education (Derived from the Multinomial Probit Regression Model in Table 3)

Propensity Score	Master's of Education		No Master's of Education (Bachelor's—Any Major)	
	Proportion	Sample Size	Proportion	Sample Size
0.01	0.116	15	0.512	1844
0.02	0.209	27	0.630	2271
0.03	0.279	36	0.712	2566
0.04	0.380	49	0.769	2771
0.05	0.457	59	0.818	2948
0.06	0.504	65	0.853	3076
0.07	0.589	76	0.880	3171
0.08	0.628	81	0.905	3262
0.09	0.667	86	0.923	3326
0.10	0.729	94	0.940	3388
0.11	0.760	98	0.951	3427
0.12	0.798	103	0.962	3467
0.13	0.829	107	0.969	3495
0.14	0.853	110	0.975	3515
0.15	0.891	115	0.981	3535
0.16	0.899	116	0.986	3553
0.17	0.907	117	0.989	3566
0.18	0.922	119	0.991	3573
0.19	0.946	122	0.995	3588
0.20	0.953	123	0.996	3590
0.21	0.977	126	0.997	3595
0.22	0.992	128	0.998	3599
0.23	1.000	129	0.999	3602
0.24	1.000	129	0.999	3603
0.26	1.000	129	1.000	3605

Source: Analyses of the second follow-up (1997) to the NCES Bachelor's and Beyond (B&B:93/97) longitudinal survey, using propensity scores generated from a multinomial probit model and panel weighted (BNBPANEL) data.

TABLE A.2. Cumulative Distribution of Estimated Propensity Scores for Receiving a Master's of Degree in Business/Management (Derived from the Multinomial Probit Regression Model in Table 3)

Propensity Score	Master's of Business/ Management		No Master's of Business/ Management (Bachelor's—Any Major)	
	Proportion	Sample Size	Proportion	Sample Size
0.01	0.176	15	0.584	2079
0.02	0.329	28	0.717	2552
0.03	0.506	43	0.798	2840
0.04	0.600	51	0.856	3048
0.05	0.624	53	0.897	3195
0.06	0.694	59	0.926	3297
0.07	0.765	65	0.947	3373
0.08	0.788	67	0.963	3428
0.09	0.871	74	0.977	3478
0.10	0.918	78	0.984	3505
0.11	0.941	80	0.991	3530
0.12	0.953	81	0.994	3541
0.13	0.953	81	0.996	3546
0.14	0.953	81	0.997	3549
0.15	0.953	81	0.997	3551
0.16	0.965	82	0.997	3552
0.17	0.988	84	0.998	3554
0.19	0.988	84	0.998	3555
0.20	0.988	84	0.999	3557
0.22	0.988	84	0.999	3558
0.23	0.988	84	1.000	3560
0.33	1.000	85	1.000	3561

Source: Analyses of the second follow-up (1997) to the NCES Bachelor's and Beyond (B&B:93/97) longitudinal survey, using propensity scores generated from a multinomial probit model and panel weighted (BNBPANEL) data.

TABLE A.3. Cumulative Distribution of Estimated Propensity Scores for Receiving a Master's Degree in "Other Areas" (Derived from the Multinomial Probit Regression Model in Table 3)

