Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools[1]

Joseph G. Altonji[2]
Todd E. Elder[3]
Christopher R. Taber[2]

March 26, 2001

[1]We thank Timothy Donohue and Emiko Usui for excellent research assistance. We also received helpful comments from Thomas DeLeire, James Heckman, Robert LaLonde, Jean-Marc Robin, George Jakubson and especially Tim Conley and Derek Neal as well as participants in seminars at American Economic Association Winter meetings (January 2000), Boston College, Cornell University, CREST-INSEE, Johns Hopkins University, Harvard University, MIT, Northwestern University, Princeton University, University of Chicago, University College London, University of Florida, University of Maryland, University of Rochester, and University of Wisconsin at Madison. We are grateful for support from the National Science Foundation (Altonji), the National Institute of Child Health and Development (Altonji and Taber), and the Institute for Policy Research, Northwestern University.

[2]Department of Economics and Institute for Policy Research, Northwestern University and NBER

[3]Department of Economics and Institute for Policy Research, Northwestern University

Abstract

In this paper we measure the effect of Catholic high school attendance on educational attainment and test scores. Because there is no valid instrumental variable for Catholic school attendance, we develop new estimation methods based on the idea that the amount of selection on the observed explanatory variables in a model provides a guide to the amount of selection on the unobservables. We show that if the observed variables are a random subset of a large number of factors that influence the outcome of interest, then the relationship between the index of observables that determines Catholic school attendance and the index that determines the outcome will be the same as the relationship between the indices of unobservables that determine the two variables. In some circumstances this fact may be used to identify the effect or to bound the effect of the endogenous variable. We also propose an informal way to assess selectivity bias based on measuring the ratio of selection on unobservables to selection on observables that would be required if one is to attribute the entire effect of the Catholic school attendance to selection bias. We use our methods to estimate the effect of attending a Catholic high school on a variety of outcomes. Our main conclusion is that Catholic high schools substantially increase the probability of graduating from high school and, more tentatively, college attendance. We do not find much evidence for an effect on test scores.

1 Introduction

Distinguishing between correlation and causality is the most difficult challenge faced by empirical researchers in the social sciences. Social scientists rarely are in a position to run a well controlled experiment. Consequently, they rely on a priori restrictions on the patterns of interaction among the variables that are observed or unobserved. These restrictions are typically in the form of exclusion restrictions, restrictions on the functional form of the model, restrictions on the distribution of the unobserved variables, or restrictions on dynamic interactions. Occasionally, the a priori restrictions are derived from a widely accepted theory or are supported by other studies that had access to a richer set of data. However, in most cases, doubt remains about the validity of the identifying assumptions and the inferences that are based on them.

The challenge of isolating causal effects is particularly difficult for the question addressed in our paper—"Do Catholic high schools provide better education than public schools?" This question is at the center of the debate in the United States over whether vouchers, charter schools, and other reforms that increase choice in education will improve the quality of education. It is also highly relevant to the search for ways to improve teaching and governance of public schools. Simple cross tabulations or multivariate regressions of outcomes such as test scores and post secondary educational attainment typically show a substantial positive effect of Catholic school attendance.[1] However, many prominent social scientists, such as Goldberger and Cain (1982), have argued that the positive effects of Catholic school attendance may be due to spurious correlations between Catholic school attendance and unobserved student and family characteristics. The argument begins with the observation that it costs parents time and money to send their children to private school. In the absence of experimental data, the challenge in addressing this potentially large bias is finding exogenous variation that affects school choice but not outcomes. Most student background characteristics that influence the Catholic school decision, such as income, attitudes, and education of the parents, are likely to influence outcomes independently of the school sector because they are likely to be related to other parental inputs. Characteristics of private and public schools that influence choice, such as tuition levels, student body characteristics, or

[1] The most influential examples are Coleman, Hoffer, and Kilgore (1982) and Coleman and Hoffer (1987). Other early example examples of studies of Catholic schools and other private schools are Noell (1982), Goldberger and Cain (1982), and Alexander and Pallas (1985). Recent studies include Evans and Schwab (1995), Tyler (1994), Neal (1997), Grogger and Neal (1999), Figlio and Stone (2000), Sander (2000) and Jepsen (2000). Murnane (1984), Witte (1992), Chubb and Moe (1990), Cookson (1993) and Neal (1998) provide overviews of the discussion and references to the literature. Grogger and Neal provide citations to a small experimental literature, which for the most part has found positive effects of Catholic school.

school policies, are likely to be related to the effectiveness of the schools.

Several recent studies, including Evans and Schwab (1995), Neal (1997), Grogger and Neal (1999), Figlio and Stone (1999) and Altonji, Elder and Taber (1999) use various exclusion restrictions to estimate the Catholic school effect on a variety of outcomes. Evans and Schwab (1995) use religious affiliation as an exogenous source of variation in Catholic school attendance and confirm the large positive estimates of Catholic school effects on high school graduation and college attendance that they obtain when Catholic school attendance is treated as exogenous. However, as Evans and Schwab recognize and Murnane et al (1985) and Neal (1997) note, being Catholic could well be correlated with characteristics of the neighborhood and family that influence the effectiveness of schools. Another influential paper by Neal (1997) uses proxies for geographic proximity to Catholic schools as an exogenous source of variation in Catholic high school attendance. The basic assumption is that the location of Catholics and/or Catholic schools was determined by historical circumstances and is independent of unobservables that influence performance in schools. He finds evidence of a positive effect of Catholic high school attendance on high school and college graduation among students in urban areas, particularly in the case of nonwhites. In Altonji, Elder and Taber (1999, 2001), we employ a similar methodology using data on zip code of residence and the zip codes of all of the Catholic high schools in the country to compute a measure of distance from the nearest Catholic high school for our samples. We conclude that the use of location or location interacted with religion is not a good way to estimate Catholic school effects.[2] Grogger and Neal (2000) come to a similar conclusion.[3] Altonji, Elder, and Taber (2001) also find that Catholic religion has a strong association with graduation rates for students who attended public eighth grades even though such students rarely attend Catholic high school. This evidence and work by Ludwig (1997) raises serious doubts about the validity of Catholic religion as an instrument.

[2]We provide evidence based on links to observed variables and to eighth grade test scores that suggests that neither distance from Catholic high schools nor the interaction between distance and religion should be excluded from the outcome equations unless detailed controls for location are included. (This informal use of observables as a guide to correlation between the instrument and the unobservables lead to the current paper.) Failure to control for these factors leads to negative biases in estimates of Catholic school effects. Unfortunately, including detailed geographic controls (such as 3 digit zip code) leads to very large standard errors. We also follow Neal (1997) and Evans and Schwab (1995) by using bivariate probit models to jointly estimate the Catholic School decision with the outcomes. We find that empirical identification comes largely from the functional form of the model rather than exclusion of the measure of distance from Catholic schools. Nonlinearities in the effects of student background rather than proxies for distance from Catholic schools seem to be the main source of identification.

[3]Grogger and Neal (2000) use NELS:88, the data set for the present study. Altonji, Elder and Taber (1999) analyze NELS:88, the National Longitudinal Survey of the High School Class of 1972, and NLSY. Neal (1997) uses the National Longitudinal Survey of Youth, 1979.

In this paper we develop new estimation strategies that may be helpful when strong prior information is unavailable regarding the exogeneity of either the variable of interest or instruments for that variable. We view this to be the situation in studies of Catholic school effects and in many other applications in economics and the other social sciences. We then use our strategies to assess the effectiveness of Catholic schools.

Our approach to estimation is based on the idea of using the degree of selection on observables as a guide to how much selection there is on the unobservables. Researchers often informally argue for the exogeneity of membership in a "treatment group" or for the exogeneity of an instrumental variable by examining the relationship between group membership or the instrumental variable and a set of observed characteristics, or by assessing whether point estimates are sensitive to the inclusion of additional control variables.[4] We provide a formal analysis confirming the intuition that such evidence can be informative. More importantly, we provide a way to quantitatively assess the degree of omitted variables bias.[5]

Not surprisingly, the empirical relevance of using the link between the endogenous variable and observed variables as a guide to the link with the unobservables hinges on how the observables and unobservables are related. Using our Catholic schools application, let the outcome Y be determined by

$$Y = \alpha CH + W'\Gamma \,, \tag{1.1}$$

where CH is an indicator for whether the student attends a Catholic high school, the parameter α is the effect of Catholic school attendance on Y, W is a vector containing the full set of other variables that influence Y, and Γ is the coefficient vector for W. Let X represent the vector of observed components of W and γ be the vector of associated elements of Γ. Then one may rewrite the above equation as

$$Y = \alpha CH + X'\gamma + \varepsilon, \tag{1.2}$$

where the vector X is observed by the econometrician and the error term ε is the index of the unobserved elements of W weighted by the corresponding elements of Γ.

[4]See for example, Currie and Duncan (1995), Engen et al (1996), Poterba et al (1994), Angrist and Evans (1998), Jacobsen et al. (1999), Bronars and Grogger (1994), or Udry (1998).

[5]Two precursors to our study are Altonji's (1988) study of the importance of observed and unobserved family background and school characteristics in the school specific variance of educational outcomes and especially Murphy and Topel's (1990) study of the importance of selection on unobserved ability as an explanation for industry wage differentials.

Let CH^* be the latent variable that determines CH, with $CH = 1(CH^* > 0)$ and consider

$$\text{Proj}(CH^*|X'\gamma, \varepsilon) = \phi_0 + \phi_{cx}X'\gamma + \phi_{c\varepsilon}\varepsilon \tag{1.3}$$

where Proj(.|.) denotes a linear projection. We call the condition

$$\phi_{cx} = \phi_{c\varepsilon} \tag{1.4}$$

"equality of selection on the observables and unobservables". This condition relates selection into Catholic school to the factors that determine outcome Y. This equation says that the coefficient ϕ_{cx} relating Catholic school attendance to the index of observables $X\gamma$ that determine the outcome Y is equal to coefficient $\phi_{c\varepsilon}$ on the index of unobservables ε. The identical coefficients on $X'\gamma$ and ε capture the idea that "selection on the unobservables is similar to selection on the observables." We show that if (i) the elements of X are chosen at random from W and (ii) the number of elements in both X and W are large so that none of the elements dominates the distribution of school choice CH or the outcome Y, then equality of selection will hold. While (1.4) is strong, we argue that is a better approximation in many circumstances than the standard assumption underlying OLS and other single equation methods, which is that

$$\phi_{c\varepsilon} = 0. \tag{1.5}$$

Major data sets with large samples and extensive questionnaires are not designed to address one relatively specific question, such as the effectiveness of Catholic schools, but rather to serve multiple purposes. As a result of limits on the number of factors that we know matter for a particular outcome, know how to collect, and can afford to collect, many elements of W are left out. This is reflected in the typically low the explanatory power of social science models of individual behavior. Furthermore, in many applications, including ours, the endogenous variable is correlated with many of the elements of X. Given the constraints that shape choice of X and the fact that many of the elements of X are systematically related to CH^*, it is unlikely that the many unobserved variables that determine ε have are unrelated to CH^*, which is what (1.5) requires. Our view is that analysis based on (1.4) is a useful complement to the standard analysis based on (1.5).

We show that in some circumstances (1.4), prior knowledge about the sign of the bias, and an additional condition on the relationship among the included and excluded variables are sufficient to identify α. In others, (1.4) and the other conditions are only sufficient to

restrict α to be the root of a cubic equation. If selection on the observables is stronger than selection on the unobservables, as we argue is true in our case, then one can identify a lower bound for α.

Operationally, we estimate the joint model of Catholic school attendance and the outcome subject to a transformation of (1.4). Similar ideas can be applied to "heterogeneous effects" models in which the benefits of Catholic school attendance and public school attendance vary across students.

We provide a similar estimation method that may be used when an excluded variable (e.g., Catholic religion or proximity to a Catholic school in the Catholic schools literature) is used to identify a model, but there are concerns about whether it is exogenous. We show that if the observables are a random subset of the variables that determine the outcome, then the instrumental variable will have the same relationship with the regression index of the observables and the regression index of the unobservables. This condition can sometimes be used to identify α even though the instrumental variable is correlated with the error term in the outcome equation.

We also propose a related but more informal way to use the relationship between the observables as a guide to endogeneity bias. It is related to the common practice of checking for a systematic relationship between CH and the mean of the elements of X. Let $E(.)$ and $Var(.)$ be mean and variance operators. We compute $\frac{E(X'\gamma|CH=1)-E(X'\gamma|CH=0)}{Var(X'\gamma)}$, which is the normalized shift in the index of observables in the outcome equation that is associated with CH, and then ask how many times larger the normalized shift in the index of the unobservables $\frac{E(\varepsilon|CH=1)-E(\varepsilon|CH=0)}{Var(\varepsilon)}$ would have to be to explain away the entire estimate of α. The null hypothesis that the single equation estimator of α is unbiased corresponds to the case in which $\frac{E(\varepsilon|CH=1)-E(\varepsilon|CH=0)}{Var(\varepsilon)}$ is 0, while the hypothesis that X is a randomly chosen subset of W implies that

$$\frac{E(\varepsilon|CH=1)-E(\varepsilon|CH=0)}{Var(\varepsilon)} = \frac{E(X'\gamma|CH=1)-E(X'\gamma|CH=0)}{Var(X'\gamma)}. \tag{1.6}$$

If selection on unobservables must be several times stronger than selection on the observables to explain for the entire estimate of α, then the case for a causal effect of Catholic school is strengthened.

In section 2 we set the stage for the development and application of our econometric methods by providing a standard multivariate analysis of the Catholic school effect using the National Educational Longitudinal Survey of 1988 (NELS:88). We present descriptive statistics on the relationship between Catholic school attendance and a broad range of

observable measures of family background, eighth grade achievement, educational expectations, social behavior, and delinquency. The descriptive statistics show huge Catholic high school advantages in high school graduation and college attendance rates, and smaller ones in 12th grade test scores. However, the evidence across the wide range of observables, which have substantial explanatory power in our outcome equations, suggests fairly strong positive selection into Catholic schools. We also find that the link between observables and Catholic high school attendance is much weaker among children who attended Catholic eighth grade. To reduce sample selection bias and to avoid confounding the effect of attending Catholic high school with the effect of Catholic elementary school, we use the Catholic eighth grade sample for much of our analysis, unlike most previous studies.

We present an initial set of regression and probit models containing detailed controls for student characteristics that are determined prior to high school. We find a small positive effect on 12th grade math scores, and a zero effect on reading scores. However, our estimates of the effect of Catholic high school point to a very large positive effect of 0.15 on the probability of attending a 4 year college 2 years after high school and 0.08 on the high school graduation rate. The estimates are not very sensitive to the addition of a powerful set of controls, particularly in the case of the high school graduate rate. The insensitivity of the results to the controls and the "modest" association between the observables that determine the outcome and Catholic high school suggests that part of the educational attainment effect is real. However, the small positive effects on math test scores could easily be accounted for by positive selection on unobservables.

In sections 3 and 4 we develop and apply our methods for using the degree of selection on observables to provide better guidance about bias from selection on unobservables. Because high school outcomes depend on many variables that are determined after the decision to attend Catholic high school is made, selection on unobservables that affect outcomes is likely to be weaker than selection on observables, with $0 \leq \phi_{c\varepsilon} \leq \phi_{cx}$. Consequently, our estimates of a joint model of Catholic high school attendance and educational attainment subject to (1.4) are likely to overstate selection and understate the Catholic school effect. The estimate of the effect of Catholic school on high school graduation declines from the univariate estimate of about 0.08, which we view as an upper bound, to 0.07 when we impose equal selection, which we view as a lower bound, although sampling error widens this range. The estimate of the effect on college attendance declines from the univariate estimate of 0.15 to 0.07 or 0.02, depending on the details of the estimation method.

Using (1.6) we estimate that selection on unobservables would have to be between 2.78

and 4.29 times stronger than selection on the observables to explain away the entire Catholic school effect on high school graduation, which seems highly unlikely. It would have to be between 1.30 and 2.30 times stronger to explain away the entire college effect, which is also unlikely. However, more modest positive selection on the unobservables could explain away the entire Catholic school effect on math scores. We conclude that Catholic high school attendance substantially boosts high school graduation rates and, more tentatively, college attendance rates.

In section 5 we extend our analysis to a subsample of urban minorities, for whom we obtain larger univariate effects but also stronger evidence of selection. In section 6 we provide conclusions and an agenda for further research on the use of observables as a guide to selection bias.

2 A Preliminary Analysis of the Catholic School Effect

In section 2.1 we describe the data. In section 2.2 we present the sample means of outcomes, measures of family background, eighth grade achievement, social behavior, and delinquency as a way of assessing the potential importance of selection bias and to motivate the choice of sample. In section 2.3 we present probit and OLS regression estimates of the Catholic school effect which serve as a benchmark for our subsequent analysis. In section 2.4 we present an analysis of the sensitivity of the Catholic high school effect to assumptions about the degree of selection on unobservables.

2.1 Data

Our data set is NELS:88, a National Center for Education Statistics (NCES) survey which began in the Spring of 1988. The base year sample is a two stage stratified probability sample in which a set of schools containing eighth grades were chosen on the basis of school size and whether they were classified as private or public. In the second stage, as many as 26 eighth grade students from within a particular school were chosen based on race and gender. A total of 1032 schools contributed student data in the base year survey, resulting in 24,599 eighth graders participating. Subsamples of these individuals were reinterviewed in 1990, 1992, and 1994. The NCES only attempted to contact 20,062 base-year respondents in the first and second follow-ups, and only 14,041 in the 1994 survey. Additional observations are lost due to attrition.

The NELS:88 contains information on a wide variety of outcomes, including test scores and other measures of achievement, high school dropout and graduation status, and post-secondary education (in the 1994 survey only). Parent, student, and teacher surveys in the base year provide a rich set of information on family and individual background, as well as pre-high school achievement, behavior, and expectations of success in high school and beyond. Each student was also administered a series of cognitive tests in the 1988, 1990, and 1992 surveys to ascertain aptitude and achievement in math, science, reading, and history, and the evolution of achievement throughout high school. We use standardized item response theory (IRT) test scores that account for the fact that the difficulty of the 10th and 12th grade tests taken by a student depends on the 8th grade scores. We use the 8th grade test scores as control variables and the 10th and 12th grade reading and math tests as outcome measures.

We also use high school graduation and college attendance as outcome measures. The high school graduation variable is equal to one if the respondent graduated high school by the date of the 1994 survey, and zero otherwise.[6] The "College attendance" indicator is one if the respondent was enrolled in a four-year university at the date of the 1994 survey and zero otherwise.[7]

The indicator variable for Catholic high school attendance, CH, is one if the current or last school in which the respondent was enrolled was Catholic as of 1990 (two years after the eighth grade year) and zero otherwise.[8]

We estimate models using a full sample, a Catholic eighth grade sample, and various other subsamples. We always exclude approximately 400 respondents who attended non-Catholic private high schools or private, non-Catholic eighth grades. Observations with

[6] We obtain similar results using a "drop out" dummy variable which equals one if a student dropped out of high school by 1992, or if the student dropped out of high school by 1990 and was not reinterviewed in 1992 or 1994, zero otherwise. This variable catches dropouts who left the survey by 1990 and were either dropped from the sample or were nonrespondents.

[7] Our major findings are robust to whether or not college attendance is limited to 4-year universities, full-time versus part-time, or enrolled in college "at some time since high school" or at the survey date.

[8] A student who started in a Catholic high school and transferred to a public school prior to the tenth grade survey would be coded as attending a public high school ($CH = 0$). If such transfers are frequently motivated by discipline problems, poor performance, or alienation from school, then misclassification of the transfers as public high school students could lead to upward bias in estimates of the effect of CH on educational attainment. We investigated this issue using an 8th grade question about whether the student expected to attend Catholic high school and information about whether the student had changed high schools prior to the 10th grade survey. Among Catholic school 8th graders for whom we have the relevant data, 832 of 889 kids (94%) who reported that they expected to attend Catholic high school actually attended Catholic high school. Among the remaining 57, only 12 students had transferred at least once and of these only 3 failed to graduate high school. Furthermore, it is quite possible that 1 or 2 of these students never started Catholic high school, perhaps because of a family move. We conclude that any bias from misclassification of students is small.

missing values of key eighth grade or geographic control variables (such as distance from the nearest Catholic high school) were dropped. Sample sizes vary across dependent variables because of data availability and are presented in the tables. The sampling probabilities for the NELS:88 followups depend on choice of private high school and the dropout decision, so sample weights are used to avoid bias from a choice based sample. Unless noted, the results reported in the paper are weighted.[9] Details regarding construction of variables and the composition of the sample are provided in Appendix B.

2.2 Characteristics of Catholic High School and Public High School Students by Eighth Grade Sector

In Table 1 we report the weighted means by high school sector of a set of family background characteristics, student characteristics, eighth grade outcomes, and high school outcomes. We report results separately for students who attended Catholic eighth grades (the "C8" sample) and for the full sample. The outcomes category displays by high school sector the college attendance rate, high school graduation rate, and 10th and 12th grade math and reading test scores for students from the NELS:88 sample.[10] Looking at the full sample, the graduation and college attendance rates differ enormously between the two sectors. Catholic high school students are one fifth as likely to drop out of high school as their public school counterparts (0.03 versus 0.15), and are twice as likely to be enrolled in a four year college in 1994 (0.59 versus 0.29). Differences in twelfth grade test scores are more modest but still substantial—about 0.4 of a standard deviation higher for Catholic high school students. In the C8 sample the gap in the dropout rate is also very large (0.02 versus 0.10), as is the gap in the college attendance rate (0.62 versus 0.39). The gap in the 12th grade math score is about 0.25 standard deviations. Table 2 shows that the gaps in

[9]In the initial sample, private schools and schools with a minority enrollment of over 19 percent were oversampled. The probability of sampling in the first and second follow-ups is smaller for high schools attended by fewer than 10 students from the NELS:88 base year sample and the weight declines with the number of sample members in the high school. This is likely to lead to undersampling of students who attend private high schools. In contrast, the third follow-up sample design oversamples those who attended private high schools. Furthermore, the sampling probability depends on whether the student was believed to have dropped out of high school. Because the sample probabilities depend on an endogenous right-hand side variable and the school attainment variables, it is necessary to weight the analysis to obtain consistent parameter estimates. We use the first follow-up panel weights for the analysis of 10th grade test scores, the second follow-up panel weights for the analysis of 12th grade scores, and the third follow-up cross section weights for the analysis of high school graduation and college attendance. The results are somewhat sensitive to the use of sample weights, although our main findings are robust to weighting. Given the sampling scheme the weighted estimates are clearly preferred.

[10]In Table 1 and Table 2 the outcome variables are weighted with the same weights used in the regression analysis, as described in the previous section. All other variables are weighted using second follow-up panel weights.

school attainment and test scores are even more dramatic for minority students in urban schools.

Tables 1 and 2 also show that the means of favorable family background measures, 8th grade test scores and grades, and positive behavior measures in eighth grade are substantially higher for the students who attend Catholic high schools. The large discrepancies for many of the variables raise the possibility that part or even all of the gap in outcomes may be a reflection of who attends Catholic high school. However, the gap is much lower for most variables in the case of Catholic eighth graders. For example, the gap in log family income is 0.49 for the full sample but only 0.19 for the C8 sample. The high school sector gap in measures of the parents' educational expectations for the child is more favorable to the students who attend Catholic high school in the full sample than in the eighth grade sample, and the difference in the student's expected years of schooling is 0.72 in the full sample but only 0.40 in the C8 sample.[11] The high school sector differential in father's education is about one year in both samples, but for mother's education it is 0.75 for the full sample and 0.54 for the C8 sample. The discrepancy in the fraction of students who repeated a grade in grades 4-8 is -0.05 in the full sample and only -0.01 in the C8 sample, and the gap in the fraction of students who are frequently disruptive is -0.05 in the full sample and 0 in the C8 sample. Both of these variables are powerful predictors of high school graduation. Finally, the gap in the 8th grade reading and math scores are 3.86 and 3.44, respectively, in the full sample, but only 1.47 and 1.09, respectively, in the C8 sample.

These results hold for most of the other variables in Table 1. Specifically, differences by high school sector among the family background characteristics and eighth grade outcomes are much smaller for Catholic eighth graders than for public eighth graders. This pattern is consistent with the presumption that since the parents of 8th graders from Catholic schools have already chosen to avoid public school at the primary level, other, arguably more idiosyncratic factors, are likely to drive selection into Catholic high schools from Catholic eighth grade. Intuitively, it seems likely that these factors could lead to less selection bias than in the full sample, although the overwhelming evidence based on very broad set of 8th grade observables is that selection bias is positive in both samples. These considerations, concern about selection bias arising from the fact that only a 0.3% of public school eighth graders in our effective sample go to Catholic high school, and the desire to

[11]Appendix B and the footnotes to Table 3 provide the complete list of variables used in our multivariate models. Many are excluded from Tables 1 and 2 to keep them manageable. The expectations variables in Tables 1 and 2 are excluded from our outcome models because if Catholic school has an effect on outcomes, this may be influence expectations.

avoid confounding the Catholic high school effect with the effect of Catholic elementary school lead us to focus on the sample of Catholic eighth graders in much of our analysis, in contrast to most previous studies.[12]

2.3 Estimates of the Effect of Catholic High Schools

In this section of the paper we present regression and univariate probit estimates of the effects of Catholic high school attendance on a set of outcomes. For reasons discussed above, we focus on the subsample who attended Catholic eighth grade, although we also present results for the full sample.

In the top panel of Table 3 we report the coefficient on CH, the Catholic high school attendance dummy, in univariate probit, OLS, and school fixed effects models for high school graduation.[13] The difference in means is 0.08 when no controls are included. In the probit model, the coefficient is 0.88 (0.25), with an associated average marginal effect on the graduation rate of 0.084, which is a huge effect given that the graduation rate is 0.947 among students from Catholic eighth grades. In this sample the family background and geographic controls explain *none* of the raw difference of 0.08 in the graduation rate. The point estimate of the marginal effect of CH declines slightly to 0.081 when we add eighth grade test scores in column 5, and increases to 0.088 when we add a large set of eighth grade measures of attendance, attitudes toward school, academic track in eighth grade, achievement, and behavioral problems. The stability of the Catholic school effect is remarkable, especially given the fact that the control variables in column 6 are quite powerful, explaining 0.32 percent of the variance in the latent variable for high school graduation when CH is excluded from the model.