Propensity Score	Master's in Other Areas		No Master's in Other Areas (Bachelor's—any major)	
	Proportion	Sample Size	Proportion	Sample Size
0.01	0.061	16	0.152	567
0.02	0.129	34	0.274	1023
0.03	0.186	49	0.394	1473
0.04	0.240	63	0.503	1881
0.05	0.300	79	0.589	2202
0.06	0.395	104	0.663	2480
0.07	0.468	123	0.733	2739
0.08	0.544	143	0.786	2938
0.09	0.597	157	0.821	3070
0.10	0.662	174	0.855	3198
0.11	0.707	186	0.883	3301
0.12	0.726	191	0.901	3369
0.13	0.779	205	0.924	3456
0.14	0.810	213	0.940	3516
0.15	0.852	224	0.953	3565
0.16	0.871	229	0.963	3600
0.17	0.894	235	0.971	3630
0.18	0.913	240	0.976	3651
0.19	0.924	243	0.981	3667
0.20	0.939	247	0.986	3685
0.21	0.947	249	0.990	3702
0.22	0.962	253	0.993	3714
0.23	0.966	254	0.995	3720
0.24	0.966	254	0.995	3722
0.25	0.985	259	0.998	3731
0.26	0.985	259	0.998	3732
0.27	0.989	260	0.998	3733
0.28	0.992	261	0.999	3734
0.29	0.996	262	0.999	3737
0.30	0.996	262	1.000	3738
0.35	1.000	263	1.000	3739

REFERENCES

- Agodini, R., and Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics* 86(1): 180–194.
- Amemiya, T. (1985). *Advanced Econometrics*, Harvard University Press, Cambridge.
- Becker, G. S. (1993). *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*, (3rd Ed.), University of Chicago Press, Chicago.
- Becker, S. O., and Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *Stata Journal* 2: 358–377.
- Bellio, R., and Gori, E. (2003). Impact evaluation of job training programmes: Selection bias in multilevel models. *Journal of Applied Statistics* 30(8): 893–907.
- Bourdieu, P. (1986). The forms of capital. In: Richardson, J. G. (ed.), *Handbook of Theory and Research for the Sociology of Education*, Greenwood Press, New York, pp. 241–258.
- Bourdieu, P., and Passeron, J. C. (1977). *Reproduction in Education, Society, and Culture*, Sage Publications, Beverly Hills, CA.
- Blundell, R., and Costa, D. M. (2000). Evaluation methods for nonexperimental data. *Fiscal Studies* 21(4): 427–468.
- Choy, S. P., and Moskowitz, R. (1998). *Student Financing of Graduate and First-Professional Education, 1995–96: With Profiles of Students in Selected Degree Programs*, National Center for Education Statistics, Washington, D.C.
- Cochran, W. G., and Cox, G. M. (1950). *Experimental Design*, John Wiley, New York.
- Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology* 94(Supplement): 95–120.
- Carneiro, P., and Heckman, J. J. (2002). The evidence on credit constraints in post-secondary schooling. *The Economic Journal* 112(482): 705–734.
- Conniffe, D., Gash, V., and O’Connell, P. J. (2000). Evaluating state programmes: “Natural experiments” and propensity scores. *The Economic and Social Review* 31(4): 283–308.
- Daganzo, C. D. (1979). *Multinomial Probit, The Theory and Its Applications to Demand Forecasting*, Academic Press, New York.
- Davidson, R., and MacKinnon, J. G. (2004). *Econometric Theory and Methods*, Oxford University Press, New York.
- Dehejia, R. H., and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94(448): 1053–1062.
- DesJardins, S. L., McCall, B. P., Ahlburg, D. A., and Moye, M. J. (2002). Adding a timing light to the “tool box”. *Research in Higher Education* 43(1): 83–114.
- Doiron, D. J. (2004). Welfare reform and the labour supply of lone parents in Australia: A natural experiment approach. *Economic Record* 80(249): 157–176.
- Eltinge, J. L., and Sribney, W. M. (1996). Svy1: Some basic concepts for design-based analysis of complex survey data. *Stata Technical Bulletin* 31: 3–6 Reprinted in *Stata Technical Bulletin Reprints*, 6, 208–213.
- Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall, New York.
- Fisher, R. A. (1935). *The Design of Experiments*, Hafner, New York, N.Y. Reprint 1971.
- Frölich, M. (2004). Finite sample properties of propensity-score matching and weighting estimators. *The Review of Economics and Statistics* 86(1): 77–90.
- Fuller, W. C., Manski, C. F., and Wise, D. A. (1982). New evidence on the economic determinants of postsecondary schooling choices. *Journal of Human Resources* 17(4): 477–498.