The second row in Table 3 is based on linear probability models of high school graduation. The coefficient on CH varies from 0.080 to 0.081 and closely agrees with the probit estimates. Row 3 of columns 4-6 adds eighth grade fixed effects to the specifications reported in row 2.[14] The fixed effects estimate is .115 for the basic specification and 0.102

[12] This is an unweighted percentage. The weighted percentage is 0.8%. We have made similar calculations based on the sample of 16,070 individuals for whom information on sector of eighth-grade and sector of 10th grade is available. The corresponding estimate of the percentage of the eighth graders from public schools who attend Catholic high schools is 0.3%. If one restricts the analysis to individuals whose parents are Catholic, only 0.7% of students who attended public eighth-grade attend a Catholic high school. The unweighted and weighted estimates of the percentage of Catholic high school 10th graders who attended Catholic eighth-grade are 95.2 percent and 84.7 percent.

[13] Huber-White standard errors are reported throughout the paper. The standard errors account for the use of weights and, with the exception of Table 7 and 8, they account for correlation among students from the same eighth grade.

[14] That is, it includes separate intercepts for each eighth grade.

when the full set of controls is included.[15]

In Table 3 we also report estimates of the effect of Catholic high school attendance on the probability that a student is enrolled in a 4 year college at the time of the 3rd follow-up survey in 1994, 2 years after most students graduate from high school. For the basic specification (column 4) the probit estimate implies that CH raises the college enrollment probability by 0.154, which compares to a raw difference of 0.23. This estimate falls to 0.149 when we add detailed controls to the model. Linear probability models yield similar estimates.

In Table 4 we report estimates of the effect of CH on 10th and 12th grade reading and math scores. In contrast to the above findings, we obtain modest **negative** estimates of the effects of Catholic high schools on 10th grade reading scores. In the simplest specification for the Catholic eighth grade subsample, we obtain a coefficient of -1.07 (0.97), which rises to -0.87 (0.77) when the full set of controls and eighth grade fixed effects are added. We obtain a small but statistically insignificant coefficient of -0.32 (1.01) in the case of math, but this estimate declines to essentially 0 when we add detailed controls.

In the bottom panel of Table 4 we report estimates of the effects on 12th grade reading and math scores. For the Catholic eighth grade sample with the full set of controls we obtain a small positive effect of 1.14 (0.46) on the math score and 0.33 (0.62) on the reading score. As Grogger and Neal (1999) emphasize, a positive effect of Catholic schools on the high school graduation rate might lead to a downward bias in the Catholic high school coefficient in the 12th grade test equations given that dropouts have lower test scores and that dropouts have a lower probability of taking the 12th grade test. However, the issue appears to be of only minor importance.[16]

---

[15]We report fixed effects results not because the use of fixed effects is necessarily a more appropriate estimator but rather to show that factors that vary across Catholic elementary schools (such as public high school quality) do not drive the large positive estimates of the Catholic high school effect. Bias from individual heterogeneity could well be more severe in the within-school than the cross-school analysis.

[16]We deal with this issue by filling in missing data for both high school graduates and dropouts using predicted values from a regression of the 12th grade score on the full set of controls in the outcome regression, plus the Catholic high school dummy and the 10th grade test scores and a dummy variable for whether the individual graduated from high school (high school graduation has a small and statistically insignificant coefficient). Using the new dependent variable and sample the estimated effect of Catholic high schools for 12th grade math and reading are 1.20 and 0.58 respectively. We obtain 1.20 and 0.56, respectively, when we use an alternative imputation in which we adjust for differences in unobservables using the assumption that the difference between dropouts with and without 12th grade test scores in the mean/variance of the regression residual from the test score prediction regression is the same as the difference in the mean/variance of the predicted values of the tests. The R^2 of the prediction equations are 0.70 for reading and 0.86 for math. The estimates of the reliability of the math test reported in the NELS:88 documentation, while probably downward biased, are in the 0.87 to 0.90 range. Consequently, a substantial part of the test score residual probably reflects random variation in test performance and is unrelated to achievement levels. For this reason selection on unobservables in the availability of test data

To facilitate a comparison to other studies, we also present estimates for the combined sample of students from Catholic and public eighth grades. For this sample the effect of Catholic high school attendance is reduced from 0.081 to 0.052 after we add the full set of controls (Table 3, columns 1-3). It is interesting to note, however, that the OLS estimate is only 0.023 once the full set of controls are added and differs substantially from the probit estimate of the average marginal effect. The college attendance results largely mirror the high school graduation results. The probit estimate of the effect of Catholic school attendance is 0.074 once the full set of controls are included, which is substantial relative to the mean college attendance probability of 0.31. Note that the controls make a much larger difference in the full sample than in the Catholic eighth grade sample, which is consistent with the evidence that selection on the observables is more severe in the full sample.

Once detailed controls for eighth-grade outcomes are included, the estimates of the effect of Catholic high schools on 10th grade math and reading scores are essentially 0, and the estimates of the effects on 12th grade reading and math are only 1.14 and 0.92, respectively. Again, there is little evidence that Catholic high schools increase achievement by 10th grade, in accordance with the findings based on the Catholic eighth grade subsample. In contrast, the 12th grade math and reading score results indicate a small but statistically significant positive effect. Given the high degree of selection into Catholic high school in the full sample on the basis of observable traits, these estimates may reflect the effects of unobserved differences between public and Catholic high school students rather than actual effects on test scores, and should be interpreted with caution.

2.4 A Sensitivity Analysis.

Based on observables, there is not much evidence of selection. However, it is possible that a small amount of selection on unobservables could explain the whole Catholic school effect. We now explore this possibility by examining the sensitivity of the estimates of the Catholic high school effect to the correlation between the unobserved factors that determine CH and the various outcomes Y. We display estimates of the Catholic school effect for a range of values of the correlation between the unobserved determinants of school choice in the outcome.

Consider the bivariate probit model

is probably less strong than selection on the predicted portion of the test scores.

$$CH_i = 1(X_i'\beta + u > 0) \tag{2.1}$$

$$Y_i = 1(X_i'\gamma + \alpha CH_i + \varepsilon > 0) \tag{2.2}$$

$$(u, \varepsilon) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right). \tag{2.3}$$

where $\mathbf{1}(\cdot)$ is the indicator function taking the value one if its argument is true and zero otherwise. While this model is formally identified without an exclusion restriction, semiparametric identification requires such an excluded variable. Furthermore, empirical researchers are highly skeptical of results from this model in the absence of an exclusion restriction. Our thought exercise in this section is to treat this model as if it were underidentified by one parameter. In particular, we act as if ρ is not identified.[17]

In Table 7 we display estimates of Catholic schooling effects that correspond to various assumptions about ρ, the correlation between the error components in the equation for CH and Y.[18] We report results for high school graduation in the top panel and college attendance in the bottom panel, and include both probit coefficients and average derivatives of the outcome probabilities (in brackets). We include family background, eighth grade tests, and other eighth grade measures. However, because of convergence problems in estimating the bivariate probit models we eliminated the dummy variables for household composition (but not marital status of parents), urbanicity, region, and indicators for "student rarely completes homework", "student performs below ability", "student inattentive in class", "a limited English proficiency index", and "parents contacted about behavior" from the set of controls. We vary ρ from 0 (the probit case that we have already considered above) to 0.5 by estimating probit models constraining ρ to the specified value. For the full sample, the raw difference in the high school graduation probability is 0.12. When $\rho = 0$ the estimated effect is 0.058, and the figure declines to 0.037 when $\rho = 0.1$ and to 0.011 if $\rho = 0.2$. Given the strong relationship between the observables that determine high school graduation and Catholic school attendance in the full sample, the evidence for a strong Catholic school effect is considerably weaker than the estimates that take Catholic school attendance as exogenous suggest.

For our preferred sample of Catholic 8th graders, the results are less sensitive to ρ. The effect on high school graduation is 0.078 when $\rho = 0$,which is slightly below the estimate we

[17]We use the bivariate probit because it is convenient. An alternative would be to treat ε and u nonparametrically subject to the normalization $var(\varepsilon) = var(u) = 1$ and the restriction $corr(\varepsilon, u) = \rho$.

[18]See Rosenbaum (1995) for examples of this type of sensitivity analysis.

obtain with the full set of controls in Table 3. It declines to 0.038 when $\rho = 0.3$ and is still positive when $\rho = 0.5$. Thus, for the Catholic 8th grade sample, the correlation between the unobservable components of Catholic school attendance and high school graduation would have to be greater than 0.5 to explain the estimated effect under the null of no "true" Catholic high school effect.

In the bottom panel of Table 7 we present the results for college attendance. For the full sample, the results are very similar to the high school graduation results. The evidence for a positive effect of CH on college attendance is stronger in the Catholic 8th grade sample than in the full sample, with the effect remaining positive until ρ is about 0.3. However, in this sample the strongest evidence is for a positive effect of CH on high school graduation.

The problem with this type of analysis is that, without prior knowledge, there is no manner to judge the magnitude of ρ. We will show in the section 3 that assuming that "selection on the unobservables is similar to selection on the observables" can solve this problem.

2.5 Lessons from the Preliminary Analysis

Our preferred results, which are based on the Catholic eighth grade sample, suggest a strong positive effect of CH on high school graduation and college attendance. The estimates of the effect on 12th grade test scores are much smaller (Tables 5 and 6 present qualitatively similar results for urban minorities, which we consider in detail in section 5). The key question is how much of the estimated high school effect on educational attainment is real, and how much is due to selection bias. We have taken advantage of the fact that the NELS:88 data set contains an unusually rich set of family background variables and eighth grade outcomes that are likely to be relevant for educational attainment and achievement to provide some guidance regarding the extent of selection bias. The means of favorable variables are typically higher for Catholic high school students, suggesting positive selection bias. However, positive selection is more modest in the sample of Catholic eighth graders, and in this sample the estimates of the effect of CH on high school graduation and college attendance, are very large. Furthermore, the estimates are not very sensitive to the addition of a powerful set control variables, especially in the high school graduation case. Finally, in Table 7 we show that even with what seems like a large amount of correlation between the observables and unobservables, we cannot explain away all of the Catholic high school effect. In this sample it seems as if the degree of selection must be quite high to explain the full Catholic high school effect. We would conclude that part of the Catholic school effect

on educational attainment is real, but could not go much beyond such a statement. This is where the typical analysis of bias due to selection on unobservables based on patterns in the observables would end.

In the remainder of the paper, we formalize the idea of using the degree of selection on the observables as a guide to bias from selection on unobservables and provide ways of formally incorporating such information into the sensitivity analysis. We then apply our methods to study the effect of CH.

3 Selection Bias and the Link Between the Observed and Unobserved Determinants of School Choice and Education Outcomes

In this section we consider ways to use the relationship among the observed determinants of Catholic school attendance and educational outcomes to provide a quantitative assessment of the importance of the bias resulting from a relationship between the unobserved determinants. In particular we show that modeling how the set of observed variables is determined can yield conditions that are useful for identification.

To motivate this section we consider a model in which CH represents a dummy variable for participation in a "program" such as Catholic high school. Let CH^* denote the latent variable such that

$$CH = 1(CH^* \geq 0).$$

Define another latent variable that depends on CH itself,

$$Y^* = \alpha CH + X'\gamma + \varepsilon. \tag{3.1}$$

Our outcome variable will be some function of this latent variable. In some cases we may be interested in a binary variable such as graduating from high school (GHS) in which the outcome may be $GHS = 1(Y^* \geq 0)$. In others the continuous variable Y^* itself may be the variable of interest (such as test scores in the analysis above).

The fundamental selection/endogeneity problem is that CH^* and thus CH may be correlated with ε. In an ideal world one could solve this problem with an instrumental variable that is strongly correlated with CH and uncorrelated with ε. Unfortunately, we do not believe a perfect instrument exists for Catholic high schools in existing data sets (see Altonji, Elder, and Taber, 2000), so we must use alternative methods for inference.

In particular we consider the following condition which was discussed above

Condition 1

$$Proj(CH^*|X'\gamma, \varepsilon) = \phi_0 + \phi_c X'\gamma + \phi_c \varepsilon. \tag{3.2}$$

This condition relates selection into Catholic school to the factors that determine the outcome Y^*. The identical coefficients on $X'\gamma$ and ε capture the idea that "selection on the unobservables is similar to selection on the observables." We are not advocating imposing Condition 1 to produce an estimate that would be interpreted as the "best" single estimate of α. Instead, we use Condition 1 to inform the type of sensitivity analysis we performed in section 2.4.

It is useful to contrast this condition with the condition which is needed to justify OLS or other standard single equation methods,

Condition 2

$$Proj(CH^*|X'\gamma, \varepsilon) = \phi_0 + \phi_c X'\gamma.$$

There is not much of a case for preferring Condition 2 a priori to Condition 1 in our Catholic school problem. Random assignment of CH, as in a social experiment, would imply that $\phi_c = 0$ in both conditions so that they would both hold. However, if ϕ_c is not zero, it is hard to justify Condition 2, particularly if a large number of covariates have the same sign in the regression and selection equations. We see no reason to expect the omitted determinants of Y^* to be uncorrelated with CH^* if the factors in our data that influence Y^* are systematically related to CH^*.

In section 3.1 we present a model of observable and unobserved variables that justifies Condition 1. We do this for two reasons. First, the model makes clear the type of assumptions that are likely to yield Condition 1 and aids interpretation of the parameters. Second, it justifies the particular multivariate regression representation that we use in Condition 1. While the assumptions are admittedly strong, it would take even assumptions to justify Condition 2. Again, our claim is not that Condition 2 is wrong and should never be used while Condition 1 is right and should always be the basis for identification. We maintain instead that estimating treatment effects under both conditions is helpful in assessing the evidence. In our application we argue that Condition 1 is likely to yield an upper bound on the amount of selection bias while Condition 2 is likely to yield a lower bound. Thus the true effect of Catholic schools is likely to lie between our two estimates. We strongly suspect this to be the case in other contexts although they should be assessed on a case by case basis.

In section 3.2 we point out that a structural model of school choice in which the odds of attending a Catholic school depend directly on the outcome can also lead to Condition 1. In our own application we interpret condition 1 as a bound on the amount of selection, but in section 3.3 we consider the implications of Condition 1 for point identification of the treatment effect. We show that Condition 1 can take the place of an exclusion restriction and deliver identification of the treatment effect.

In section 3.4 we extend the analysis in a few directions. First, we consider the case of continuous dependent variables. Second, we discuss cases in which a possibly invalid instrumental variable is available. An alternative condition for identification of the model above is that one has an instrument that is related to CH, but uncorrelated with the error term conditional on X. This is the usual assumption in studies that use an IV strategy and is equivalent to assuming that

$$\text{Proj}(Z|X'\gamma, \varepsilon) = \phi_0 + \phi_z X'\gamma,$$

where Z is the excluded instrument. We show that the type of data set generation process that leads to Condition 1 also implies that

$$\text{Proj}(Z|X'\gamma, \varepsilon) = \phi_0 + \phi_z X'\gamma + \phi_z \varepsilon.$$

Many of the points made above about Condition 1 and 2 apply to these conditions. It is very hard to argue that an instrument should be orthogonal to unobservable determinants of Y^* when it is correlated with a broad set of observable determinants of Y^*. The third part of the section discusses heterogeneous treatment effects.

In section 3.5 we discuss the relevance of our analysis for studies of the Catholic school effect. We argue that the truth is somewhere between Conditions 1 and 2. Using this argument we show how one can identify a set of permissible values of α.

3.1 A Model of Observed and Unobserved Variables

The goal of this subsection is to present a model that formalizes the idea that "selection on the unobservables is similar to selection on the observables." We derive the model and then show that it implies Condition 1. Let W be the full set of variables that determine Y^* according to

$$Y^* = \alpha CH + W'\Gamma, \tag{3.3}$$

where Γ is a conformable coefficient vector. We assume that Γ is random, but is drawn once and is identical for everyone in the population. However, W and CH are random

variables that vary across members of the population, so that each individual obtains a separate value of W and CH but common values of Γ and α.

Assume that some of the elements of W are observable to the econometrician and others are not (or that the econometrician does not know that some of the observed variables belong in the model for Y^*). Denote the observable portion of W as X and the corresponding elements of Γ as γ so that

$$Y^* = \alpha CH + X'\gamma + \varepsilon, \tag{3.4}$$

where ε is unobserved. That is, for each potential covariate, W_j, let S_j be a dummy variable indicating whether W_j is observable. Then

$$X'\gamma = \sum_{j=1}^{K} S_j W_j \Gamma_j, \quad \varepsilon = \sum_{j=1}^{K} (1 - S_j) W_j \Gamma_j.$$

Like Γ_j, S_j does not vary across the population.

We do not know of a formal discussion of how variables are chosen for inclusion in data sets. However, we make a few general comments that apply to many social science data sets, including NELS:88. First, most large scale data sets such as NLSY, NELS:88, the PSID, and the German Socioeconomic Panel are collected to address many questions. Data set content is a compromise among the interests of multiple research, policy making, and funding constituencies. Burden on the respondents, budget, and access to administrative data sources serve as constraints. Obviously, content is also shaped by what is known (i.e. the factors that really matter for particular outcomes) and by variation in the feasibility of collecting useful information on particular topics. Explanatory variables that influence a large set of important outcomes (such as family income, race, education, gender, or geographical information) or those that are themselves of interest as outcomes, are more likely to be collected. Major data sets with large samples and extensive questionnaires are not designed to address one relatively specific question, such as the effectiveness of Catholic schools, but rather to serve multiple purposes. As a result of the limits on the number of the factors that we know matter and that we know how to collect and can afford to collect, many elements of W are left out. This is reflected in the relatively low explantory power of social science models of individual behavior. Furthermore, in many applications, including ours, the endogenous variable is correlated with many of the elements of X.

These considerations suggest that Condition 2, which underlies single equation methods in econometrics, will rarely hold in practice. The optimal survey design for estimation of α would be to assign the highest priority to variables that are important determinants of

both CH^* and Y^* when choosing S (it would also be to include variables that determine CH^* but not Y^* that can serve as instruments). However, many factors that influence Y^* and are correlated with CH^* and/or X are left out.

An alternative approximation is that the constraints on data collection are sufficiently severe that it is better to think of the elements of X are a more or less a random subset of the elements of W, rather than a set that has been systematically chosen to eliminate bias. We show below that random selection from W leads to Condition 1. Indeed, a natural way to formalize the idea that "selection on the observables is the same as selection on the unobservables" is to treat observables and unobservables symmetrically by assuming that the observables are a "random subset" of a large number of underlying variables. In our notation this amounts to assuming that S_j is an *iid* binary random variable which is equal to one with probability P_S. The outcome of S_j determines whether covariate W_j is observed. Of course, there are other ways to capture the idea of equality of selection on observables and unobservables. For example, P_S may vary across types of variables but have no systematic relationship with the values of Γ_j relative to the influence of the variables on CH^* (β_j below). An important question for future research is the degree to which our results change under different assumptions about the data set generation. To the extent that the data set was designed for the study of the effect of CH^* on Y^*, one might expect that in (3.2) the coefficient on $X'\gamma$ would exceed the coefficient on ε. For this and other reasons discussed in section 3.5, we use (3.2) as the basis for a bound rather than for point identification.

To see the intuition for the link between random choice of variables and condition 1, define ϕ_c and ω such that

$$\text{Proj}\,(CH^* \mid W'\Gamma) = \phi_0 + \phi_c W'\Gamma \tag{3.5}$$

$$\omega = CH^* - \phi_0 - \phi_c W'\Gamma. \tag{3.6}$$

Notice that

$$\begin{aligned} E(X'\gamma\omega) &= E\left(\sum_{j=1}^{K} S_j W_j \Gamma_j \omega\right) \\ &= P_S E\left(\omega \sum_{j=1}^{K} W_j \Gamma_j\right) \\ &= P_S E(\omega W'\Gamma) = 0. \end{aligned} \tag{3.7}$$

Similar logic yields $E(\varepsilon\omega) = 0$. Consequently, since

$$\begin{aligned} CH^* &= \phi_0 + \phi_c W'\Gamma + \omega \\ &= \phi_0 + \phi_c X'\gamma + \phi_c \varepsilon + \omega \end{aligned}$$

and since $X'\gamma$ and ε are orthogonal to ω, Condition 1 holds on average over draws of the vector $\{S_1....S_K\}$. This result in itself is not useful in practice because we only observe one draw of the sequence of S_j and Γ_j. To justify use of Condition 1 we now show that as the number of covariates W gets large, Condition 1 will become approximately true for a given draw of S_j and Γ_j, and equality of selection on observables and unobservables holds.

We will define Y_K^*, CH_K, and CH_K^* as outcomes for a sequence of models where there are K factors that determine Y_K^*.[19] A natural part of the thought experiment in which K varies across models is the idea that the importance of each individual factors declines with K. That is,

$$Y_K^* = \alpha CH_K + \sum_{j=1}^{K} W_j^K \Gamma_j^K, \tag{3.8}$$

where either W_j^K or Γ_j^K depend on K. To insure that Y_K^* is well behaved as K gets large, we specify that the net effect of the change in scale of W_j^K and/or Γ_j^K on the scale of $W_j^K\Gamma_j^K$ is inversely proportional to K, which means that the above equation may be rewritten as

$$Y_K^* = \alpha CH_K + \frac{1}{\sqrt{K}} \sum_{j=1}^{K} W_j \Gamma_j. \tag{3.9}$$

We restrict $W_j\Gamma_j$ in this sequence to be stationary so that the first variable included in the data set will be treated symmetrically with the last variable included so that no particular covariate will be any more important ex-ante than others. This embodies the idea that a large number of factors are important in determining outcomes in social science data and that none dominate. We also need to normalize the model by assuming that $E(W_j\Gamma_j) = 0$.

The variables CH_K and CH_K^* also must be well behaved as the number of covariates gets large. Since these are typically not linear functions of variables this will be done in a very general manner. After presenting the theorem we give an example of a specification for CH_K that satisfies the condition. We put no restrictions on the relationship between CH_K

[19]The "local to unity" literature in time series econometrics and the "weak instruments" literatures (Staiger and Stock, 1997) are other examples in econometrics in which the asymptotic approximation is taken over a sequence of models, which in the case of those literatures, depend on sample size.

and CH_K^*, so our notation is general enough to include the specification $CH_K = CH_K^*$ as well as $CH_K = 1\left(CH_K^* \geq 0\right)$.

We now show that under certain assumptions Condition 1 will hold as the number of elements of W gets large. Note that our asymptotic analysis is nonstandard. First, we are allowing the number of underlying factors, K, to get large. Second, the random variable W_j is different in a sense than random variables Γ_j and S_j. For each j we draw one observation on Γ_j and S_j which are the same for every person in the population; however, each individual will draw their own W_j. Consider the projection of CH_K^* on the observable portion of Y_K, $\frac{1}{\sqrt{K}}\sum_{j=1}^{K} S_j W_j \Gamma_j$, and the unobservable portion, $\frac{1}{\sqrt{K}}\sum_{j=1}^{K}\left(1-S_j\right) W_j \Gamma_j$. This projection is meant to be the population projection (i.e., for a very large number of persons) but with K fixed. That is, this projection conditions on a particular realization of Γ_j and S_j, $j = 1...K$. The theorem states that as K gets large the projection coefficients on $\frac{1}{\sqrt{K}}\sum_{j=1}^{K} s_j W_j \Gamma_j$ and $\frac{1}{\sqrt{K}}\sum_{j=1}^{K}\left(1-s_j\right) W_j \Gamma_j$ will approach each other with probability one.

Theorem 1 *Define*

$$Y_K = \alpha CH_K^* + \frac{1}{\sqrt{K}}\sum_{j=1}^{K} W_j \Gamma_j,$$

where W_j and β_j are independent nondegenerate, stationary, ergodic processes that satisfy the conditions for White's (1984) Central Limit Theorem 5.15, $E(W_j\Gamma_j) = E(CH_K^*) = 0$, and S_j is independent and identically distributed with $0 < Pr(S_j = 1) < 1$.

Let

$$V_j \equiv \left\{\underset{K\to\infty}{plim}\sqrt{K}E\left(CH_K^* W_j \mid \Gamma_1,...\Gamma_K\right)\right\}.$$

For each j, $E(V_j) < \infty$, the sequence $\{\Gamma_j V_j\}$ satisfies the mixing conditions specified in McLeish's (1975) law of large numbers, and

$$\underset{K\to\infty}{plim}\sup_{j}\left|\Gamma_j\left(V_j - \sqrt{K}E\left(CH_K^* W_j \mid \Gamma_1,...\Gamma_K\right)\right)\right| = 0.$$

Define ϕ_{1K} and ϕ_{2K} such that conditional on $S_1,...,S_K,\Gamma_1,...,\Gamma_K$,

$$Proj\left(CH_K^* \mid \frac{1}{\sqrt{K}}\sum_{j=1}^{K} S_j W_j \Gamma_j, \frac{1}{\sqrt{K}}\sum_{j=1}^{K}\left(1-S_j\right) W_j \Gamma_j\right)$$
$$= \phi_{1K}\frac{1}{\sqrt{K}}\sum_{j=1}^{K} S_j W_j \Gamma_j + \phi_{2K}\frac{1}{\sqrt{K}}\sum_{j=1}^{K}\left(1-S_j\right) W_j \Gamma_j.$$

Then as K gets large, $\left(\phi_{1K} - \phi_{2K}\right)$ converges in probability to zero.

(Proof in Appendix A.1)

One nonstandard assumption in the theorem is that

$$V_j \equiv \left\{ \operatorname*{plim}_{K\to\infty} \sqrt{K} E\left(CH_K^* W_j \mid \Gamma_1, \ldots \Gamma_K\right) \right\}$$

is well behaved. To motivate that assumption suppose that CH_K^* is treated symmetrically with $(Y_K - \alpha CH_K)$ so that

$$CH_K^* = \frac{1}{\sqrt{K}} \sum_{j=1}^{K} W_j \beta_j \tag{3.10}$$

where $E(W_j \beta_j) = 0$ and the sequence $\{W_j \Gamma_j\}$ is stationary. Under standard stationarity assumptions about W_j and $\{\Gamma_j, \beta_j\}$ this will satisfy the conditions since

$$\begin{aligned} \operatorname*{plim}_{K\to\infty} \sqrt{K} E\left(CH_K^* W_{j_1} \mid \Gamma_1, \ldots \Gamma_K\right) &= \operatorname*{plim}_{K\to\infty} E\left(\sum_{j_2=1}^{K} W_j W_{j_2} \beta_{j_2} \mid \Gamma_1, \ldots \Gamma_K\right) \\ &= E\left(\sum_{j_2=1}^{\infty} W_j W_{j_2} \beta_{j_2} \mid \Gamma_1, \Gamma_2, \ldots\right). \end{aligned}$$

Conditional on a sequence of Γ_j this will be finite. Since Γ_j is random, V_j will be a random variable. It is then straightforward to provide conditions about $\{\Gamma_j, \beta_j\}$ under which $\{\Gamma_j V_j\}$ will satisfy the conditions for the law of large numbers.