- Garen, J. E. (1984). The returns to schooling: A selectivity bias approach with a continuous choice variable. *Econometrica* 52(5): 1199–1217.
- Glenn, D. (2005, March). New federal policy favors randomized trials in education research. The Chronicle of Higher Education, Retrieved March 25, 2005 from <http://www.chronicle.com>.
- Greene, W. H. (2000). *Econometric Analysis*, (4th Ed.), Prentice-Hall, Upper Saddle River, NJ.
- Hagen, T. (2002). Do temporary workers receive risk premiums? Assessing the wage effects of fixed- term contracts in West Germany by a matching estimator compared with parametric approaches. *Labour* 16(4): 667–705.
- Hausman, J. A., and Wise, D. A. (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica* 46(2): 403–426.
- Heckman, J. J. (1974). Shadow prices, market wages, and labor supply. *Econometrica* 42(4): 679–694.
- Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* 47(1): 153–162.
- Heckman, J. (1997). Instrumental variables: A study in implicit behavioral assumptions used in making program evaluations. *The Journal of Human Resources* 32(3): 441–462.
- Heckman, J., Ichimura, H., and Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies* 64(4): 605–654.
- Heckman, J. J., and Li, X. (2004). Selection bias, comparative advantage and heterogeneous returns to education: Evidence from China. *Pacific Economic Review* 9(3): 155–171.
- Heckman, J. J., and Robb, R. (1985). Alternative methods for evaluating the impact of interventions. In: Heckman, J. J., and Singer, B. (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge University Press, Cambridge, UK, pp. 156–245.
- Hill, J. L., Reiter J. P., and Zanutto, E. L. (2004). A comparisons of experimental and observational data analyses. In: Gelman A., and Meng. X. (eds.), *Applied Bayesian Modeling and Causal Inference from an Incomplete-Data Perspective*. Wiley, pp. 44–56.
- Idson, T., and Feaster, D. (1990). A selectivity model of employer-size wage differentials. *Journal of Labor Economics* 8(1): 99–122.
- John, E. P., and Asker, E. H. (2001). The role of finances in student choice: A review of theory and research. In: Paulsen, M. B., and Smart, J. C. (eds.), *The finance of higher education: Theory, research policy, and practice*, Agathon Press, New York, pp. 419–438.
- Kemphorne, O. (1952). *Design and Analysis of Experiments*, John Wiley, New York.
- Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *The Review of Economics and Statistics* 84(2): 205–220.
- Lee, L. F. (1978). Unionism and wage rates: A Simultaneous equations model with qualitative and limited dependent variables. *International Economic Review* 19(2): 415–433.
- Leuven, E., and Sianesi, B. (2004). “PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.” EconPapers. <http://www.econpapers.repec.org/software/bocbocode/S432001.htm>.
- Levy, P., and Lemeshow, S. (1999). *Sampling of Populations*, (3rd Ed.), John Wiley & Sons, New York.
- Lin, N. (2001). Building a network theory of social capital. In: Lin, N., Cook, K., and Burt, R. S. (eds.), *Social Capital: Theory and Research*, Aldine De Gruyter, New York, pp. 3–29.
- Louviere, J. J., Hensher, D. A., and Swait, J. D. (2000). *Stated Choice Methods: Analysis and Applications*, Cambridge University Press, Cambridge.
- Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*, Harvard University Press, Cambridge.