3.2 Structural Models of School Choice and Condition 1

A conventional path to identification of causal effects in the presence of endogeneous variables is through the use of an economic model as a source of informed restrictions. Here we digress briefly to show that this kind of approach can also deliver restrictions like Condition 1. Suppose that Catholic school attendance depends on X and ε only through Y^*. In addition, CH^* may depend on some additional unobserved variables that are unrelated to X and ε. In this case, the equation for CH^* would take the form:

$$CH^* = a_1 \left(Y - \alpha CH\right) \; + \varsigma, \tag{3.11}$$

where ς is uncorrelated with X and ε. Combining these equations one obtains

$$CH^* = \phi_c X' \gamma + \phi_c \varepsilon + \varsigma, \tag{3.12}$$

where $\phi_c = a_1$, and Condition 1 is satisfied. However, 3.11 is much stronger than Condition 1 because it implies CH^* is linear with coefficient $\beta_j = \phi_c \gamma_j$ for all j.

The above model might be a plausible approximation of the decision making process of the schools, parents, and children in situations in which schools are oversubscribed and select students to maximize outcomes such as achievement or college attendance. Many Catholic high schools give admissions tests and base decisions in part on the results, so the criterion of the high schools is partly related to 10th grade or 12th grade test performance. But particular elements of X may influence CH^* quite differently from the way in which they influence the secondary school outcomes.[20] Our point is simply to establish that structural models of school choice and outcomes may also lead to Condition 1. Our model of the data generation process is sufficient for Condition 1, it is not necessary.

3.3 Identification Based on Condition 1

In this section we show how one can use Condition 1 to obtain point identification of the model. First, we will strengthen Condition 1 to account for the fact that the index $X'\gamma$ is not identified unless $E(\varepsilon|X) = 0$ even if α is known. Then we show how the modified condition can help solve the identification problem.

A problem that arises is that Condition 1 is not operational because γ is not identified unless $E(\varepsilon|X) = 0$. Mean independence of ε and X is maintained in virtually all studies of selection problems, because without it, α is not identified even if one has a valid exclusion restriction.[21] Our discussion of how the observables are arrived out makes clear that it is hard to justify in most settings, including ours. If the observables are correlated with one another, as in most applications, then observed and unobserved variables are likely to be correlated.

Assume that the conditional expectation is linear and define $\hat{\gamma}$ and $\hat{\varepsilon}$ to be the slope vector and error term of the "reduced form"

$$\begin{aligned} E\left(Y^* - \alpha CH \mid X\right) + \varepsilon &\equiv X\hat{\gamma} \\ Y^* - \alpha CH &= X\hat{\gamma} + \hat{\varepsilon}. \end{aligned}$$

In appendix A.2 we consider the closely related condition

[20]For example, the relative effects of specific variables such as religion, race, parental education, and the ability and motivation of the child on sector choice and outcomes may be different. Allowing the effects of a subset of the observed variables to enter freely into (3.11) may not be sufficient and one would require a priori information about which variables to enter. The implicit restrictions on the unobservables embodied in (3.11) also pose a problem, since whether or not a student graduates from high school or attends college will be influenced by many factors that are determined after the child decides whether to attend a Catholic high school.

[21]The exception is when the instrument is uncorrelated with X as well as ε, as when the instrument is randomly assigned in an experimental setting.

Condition 1′

$$Proj(CH^* | X'\widehat{\gamma}, \widehat{\varepsilon}) = \phi_0 + \phi_c X'\hat{\gamma} + \phi_c \hat{\varepsilon}.$$

We consider the case in which CH^* is linear as defined in (3.10) and consider our data generation process defined above. We show that in this case

$$\frac{\sum_{\ell=-\infty}^{\infty} E\left(W_j W_{j-\ell}\right) E\left(\beta_j \gamma_{j-\ell}\right)}{\sum_{\ell=-\infty}^{\infty} E\left(W_j W_{j-\ell}\right) E\left(\gamma_j \gamma_{j-\ell}\right)} = \frac{\sum_{\ell=-\infty}^{\infty} E\left(\widetilde{W_j}\widetilde{W_{j-\ell}}\right) E\left(\beta_j \gamma_{j-\ell}\right)}{\sum_{\ell=-\infty}^{\infty} E\left(\widetilde{W_j}\widetilde{W_{j-\ell}}\right) E\left(\gamma_j \gamma_{j-\ell}\right)}, \tag{3.13}$$

is a sufficient condition for Condition 1′ where $\tilde{W}_j$ is the component of W_j that is orthogonal to X. This condition will hold under that standard assumption $E(\varepsilon \mid X) = 0$, in which case $\gamma = \hat{\gamma}$ and $\varepsilon = \hat{\varepsilon}$ implying that Condition 1 and Condition 1′ are identical. However, $E(\varepsilon \mid X) = 0$ is not necessary for (3.13). For example, (3.13) will also hold if $E\left(\beta_j \gamma_{j-\ell}\right)$ is proportional to $E\left(\gamma_j \gamma_{j-\ell}\right)$ regardless of the correlations among the W_j. Furthermore, in an informal Monte Carlo analysis not reported, we did not obtain large biases even when the unobservables were correlated with the observables in the original data generating process, which provides some additional reassurance.

We are now ready to discuss identification. Model (3.1) developed above is linear; however, in studying identification we want to isolate the contribution of Condition **1** or **1′** from the role of linearity or large sample properties (e.g., normality of ε is implied by our model as the number of factors gets large).[22] We only want to rely on an assumption about the relationship between unobservables and observables rather than all of the implications of the model.[23] We also wish to avoid some of the complications that arise in studying nonparametric identification of discrete choice models. Consequently, we study identification of α using the familiar "treatment effect" model without exclusion restrictions:

[22] For example, as long as the probability of going to a Catholic school is nonlinear, linearity of g in (3.11) below is sufficient for identification of α and one does not need an exclusion restriction. The propensity score could be used as an instrument.

[23] The use of a subset of restrictions implied by a model for identification is common in applied work. For example, consider a standard linear model with one endogenous variable such as

$$Y = \alpha CH + X'\gamma + \varepsilon,$$

where we are concerned about endogeneity of CH but not X. One could always use nonlinear functions of covariates in X as instruments, but this is not deemed appropriate. We are taking a similar approach here in that we do not want identification to come from the linearity assumption, but rather from the relationship between observables and unobservables.

$$CH = 1(CH^* \geq 0) \tag{3.14}$$

$$Y = \alpha CH + g(X) + \varepsilon \tag{3.15}$$

$$E(\varepsilon \mid X) = 0 \tag{3.16}$$

where $E(\varepsilon \mid X) = 0$. The econometrician observes (X, CH, Y), but not ε or the latent variable CH^*.[24]

It is well known that (3.14)-(3.15) is not nonparametrically identified without an exclusion restriction. Even if $E(\varepsilon \mid X) = 0$, we are essentially one parameter (or one equation) short of identification. This result suggests that one more restriction on this set of equations may deliver identification of α. We now show that prior information about how the observables are chosen can sometimes suffice. It should be understood that if $E(\varepsilon \mid X) \neq 0$ we require that a condition analogous to Condition 1′ holds. However, to avoid further notation we leave implicit the fact that in this case we could redefine $g(X)$ to be $E(Y^* - \alpha CH|X)$ and ε to be $(Y^* - \alpha CH) - E(Y^* - \alpha CH|X)$ and work with the analog to Condition 1′ rather than the analog to Condition 1.

In the notation of (3.14)-(3.15) Condition 1 (or Condition 1′ after redefinition of $g(X)$ and ε) is

Condition 1″

$$Proj(CH^*|g(X), \varepsilon) = \phi_0 + \phi_c g(X) + \phi_c \varepsilon.$$

Since we are assuming that $E(\varepsilon \mid X) = 0$ this is equivalent to the condition.

$$\frac{cov(CH^*, g(X))}{var(g(X))} = \frac{cov(CH^*, \varepsilon)}{var(\varepsilon)}. \tag{3.17}$$

It turns out that Condition 1″ sometimes delivers point identification and always restricts the model so that the solutions α^* for α are the roots of a cubic.

Theorem 2 *In the selection model (3.14)-(3.15)) let α be the true value of the treatment effect. Under Condition 2, in the data we can identify a set $\mathcal{A}$ of which α is a member. Define $p(X)$ as the propensity score $p(X) \equiv \Pr(CH = 1 \mid X)$. The elements α^* of $\mathcal{A}$ are*

[24] At this point we abstract from most of the recent literature on program evaluation by assuming that α does not vary across individuals. We discuss how one might extend our model into this framework in section 3.4.3.

roots of the cubic

$$
\begin{aligned}
0 = (\alpha-\alpha^*)^3 & \left[\frac{var(CH-p(X))}{var(\varepsilon)}\frac{cov(CH^*,p(X))}{var(g(X))} - \frac{var(p(X))}{var(g(X))}\frac{cov\left(CH^*,CH-p(X)\right)}{var(\varepsilon)}\right] \\
+(\alpha-\alpha^*)^2 & \left[\phi\frac{var(CH-p(X))}{var(\varepsilon)} + 2\frac{cov(\varepsilon,CH-p(X))}{var(\varepsilon)}\frac{cov(CH^*,p(X))}{var(g(X))}\right. \\
& \left. -\phi\frac{var(p(X))}{var(g(X))} - 2\frac{cov(g(X),p(X))}{var(g(X))}\frac{cov\left(CH^*,CH-p(X)\right)}{var(\varepsilon)}\right] \\
+(\alpha-\alpha^*) & \left[\frac{cov(CH^*,p(X))}{var(g(X))} + 2\phi\frac{cov(\varepsilon,CH-p(X))}{var(\varepsilon)}\right. \\
& \left. -\frac{cov\left(CH^*,CH-p(X)\right)}{var(\varepsilon)} - 2\phi\frac{cov(g(X),p(X))}{var(g(X))}\right]
\end{aligned}
$$

(Proof in Appendix A.3)

Since there is no constant term, $\alpha^* = \alpha$ is one root of the cubic. Except for pathological cases, there will be either no other real roots, or two others.[25]

To understand why Condition 1$''$ does not always yield point identification, note that in (3.16) the denominators, $var(g(X))$ and $var(\varepsilon)$, are not identified without knowledge of α. In particular, defining $(\alpha^*, g^*, \varepsilon^*)$ to be an alternative possibility for (α, g, ε), one may write $var(g^*(X))$ as

$$
\begin{aligned}
(3.18) \qquad var(g^*(X)) &= var\left(g(X) + (\alpha-\alpha^*)\,p(X)\right) \\
&= var(g(X)) + 2\,(\alpha-\alpha^*)\,cov(g(X),p(X)) + (\alpha-\alpha^*)^2\,var(p(X)).
\end{aligned}
$$

Equation (3.16) may be rewritten as

$$cov(b(X), g^*(X))var(\varepsilon^*) = cov(u, \varepsilon^*)var(g^*(X)).$$

The right hand side is the product of $var(g^*(X))$, which is quadratic in $(\alpha-\alpha^*)$, and $cov(u,\varepsilon^*)$, which is linear in $(\alpha-\alpha^*)$. This yields a cubic.

It is not clear how much we should worry about this potential problem even when one is using the condition for point identification. Consider equation (3.17). We suspect that in typical applications, the contribution of $(\alpha^*-\alpha)p(X)$ to the variance of $g^*(X)$ will be small relative to $var(g(X))$ when α^* remains within a reasonable range. In this case the other two roots are not worrisome since they involve changes in $var(g^*(X))$ outside the range of plausibility. In our empirical work we have found that $var(g^*(X))$ is insensitive to reasonable values of α^*, but the question of whether this is true in most applications can only be answered through empirical implementation.

[25] If all three coefficients of the cubic are 0, there are infinitely many solutions. If the cubic is tangent to 0, there can be two roots. While both of these cases are possible, they are very special.

3.4 Extensions

3.4.1 Continuous Endogenous Variables

The discussion in the previous subsection focused on a model such as Catholic schooling in which CH is binary and the restriction applies to the underlying latent variable. However, the link between CH and CH^* in the theory section is not restricted to $CH = 1(CH^* > 0)$. Many potential applications of the idea involve continuous endogenous variables. We maintain the model

$$Y^* = \alpha CH + g(X) + \varepsilon$$

but no longer require that CH be binary but instead assume that

$$CH = CH^*.$$

Define

$$\begin{aligned} b(X) &= E(CH \mid X) \\ u &= CH - b(X). \end{aligned}$$

In this case we obtain a stronger identification result using Condition 1″ and one additional assumption:

Theorem 3 *Assume Condition* 1″ *and that*

$$\frac{var(u)}{var(b(X))} \neq \frac{var(\varepsilon)}{var(g(X))}.$$

We can identify the set $\mathcal{A}$ *which includes two values, the true* α *and* $\alpha + \frac{var(\varepsilon)}{cov(u,\varepsilon)}$.

(Proof in Appendix A.4)

Although there are two roots, this result is very useful. In most cases in which an applied researcher is worried about the bias in a regression type estimator, he or she has a strong prior about the sign of the bias, which is the sign of $cov(u, \varepsilon)$. Imposing an assumption about the sign of $cov(u, \varepsilon)$ on the data delivers point identification; if one imposes that $cov(u, \varepsilon)$ is positive (negative), then the smaller (larger) of the two elements in $\mathcal{A}$ is the true value.

3.4.2 Using an Invalid Instrumental Variable

The results above extend to the case in which the researcher works with an invalid instrumental variable Z that is correlated with the error term in the outcome equation. For simplicity we focus on the linear case and maintain our notation

$$Y^* = \alpha CH + X'\gamma + \varepsilon,$$

where X is observable but ε is not. CH is a binary variable in our case but could also be continuous. We assume that X is uncorrelated with ε, but CH is potentially endogenous and thus correlated with ε. We assume our instrument Z does not influence Y directly, but is correlated with CH. However, Z is not necessarily a valid instrument because it might be correlated with ε. We extend the idea of using the data generation process for identification by showing that if the relationship between $X'\gamma$ and Z is similar to the relationship between ε and Z, then we can sometimes obtain identification.

Define β and π to come from least squares projection such that

$$\text{Proj}\,(Z \mid X) = X'\pi, \tag{3.19}$$

$$\text{Proj}\,(CH \mid X, Z) = X'\beta + \lambda Z, \tag{3.20}$$

and define v and u as the residuals of these regressions, so that

$$Z = X'\pi + v, \tag{3.21}$$

$$CH = X'\beta + \lambda Z + u. \tag{3.22}$$

Consider the regression of Y onto the predicted value $\text{Proj}(CH \mid X, Z)$ and X. The coefficient on the predicted value in this regression converges to

$$\widehat{\alpha} = \alpha + \frac{cov\,(v, \varepsilon)}{\lambda var(v)}.$$

If Z is a valid instrument, v would be uncorrelated with ε and $\widehat{\alpha}$ would equal α.

Can our assumption about the relationship between unobservables and observables pin down this bias? In general it helps but may or may not be sufficient for point identification. As above,

Condition 3

$$\frac{cov(X'\pi, X'\gamma)}{var(X'\gamma)} = \frac{cov(v, \varepsilon)}{var(\varepsilon)}$$

restricts the solutions α^* to be the solutions of a cubic equation, one of which is α. This means that typically there are either three solutions (i.e. three values of α^* that we can not distinguish between) or there is a unique solution that equals α. The details are in Appendix A.5.

3.4.3 Heterogeneity in the Effects of Catholic Schools

Our analysis extends in a natural way to the case of heterogeneity in the effect of attending Catholic school. Let Y_{ch}^* and Y_p^* be the outcomes conditional on choice of Catholic high school and public high school, respectively, for a given student. As above let W be the set of covariates that fully determine Y_{ch}^* and Y_p^* and let X be the observed components of W. The heterogeneous effects model may be written as

$$Y_{ch}^* = g_{ch}(X) + \varepsilon_{ch} \tag{3.23}$$

$$Y_p^* = g_p(X) + \varepsilon_p \tag{3.24}$$

where Y_{ch}^* is observed if $CH^* \geq 0$, in which case $CH = 1$, and Y_p is observed otherwise. Our previous specification is a special case of this model in which $g_{ch}(X) - g_p(X)$ is constant and $\varepsilon_{ch} = \varepsilon_p$. Treating the data generation processes for Y_{ch}^* and Y_p^* as equivalent to the data generation process for $Y^* - \alpha CH$ above and applying Theorem 1, we obtain the restrictions

$$\begin{aligned}
\text{Proj}(CH^*|g_{ch}(X), \varepsilon_{ch}) &= \phi_{ch} g_{ch}(X) + \phi_{ch}\varepsilon_{ch} \\
\text{Proj}(CH^*|g_p(X), \varepsilon_p) &= \phi_p g_p(X) + \phi_p \varepsilon_p.
\end{aligned}$$

These restrictions can be used to help identify the model in a way that is directly analogous to our use of Condition 1 to identify the model in the homogeneous effects case. We conjecture that if the components of X are a random subset of the components of W and if the number of elements of W and X are large, then the joint distribution of $(b(X), g_{ch}(X), g_p(X))$ is the same as the joint distribution of $(u, \varepsilon_{ch}, \varepsilon_p)$ up to a scale parameter that depends on the fraction of elements of W that are observed. If this is the case, then a nonparametric or semiparametric analysis may be possible, at least in theory. We leave a full analysis of this case to future work.

3.5 Relevance to the Study of the Catholic School Effect and Bounding the Results

We have data on a broad set of family background measures, teacher evaluations, test scores, grades, and behavioral outcomes in eighth grade, as well as measures of proximity

to a Catholic high school. These measures have substantial explanatory power for the outcomes that we examine, and a large number of the variables play a role, particularly in the case of high school graduation and college attendance. Once we restrict the sample to Catholic eighth graders and condition on Catholic religion and distance from a Catholic high school, a broad set of variables make minor contributions to the probability of Catholic high school attendance. The relatively large number and wide variety of observables that enter into our problem suggests that the observables may provide a useful guide to the unobservables.

However, our "random selection of observables" model is not to be taken literally. There are in fact strong reasons to expect that the relationship between the unobservables will be weaker than the relationship between the observables. The most important is that shocks that occur after eighth grade are excluded from X.[26] These will influence high school outcomes but not the probability of starting a Catholic high school.[27] To see this, return to the linear index formulation (3.1) and augment the model by rewriting ε as $\varepsilon = \varepsilon_1 + \varepsilon_2$, where ε_1 is determined during eighth grade and ε_2 is the independent innovation in the error term that is determined during high school. Since the observables X and the unobservable u are determined during eighth grade, we can impose our data generation condition on the variables determined prior to high school, in which case

$$\frac{cov(CH^*, X'\gamma)}{var(X'\gamma)} = \frac{cov(CH^*, \varepsilon_1)}{var(\varepsilon_1)} \geq \frac{cov(CH^*, \varepsilon_1 + \varepsilon_2)}{var(\varepsilon_1 + \varepsilon_2)}. \tag{3.25}$$

We will show how this condition can allow us to identify a set of permissible values of α. We use the following alternative to Condition 1″.

Condition 4

$$\begin{aligned} Proj(CH^*|g(X), \varepsilon) &= \phi_0 + \phi_{c_1} g(X) + \phi_{c_2}\varepsilon, \\ 0 &\leq \phi_{c_2} \leq \phi_{c_1}. \end{aligned} \tag{3.26}$$

[26]A second reason is that it is quite possible that among Catholic eighth graders the decision to attend Catholic high school is influenced by highly idiosyncratic preference variables, such as the religious beliefs of the parents, whether close friends of the student are going to Catholic high school, whether the parents attended Catholic high school, the influence of a particular eighth grade teacher or minister, the quality of the school band or sports teams, the logistics of getting to and from the school, or transitory variation in the finances of the family. We suspect that many of these factors have coefficients in the outcome and school choice equations that are quite different with some being positively correlated with outcomes and others negatively correlated. This would also lead selection on the unobservables to be weaker than selection on the observables.

[27]In the case of the 10th and 12th grade test scores, ε will also reflect variability in test performance on a particular day, which presumably has nothing to do with the decision to start Catholic high school.

Theorem 4 *For any value α^* there is a unique value of g and ε consistent with the selection model (3.14)-(3.15). Define g_{α^*} and ε_{α^*} as these values. We can identify the set*

$$\mathcal{A} = \left\{ \alpha^* \in \Re : 0 \leq \frac{cov(CH^*, \varepsilon_{\alpha^*})}{var(\varepsilon_{\alpha^*}))} \leq \frac{cov(CH^*, g_{\alpha^*}(X))}{var(g_{\alpha^*}(X))} \right\}$$

Under Condition 4 the true value α is a member of this set.

Treating the restriction as a bound actually simplifies the identification procedure somewhat. In this case all we could ever hope to identify is a set of values of α that are consistent with (3.24). The identification proof is constructive in suggesting a manner for testing hypotheses about α (or constructing confidence intervals). Following the theorem for any potential value α_0 we can construct g_{α_0} and ε_{α_0} and then test whether the restriction holds for those values.

After experimenting with our data we find that the upper bound on α occurs when one assumes that $\frac{cov(CH^*,\varepsilon)}{var(\varepsilon)} = 0$ and the lower bound comes when one assumes that $\frac{cov(CH^*,X'\gamma)}{var(X'\gamma)} = \frac{cov(CH^*,\varepsilon)}{var(\varepsilon)}$. Thus, in the empirical work below, we interpret estimates of α that incorporate Condition 1$''$ as a lower bound for α and single equation estimates with CH treated as exogenous as an upper bound. This simplifies the analysis substantially. If the lower bound estimates point to a substantial Catholic school effect, we interpret this as strong evidence in favor of such an effect. As it turns out, for some outcomes and samples, such as high school graduation, the single equation estimates are so large relative to the degree of selection on the observables that the lower bound estimate is still substantial. In other cases, even an amount of selection on the unobservables that is small relative to the selection on the observables is sufficient to eliminate the entire Catholic School effect.

4 Adjusting for Selection Bias Using Selection on the Observables

4.1 Using the indices of Observables in the School Choice and Outcome Equations as a value for ρ.

We now return to the bivariate probit model given by (2.1), (2.2), and (2.3). We argue in the theory section that Condition 1 represents an extreme assumption about the degree of selection and that the true amount of selection is somewhere between independence and Condition 1 which we formalized as Condition 4. In practice we have found the model

to be monotonic so the highest value of the treatment effect α occurs at $\rho = 0$ while the minimum value occurs when condition 1 is binding, so we focus on estimating the model while imposing condition 1 and interpret the result as a low bound on α. In the bivariate probit case, Condition $1'$ may be re-written[28] as

$$\rho = \frac{Cov(X_i'\beta, X_i'\gamma)}{Var(X_i'\gamma)}. \tag{4.27}$$

We take two approaches to estimating the model while imposing this restriction. The first is to use the Catholic eighth grade sample directly and maximize the likelihood subject to (4.26). To improve precision of the estimates of α and as a check on the robustness of the results, we employ an alternative method using information contained in the public 8th grade sample. We partition X and γ into the subvectors $\{X_1, X_{2..}X_G\}$ and $\{\gamma_1, \gamma_2, ..., \gamma_G\}$ consisting of variables and parameters that fall into the same category. In practice, G is 6. We estimate γ on the public 8th grade sample on the grounds that very few such students go to Catholic school, and so selectivity will not influence the estimates of γ even though the mean of the error term may be different for this sample. We then assume that the values of γ are the same for students from Catholic and public 8th grades, up to a proportionality factor for each subvector. Note that the univariate models reported above for the full sample implicitly assume that γ does not depend on the sector of the 8th grade. We are relaxing that assumption to some extent.[29]

In Table 8, we present estimates using methods 1 and 2 to impose the restriction, focusing on the results for the Catholic eighth grade sample. The estimate of ρ is 0.24, the estimate of α is 0.59 (0.33), which implies an effect of 0.07 on the probability of high school graduation. Consequently, even with the extreme assumption imposed, there is evidence of a large positive effect of attending Catholic high school on high school graduation.

The results for college attendance follow a similar pattern. The regression relationship between the indices of observables that determine CH and college attendance is sufficiently strong that imposing the restriction leads to a reduction in the estimated effect of Catholic schooling. The point estimate of 0.07 is substantial, although it is not statistically significant given our sample size.

[28] Keep in mind that in the binary probit the variances of ε and u are normalized to 1.

[29] The restrictions pass with a p-value of .12 in the high school graduation case, but fail with a p-value of .03 in the college attendance case, largely because the restriction fails for the coefficients on distance from Catholic school. Details are in Table 8 note 4.

When we use method 2, we obtain qualitatively similar results that point to an even larger effect of Catholic schooling on high school graduation—in this specification, ρ is only 0.09 and the estimate of the effect on the graduate probability is 0.09. The college effect is only 0.02. The restrictions on γ restrictions pass with a p-value of .12 in the high school graduation case, but fail with a p-value of .03 in the college attendance case, so perhaps the method 2 results for college attendance should be discounted. Details are in Table 8 note 4.[30]

4.2 The Relative Amount of Selection on Unobservables Required to Eliminate the Catholic School Effect

In this section we provide a different, more informal way to use information about selection on the observables as a guide to selection on the unobservables that permits us to use the Catholic high school indicator directly. Consider the alternative restriction,

Condition 5

$$\frac{E(\epsilon_i \mid CH_i = 1) - E(\epsilon_i \mid CH_i = 0)}{var(\epsilon_i)} = \frac{E(X_i'\gamma \mid CH_i = 1) - E(X_i'\gamma \mid CH_i = 0)}{var(X_i'\gamma)}.$$

This condition implies that the relationship between Catholic high school and the location of the distribution of the index of the observables that determine outcomes and the index of unobservables is the same, after adjusting for differences in the dispersion of these distributions. We justify this condition informally in Appendix A.. It will hold under the assumptions that lead to Conditions 1 and 2. However, it requires that X be uncorrelated with ε.