- Maddala, G. S. (1998). Recent developments in dynamic econometric modeling: A personal viewpoint. *Political Analysis* 7(1): 59–87.
- Manski, C. F. (1993). Adolescent econometricians: How do youth infer the returns to schooling? In: Clotfelter, C. T., and Rothschild, M. (eds.), *Studies of Supply and Demand in Higher Education*, University of Chicago Press, Chicago.
- Main, B., and Reilly, B. (1993). The employer size-wage gap: Evidence for Britain. *Economica* 60(238): 125–142.
- Manski, C. F., and Wise, D. A. (1983). *College Choice in America*, Harvard University Press, Cambridge.
- McDowell, A., and Pitblado, J. (2002). From the help desk; It's all about the sampling. *The Stata Journal* 2: 190–201.
- McMillen, D. P. (1995). Selection bias in spatial econometric models. *Journal of Regional Science* 35(3): 417–436.
- Mincer, J. (1974). *Schooling, Experience, and Earnings*, Brookfield, VT, Ashgate Publishing Company.
- Morgan, S. L. (2001). Counterfactuals, causal effect, heterogeneity, and the catholic school effect on learning. *Sociology of Education* 74(4): 341–374.
- National Center for Education Statistics (1999). *Baccalaureate and Beyond Longitudinal Study: 1993/97 Second Follow-up Methodology Report*, U.S. Department of Education, Washington, D.C.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. (translated in 1990). *Statistical Science* 5: 465–480.
- Neyman, J. (1935). Statistical problems in agricultural experimentation (with discussion). *Supplement to the Journal of the Royal Statistical Society* 2(Supplement): 107–180.
- Olsen, J. R. (1980). A least squares correction for selectivity bias. *Econometrica* 48(7): 1815–1820.
- Paulsen, M. B., and John, E. P. (2002). Social class and college costs: Examining the financial nexus between college choice and persistence. *The Journal of Higher Education* 73(2): 189–236.
- Pindyck, R. S., and Rubinfeld, D. L. (1998). *Econometric Models and Economic Forecasts*, (4th Ed.), Irwin/McGraw-Hill, Boston.
- Perna, L. W. (2000). Differences in the decision to attend college among African Americans, Hispanics, and Whites. *The Journal of Higher Education* 71(2): 117–141.
- Perna, L. W. (2004). Understanding the decision to enroll in graduate school: Sex and racial/ethnic group differences. *The Journal of Higher Education* 75(5): 487–527.
- Portes, A. (1988). Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology* 24: 1–24.
- Porter, S. R. (2006). Institutional structures and student engagement. *Research in Higher Education* 47(5): 521–556.
- Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association* 67(338): 306–310.
- Rosenbaum, P. (2002). *Observational Studies*, Springer, New York.
- Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association* 79(385): 41–48.
- Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 51–55.
- Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3(2): 135–146.

- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688–701.
- Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* 2(1): 1–26.
- Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal Educational Statistics* 6(4): 377–400.
- Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association* 81(396): 961–962.
- Rubin, D. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference* 25(3): 279–292.
- Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* 47(4): 1213–1234.
- Rubin, D. B., and Thomas, N. (1996). Matching using propensity scores, relating theory to practice. *Biometrics* 52(1): 249–264.
- Rubin, D. B., and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* 95(450): 573–585.
- Saiz, A., and Zoido, E. (2005). Listening to what the world says: Bilingualism and earnings in the United States. *Review of Economics and Statistics* 87(3): 523–538.
- Schwartz, J. B. (1985). Student financial aid and the college enrollment decision: The effects of public and private grants and interest subsidies. *Economics of Education Review* 4(2): 129–144.
- Thomas, S. L., and Perna, L. W. (2004). The opportunity agenda: A reexamination of postsecondary reward and opportunity. In: Smart, J. C. (ed.), *Higher Education: Handbook of Theory and Research* (Vol. XIX), Kluwer Academic Publishers, Netherlands, pp. 43–84.
- Whitehurst, G. (2002). <http://www.ed.gov/nclb/methods/whatworks/eb/edlite-slide021.html>.
- Willis, R. J. (1986). Wage determinants: A survey and reinterpretation of human capital earnings functions. In: Ashenfelter, O., and Layard, R. (eds.), *Handbook of Labor Economics*, Elsevier Science Pub. Co, New York.
- Willis, R. J., and Rosen, S. (1979). Education and self-selection. *Journal of Political Economy* 87(5): S7–S36 part 2.
- Winship, C., and Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology* 25: 659–706.

Received October 7, 2005.