For reasons discussed earlier, the standardized difference in the mean of the unobservables that determine is Y is likely to be smaller than the standardized difference in the index of observables, because many post-eighth grade factors influence the outcomes, and many hard-to-observe factors influence high school choice. One way to gauge the strength

[30]For completeness, we also present estimates of α and ρ from an unrestricted bivariate probit on the Catholic school sample. The estimates α and ρ for high school graduation are quite close to the restricted estimates, although this is a matter of luck in view of the large standard errors. In the college attendance case we obtain a large and implausibly negative value of ρ equal to -0.52 and an implausibly large but very imprecise estimate of α equal to 1.18. As Grogger and Neal (1999) note, a finding of negative selection on unobservables based on bivariate probit models is not uncommon in the Catholic schools literature and is sometimes attributed to pre-existing differences in student motivation or discipline that are poorly captured in existing data sets. We are very skeptical of this interpretation because the rich set of 8th grade student behavior measures in NELS:88 point to positive selection more or less across the board. Our view is that without exclusion restrictions or a restriction such as Condition 2, identification of α and ρ is very tenuous. We place little weight on the unrestricted estimates.

of the evidence for a Catholic school effect is to see how much of it would remain if Condition 4 were true, and to ask how large the ratio on the left would have to be relative to the ratio on the right to eliminate the entire Catholic school effect. An advantage of this approach is that we do not have to simultaneously estimate the parameters of the CH and Y equations subject to (4.26). Consequently, we are able to use the full control set used in columns 3 and 6 of Tables 3 and 4. In Altonji, Elder, and Taber (2001) we expand on this approach by showing how it can be used to evaluate an instrumental variable.

To gauge the role of selection bias in a simple way we ignore the fact that Y_i is estimated by a probit and treat α as if it were estimated by a regression of the latent variable $Y_i^* = X_i'\gamma + \alpha CH_i + \varepsilon_i$ on X_i and CH_i. Let $X'\beta$ and $\widetilde{CH}_i$ represent the predicted value and residuals of a regression of CH_i on X_i so that $CH_i = X'\beta + \widetilde{CH}_i$. Then,

$$Y_i^* = X_i'[\gamma + \alpha\beta] + \alpha\widetilde{CH}_i + \varepsilon_i.$$

Assuming that the bias in a probit is close to the bias in OLS applied to the above model and using the fact that $\widetilde{CH}_i$ is orthogonal to X_i leads to

$$\begin{aligned}
\text{plim}\ \widehat{\alpha} &\simeq \alpha + \frac{cov(\widetilde{CH}_i, \varepsilon)}{var\left(\widetilde{CH}_i\right)} \\
&= \alpha + \frac{var\left(CH_i\right)}{var\left(\widetilde{CH}_i\right)}\left[E(\epsilon_i \mid CH_i = 1) - E(\epsilon_i \mid CH_i = 0)\right].
\end{aligned}$$

Thus, subject to Condition 4 one can estimate $E(X_i'\gamma \mid CH_i = 1) - E(X_i'\gamma \mid CH_i = 0)$ and estimate the magnitude of this bias.

We use the single equation estimates of α obtained under the assumption that Catholic schooling is exogenous in the outcome equation. A problem with using Condition 4 is that bias in α will lead to bias in the estimates of γ, which are required to evaluate the left hand side of the equation. We believe that in many applications this problem will be minor. However, as a robustness check we try three alternative ways to obtain γ. The first method is use the γ from the public eighth grade sample to form the index $X_i'\gamma$ for each Catholic 8th grade student. The results are reported in the first row of Table 9. In the case of high school graduation, the estimate of $(E(X_i'\gamma \mid CH_i = 1) - E(X_i'\gamma \mid CH_i = 0))\ /Var(X_i'\gamma)$ is 0.30. That is, the mean/variance of the probit index of X variables that determine high school graduation is 0.28 higher for those who attend Catholic high school than for those who do not. Since the variance of ε_i is 1.00, the implied estimate of $E(\epsilon_i \mid CH_i = 1) - E(\epsilon_i \mid CH_i = 0)$ if Condition 4 holds is 0.30 (row 1, column 3). Multiplying by $var\left(CH_i\right)/var\left(\widetilde{CH}_i\right)$ yields a bias of 0.37, while the estimate of the α is 1.03. The last column of the table reports

that the ratio $\widehat{\alpha}/[\frac{var(CH_i)}{var(\widetilde{CH_i})}(E(\epsilon_i \mid CH_i = 1) - E(\epsilon_i \mid CH_i = 0))] = (1.03/.37) = 2.78$. That is, the normalized shift in the distribution of the unobservables would have to be 2.78 times as large as the shift in the observables to explain away the entire Catholic school effect. This seems highly unlikely.

The second row of Table 9 reports the results when the left hand side of Condition 4 is evaluating using the estimate of γ obtained from the single equation probit estimate of the high school graduation equation on the Catholic school sample. The third row uses the estimate of γ when α is constrained to be 0. For these methods, the implied ratios are 4.29 and 3.55. The results in Table 9 suggest that a substantial part of the effect of CH on high school graduation is real.

For college attendance the ratios range between 1.30 and 2.03 depending on how we estimate γ (rows 4, 5, and 6). Since the ratio of selection on unobservables relative to selection on observables is likely to be less than 1, part of the Catholic school effect on college graduation is probably real.

Table 10 presents 10th and 12th grade test score results using the same methodology described above. The coefficient on CH_i has a positive and statistically significant coefficient only in the case of 12th grade math scores. However, this effect is small (1.14) and would be almost completely eliminated assuming the upper bound Condition 4 holds. Even if selection on unobservables is only one half as strong as that on observables, the effect of Catholic schooling would be negligible. Given the weak evidence from the univariate models and the likelihood of some positive bias, we conclude that Catholic high school probably has little effect on test scores.

5 Results by minority status and urbanicity

A number of studies, including Evans and Schwab (1995), Neal (1997), and Grogger and Neal (1999) using NELS:88 have found much stronger effects of Catholic schooling for minority students in urban areas than for other students. Table 2 reports differences in the means of outcomes and control variables, by high school type, for all urban minority students and for urban minority students who attended Catholic eighth grades. Note that 54 of the 56 minority students who attended Catholic high school came from Catholic eighth grades. Only 15 of the 700 urban minority students in public 10th grades came from Catholic 8th grades, which is too few observations to support an analysis on the Catholic eighth grade subsample. In the full urban minority sample the control variables provide

evidence of strong positive selection into Catholic high schools. The gaps in mother's education and father's education are 0.66 years and 1.69 years, respectively. The gap in the log of family income is 0.83. There are also very large discrepancies in the base year measures of parental expectations for schooling and student expectations for schooling and white-collar work, large gaps in the eighth-grade behavioral measures, and gaps of 6.49 and 3.28 in the eighth grade reading and math tests, respectively. Since there is more selection on observable variables for this subsample it is quite plausible that there could be more selection on unobservables as well and that this could explain the large measured Catholic schooling effects.

In Table 5 we report models of the high school graduation probability estimated using the urban sample of white students as well as the urban sample of minorities. All of the regression models include our full set of controls. For the minority sample, the average derivative implied by the probit estimate of the Catholic high school effect on high school graduation is 0.191, while the linear probability model estimate is 0.133 (0.056).[31] Turning to the bottom panel of Table 5, we find a substantial effect of Catholic high school on college attendance, with estimates for the urban minority sample varying from 0.144 to 0.182 depending on the estimation methods. Consistent with previous work, the effects are generally larger for minorities than for the samples of whites. However, since there is more selection on observable variables for this subsample it seems quite plausible that there could be more selection on unobservables as well and that this could explain the large measured Catholic schooling effects.

Table 6 presents test score results for the urban minority sample. As shown in the second column of the table, we obtain negative but small and statistically insignificant estimates of the effect of Catholic schooling on both the math and reading 10th grade tests, which agrees with the analysis based on both the full NELS:88 sample and the Catholic eighth grade subsample. We obtain a coefficient of -0.19 (1.39) for the 12th grade reading score as well, and a coefficient of 1.25 (1.09) for the 12th grade math score. Evidently, most or all of the substantial Catholic high school advantage for urban minorities in test scores disappears once we control for family background and 8th grade outcomes. This result reflects the large gap in the means of the controls in favor of minorities attending Catholic high school. As one can see in the table, we obtain similar results when we add suburbanites

[31] The estimate including eighth grade school fixed effects is essentially zero, which leaves open the possibility that cross-school variation in the opportunities available to urban minority students may be responsible for the positive estimated Catholic high school effects. However, the standard error of the fixed effects estimate is quite large (.107), so one should not make too much of this result.

and extend our analysis to a pooled urban/suburban minority subsample.

We also perform a sensitivity analysis of the kind described above for the urban minority sample. Turning again to Table 7, note that the raw differential in the high school graduation probability is 0.22 and the estimate of the Catholic school effect under the assumption $\rho = 0$ is 0.176. The estimate is 0.132 when ρ is 0.2, and 0.013 when ρ is 0.5. Thus, the correlation between the unobservables would have to be in the neighborhood of 0.5, a very large correlation, for one to conclude that the true effect of Catholic schools on the graduation rates of urban minorities is 0. This value seems unreasonable.

In Table 11, we conduct an analysis involving the differences in indices of observable variables based on Condition 4. In rows 2 and 4 we form the index of selection on observables using the estimates of γ from the urban minority public 8th grade sample. For this sample under Condition 4 the implied shift in $(E(\varepsilon_i \mid CH_i = 1) - E(\varepsilon_i \mid CH_i = 0))$ is 0.56 in the case of high school graduation and 0.72 in the case of college attendance, which reflects strong selection on the observables that influence these outcomes. Still, selection on the unobservables would have to be 2.37 times as strong as selection on the observables to explain away the entire high school graduation effect. This seems very unlikely to us; the evidence suggests that a substantial part of the estimated effect of Catholic schooling on graduation would remain for this group, even if there was a high degree of sample selection bias. On the other hand, we cannot rule out the possibility that much of the effect of CH on college attendance is due to selection bias.

In Table 12 we report the results of an analysis of test scores. As we have already noted, there is little evidence that Catholic high school improves the reading scores of minorities. The table shows that in the case of 12th grade reading scores $(E(X_i'\gamma \mid CH_i = 1) - E(X_i'\gamma \mid CH_i = 0)) \,/Var(X_i'\gamma)$ is 0.090. Under Condition 4 this amount of favorable selection on the observables implies an estimate of $(E(\varepsilon_i \mid CH_i = 1) - E(\varepsilon_i \mid CH_i = 0))$ equal to 2.76. Since the point estimate of α is already negative, there is certainly no evidence that Catholic schools boost 12th grade reading scores.

In the case of 12th grade math, the point estimate of a is 1.82 and the implied estimate of $(E(\varepsilon_i \mid CH_i = 1) - E(\varepsilon_i \mid CH_i = 0))$ under Condition 4 is 1.17, and the implied ratio of selection on unobservables to selection on observables required to explain away the entire estimate of α is 0.89. Consequently, we would not rule out a small positive effect on math but overall conclude that there is not much evidence that Catholic high schools boost the test scores of urban minorities.[32]

[32] These test score findings are robust to the imputation procedures for dropouts described in Section 2.3. In contrast, Grogger and Neal (1999) find some evidence for a Catholic school effect on minority test scores

6 Conclusion

Our analysis of the Catholic school effect is guided by three premises. The first is that the exclusion restrictions used in previous studies, including Altonji, Elder and Taber (1999), do not provide a reliable means of identifying the Catholic school effect. The second premise is that in the absence of a bulletproof instrument, it is important to start with a rich set of control variables and with a group of students who do not differ dramatically by whether or not they attended Catholic high school. This leads us to focus on students from Catholic eighth grades. Focusing on Catholic eighth graders allows us to avoid concerns about lack of comparability between the tiny fraction of students from public primary schools who attend Catholic high school and other students. It also allows us to isolate the effect of Catholic high school from the effect of Catholic primary school.

The third premise is that the degree of selection on the observables is informative about selection on the unobserved characteristics. As we noted in the introduction, it is standard procedure to consider the relationship between an explanatory variable or an instrumental variable and the observed variables in the model in discussions of exogeneity. The methodological contribution of this paper is to formalize the use of such information and to provide a way to assess quantitatively the degree of selection bias. We make the theoretical point that knowledge of how the observable variables are chosen from the full set of variables can be sufficient to identify the effect of an endogenous variable. We illustrate this by establishing identification in the case in which selection on observables and unobservables is the same in the sense that unit shifts in the indices of observables and unobservables that determine the outcome have the same effect on school choice. We estimate our model subject to the restrictions imposed by equal selection. We argue that in the Catholic school case, selection on the observables is likely to be stronger than selection on the unobservables. Consequently, we use the restricted estimate as a lower bound estimate of the effect of Catholic schools and use the single equation estimates as an upper bound. We also propose an informal way to assess selectivity bias based on a measure of the ratio of selection on unobservables relative to selection on observables that would be required if one is to attribute the *entire* Catholic school effect to selection bias.

We have three main substantive findings regarding Catholic schools. First, attending

using median regression, particularly when they restore high school dropouts with missing test score data to the sample by simply assigning them 0. We have not fully investigated the source of the discrepancy, but suspect that our use of a more extensive set of control variables, our imputation process, differences in the samples used, and differences between mean and median regression all play a role.

Catholic high school substantially raises high school graduation rates. In the Catholic eighth grade sample, none of the 0.08 Catholic high school advantage in graduation rates is explained by eighth grade outcomes or family background and we obtain a lower bound estimate of 0.07 when we impose equality of selection of observables and unobservables. While estimates that treat Catholic school attendance as exogenous almost certainly overstate the effect of Catholic high schools, the degree of selection on the unobservables would have to be much stronger than the degree of selection on the observables to explain away the entire effect. We also find that the effect of Catholic school on the probability of college attendance is very large (0.15) when Catholic school attendance is treated as exogenous, but the lower bound estimates range between 0.07 and 0.02 depending on estimation details. We conclude that part of the effect of CH on college attendance is probably real, but the evidence is less strong than in high school graduation case.

Second, we find little evidence that Catholic high schools raise reading scores. In fact, most of our point estimates are negative. The single equation estimates point to a positive effect of about 0.1 standard deviations on the 12th grade math score. However, given sampling error and evidence of positive selection bias, we do not have much evidence that Catholic high schools boost test scores as well as high school graduation rates.

Third, our results for urban minorities suggest that Catholic high school attendance substantially raises the probability of high school graduation for this group. Single equation estimates of the impact on college attendance are also very large, but the degree of positive selection on the observables that determine college attendance is sufficiently large that one could not rule out selection bias as the full explanation for the Catholic school effect on college attendance. One problem is that our sample of urban minorities who attended Catholic eighth grade is not big enough to permit us to perform the analysis on the Catholic eighth grade sample. Unfortunately, in the full urban minority sample, differences by high school sector in family background characteristics and eighth grade performance are very large. The assumption that the selection on the unobservables mirrors selection on the observables results in a larger selectivity bias correction for this group. While we believe that selection on the unobservables is less strong, the evidence for a Catholic school effect on college attendance is weaker for this group. In general, we find smaller differences between urban minorities and other groups in the Catholic school effect than other recent studies.

The next step on the empirical side of the project is to examine the mechanism through which Catholic schools affect high school graduation in light of the literature on Catholic schools and the data on school characteristics and student behavior during the high school

years in NELS:88. Multivariate analysis of the effect of differences in background and eighth grade social behavior suggests that such differences are more important for graduation than for test scores (not reported). Many of the traits of Catholic schools stressed by Bryk et al (1993) and Coleman and Hoffer (1987) may work to reduce the dropout probability among low achieving students or students with behavioral problems. The more structured and communitarian environment normally found in Catholic high schools may be effective in reducing dropout rates and increasing college attendance.

There is a long agenda for future research on the econometric methods that we propose. With regard to the theoretical foundations, high priorities include additional analysis of identification in both single equation and instrumental variables settings and a full analysis of the heterogeneous effects case introduced in section 3.6. Our theoretical analysis suggests that the observables may not have much to say about bias from selection on unobservables in situations in which only a handful of variables dominate the distribution of the outcome (a situation in which structural economic model may be feasible to develop and estimate), or in which the set of observables is small. In our application, the measure of the relative degree of selection on observables and unobservables is not very sensitive to how we compute γ, the parameters of the outcome equation, and we were able to use the public 8th grade sample as a benchmark for γ in any case. However, a theoretical analysis of conditions under which bias in the estimates of γ is important would be helpful.

With regard to the art of assessing when and how to use the methods that we describe, a monte carlo analysis of how the methods perform in the context of real world examples would prove informative, particularly in those cases in which concern about identification is a first order issue. One could also do a monte carlo analysis in which one samples at random from the hundreds of 8th grade family background and student characteristics available in the NELS 88, although this would be taking too literally the idea of random inclusion of variables.

In closing, we caution against the potential for misuse of the idea using observables to draw inferences about selection bias. The conditions required for Theorem 1 imply that it is dangerous to infer too much about selection on the unobservables from selection on the observables when the observables are small in number and explanatory power or if they are unlikely to be representative of the full set of factors that determine an outcome.

Appendix A

A.1 Proof of Theorem 1

Proof. We simplify the notation by defining

$$E^K(\cdot) \equiv E(\cdot \mid S_1, , ..., S_K, \Gamma_1, ...\Gamma_K)$$

Define

$$\text{(A-1)} \qquad \begin{aligned} \phi &= \frac{\operatorname*{plim}_{K\to\infty} E^K\left(CH_K^*\left(Y_K - \alpha CH_K\right)\right)}{\operatorname*{plim}_{K\to\infty} E^K\left(\left(Y_K - \alpha CH_K\right)^2\right)} \\ &= \frac{\operatorname*{plim}_{K\to\infty} \frac{1}{K}\sum_{j=1}^K \Gamma_j \sqrt{K} E^K\left(CH_K^* W_j\right)}{\operatorname*{plim}_{K\to\infty} \frac{1}{K}\sum_{j_1=1}^K \sum_{j_2=1}^K \Gamma_{j_1}\Gamma_{j_2} E^K\left(W_{j_1} W_{j_2}\right)} \\ &= \frac{\operatorname*{plim}_{K\to\infty} \frac{1}{K}\sum_{j=1j}^K \Gamma_j V_j + \operatorname*{plim}_{K\to\infty} \frac{1}{K}\sum_{j=1j}^K \Gamma_j\left(\sqrt{K} E^K\left(CH_K^* W_j\right) - V_j\right)}{\sum_{\ell=-\infty}^{\infty} E(\Gamma_j\Gamma_{j-\ell} W_j W_{j-\ell})} \\ &= \frac{\operatorname*{plim}_{K\to\infty} \left\{\frac{1}{K}\sum_{j=1}^K E\left(W_j\Gamma_j V_j\right)\right\}}{\sum_{\ell=-\infty}^{\infty} E(\Gamma_j\Gamma_{j-\ell} W_j W_{j-\ell})}. \end{aligned}$$

The term $\frac{1}{K}\sum_{j=1j}^K \Gamma_j\left(\sqrt{K} E^K\left(CH_K^* W_j\right) - V_j\right)$ goes to zero as a result of our assumption about $\operatorname*{plim}_{K\to\infty} \sup_j \left|\Gamma_j\left(V_j - \sqrt{K} E\left(CH_K^* W_j \mid \Gamma_1, ...\Gamma_K\right)\right)\right|$. We apply the central limit theorem to $W_j\Gamma_j$ in deriving the denominator and apply the law of large numbers for the numerator. Under the assumptions of the theorem both the numerator and denominator are finite.

To simplify the exposition define

$$\Psi_K = \begin{bmatrix} \frac{1}{\sqrt{K}}\sum_{j=1}^K S_j W_j \Gamma_j \\ \frac{1}{\sqrt{K}}\sum_{j=1}^K \left(1 - S_j\right) W_j \Gamma_j \end{bmatrix}$$

By definition of the projection of interest

$$\begin{aligned} \begin{bmatrix} \phi_{1K} \\ \phi_{2K} \end{bmatrix} &= \left[E^K\left(\Psi_K \Psi_K'\right)\right]^{-1} E^K\left(\Psi_K CH_K^*\right) \\ &= \begin{bmatrix} \phi \\ \phi \end{bmatrix} + \left[E^K\left(\Psi_K \Psi_K'\right)\right]^{-1} E^K\left(\Psi_K\left(CH_K^* - \Psi_K' \begin{bmatrix} \phi \\ \phi \end{bmatrix}\right)\right). \end{aligned}$$

From the conditions in the theorem $\left(E\Psi^K\Psi^{K\prime} \mid S^K, \Gamma^K\right)$ is finite and positive definite.

To see that $E^K\left(\Psi_K\left(CH_K^* - \Psi_K'\begin{bmatrix}\phi\\\phi\end{bmatrix}\right)\right)$ converges to zero note that

$$
\begin{aligned}
&\operatorname*{plim}_{K\to\infty} E^K\left(\frac{1}{\sqrt{K}}\sum_{j=1}^{K} S_j W_j \Gamma_j\left(CH_K^* - \Psi_K'\begin{bmatrix}\phi\\\phi\end{bmatrix}\right)\right)\\
=&\operatorname*{plim}_{K\to\infty}\frac{1}{\sqrt{K}}\sum_{j=1}^{K} S_j\Gamma_j E^K\left(W_j CH_K^*\right)\\
&-\operatorname*{plim}_{K\to\infty} E^K\left(\left(\frac{1}{\sqrt{K}}\sum_{j=1}^{K} S_j W_j\Gamma_j\right)\left(\frac{1}{\sqrt{K}}\sum_{j=1}^{K} W_j\Gamma_j\phi\right)\right)\\
=&\operatorname*{plim}_{K\to\infty}\frac{1}{K}\sum_{j=1}^{K} E\left(S_j\Gamma_j V_j\right)\\
&-\operatorname*{plim}_{K\to\infty}\frac{1}{K}\sum_{j_1=1}^{K}\sum_{j_2=1}^{K} S_{j_1}\Gamma_{j_1}\Gamma_{j_2}E^K\left(W_{j_1}W_{j_2}\right)\\
=&E\left(S_j\right)\operatorname*{plim}_{K\to\infty}\left\{\frac{1}{K}\sum_{j=1}^{K} E\left(W_j\Gamma_j V_j\right)\right\} - E(S_j)\left(\sum_{\ell=-\infty}^{\infty} E(\Gamma_j\Gamma_{j-\ell}W_j W_{j-\ell})\right)\phi\\
=&0
\end{aligned}
$$

where we have used $(A-1)$. By virtually the same argument

$$\operatorname*{plim}_{K\to\infty} E^K\left(\frac{1}{\sqrt{K}}\sum_{j=1}^{K}\left(1-S_j\right)W_j\Gamma_j\left(CH_K^* - \Psi_K'\begin{bmatrix}\phi\\\phi\end{bmatrix}\right)\right) = 0.$$

Thus

$$
\begin{aligned}
\operatorname*{plim}_{K\to\infty}\{\phi_{1K}\} &= \operatorname*{plim}_{K\to\infty}\{\phi_{2K}\}\\
&= \phi
\end{aligned}
$$

■

A.2 Justification for Condition 1′

When implementing our model we have assumed that the error terms are uncorrelated with the regressors, but this is not a property of the data generation process that we defined in Theorem 1 nor is it required for Condition 1. We briefly discuss the conditions under which our assumption is consistent with the previous analysis.

As above treat the model as

$$
\begin{aligned}
CH_K^* &= \frac{1}{\sqrt{K}}\sum_{j=1}^{K} W_j\beta_j\\
Y_K^* &= \frac{1}{\sqrt{K}}\sum_{j=1}^{K} W_j\gamma_j
\end{aligned}
$$

where we have incorporated αCH_K into Y_K^* to simplify the notation. In this section we use notation that differs somewhat from the text. Throughout this section we use "hats" to define the predicted value from a least square regression of a variable onto the observable covariates in W and "tildes" to denote the residual from that regression. For example $Y_K^* = \widehat{Y_K^*} + \widetilde{Y_K^*}$ where $\widehat{Y_K^*}$ is the linear prediction from a regression of Y_K^* on the observables. In the notation of the text

$$\begin{aligned} \widehat{Y_K^*} &= X'\widehat{\gamma} \\ \widetilde{Y_K^*} &= \widehat{\varepsilon} \end{aligned}$$

Furthermore we simplify the notation by dropping the K subscript when we mean the probability limit of the variable so $Y^* \equiv \text{plim}\{Y_K^*\}$.

In this notation Condition $1'$ can be written as

$$\frac{cov(CH^*, \widehat{Y^*})}{var(\widehat{Y^*})} = \frac{cov(CH^*, \widetilde{Y^*})}{var(\widetilde{Y^*})}. \tag{A-2}$$

When does our data generation process yield (A-2)? It is straightforward to verify that (A-2) is equivalent to

$$\frac{cov(CH^*, Y^*)}{var(Y^*)} = \frac{cov(\widetilde{CH^*}, \widetilde{Y^*})}{var(\widetilde{Y^*})}.$$

Since CH_K^* and Y_K^* are linear we can also write

$$\begin{aligned} \widetilde{CH_K^*} &= \frac{1}{\sqrt{K}} \sum_{j=1}^{K} \widetilde{W}_j \beta_j + u_K \\ \widetilde{Y_K^*} &= \frac{1}{\sqrt{K}} \sum_{j=1}^{K} \widetilde{W}_j \gamma_j \end{aligned} \tag{A-3}$$

Under the assumptions in Theorem 1 as the number of regressors gets large for any j,

$$\begin{aligned} \frac{cov(CH^*, Y^*)}{var(Y^*)} &\approx \frac{\sum_{\ell=-\infty}^{\infty} E\left(W_j \beta_j W_{j-\ell} \gamma_{j-\ell}\right)}{\sum_{\ell=-\infty}^{\infty} E\left(W_j \gamma_j W_{j-\ell} \gamma_{j-\ell}\right)} \\ &= \frac{\sum_{\ell=-\infty}^{\infty} E\left(W_j W_{j-\ell}\right) E\left(\beta_j \gamma_{j-\ell}\right)}{\sum_{\ell=-\infty}^{\infty} E\left(W_j W_{j-\ell}\right) E\left(\gamma_j \gamma_{j-\ell}\right)}, \end{aligned}$$

where the expectation is over both $(W_j W_{j-\ell})$ and $\left(\beta_j \gamma_{j-\ell}\right)$.

Similarly,

$$\frac{cov(\widetilde{CH^*}, \widetilde{Y^*})}{var(\widetilde{Y^*})} \approx \frac{\sum_{\ell=-\infty}^{\infty} E\left(\widetilde{W_j}\widetilde{W_{j-\ell}}\right) E\left(\beta_j \gamma_{j-\ell}\right)}{\sum_{\ell=-\infty}^{\infty} E\left(\widetilde{W_j}\widetilde{W_{j-\ell}}\right) E\left(\gamma_j \gamma_{j-\ell}\right)}.$$

Since in general the autocovariance structure of W_j will be different from the autocovariance structure of $\widetilde{W_j}$, these will be different and the restriction (A-2) will not be valid. However, we can give two examples for which (A-2) holds.

The first example is the independence case. Suppose there is no serial correlation in the W's so that the unobservables are uncorrelated with the observables. If that is the case

$$\begin{aligned}\frac{cov(CH^*,Y)}{var(Y^*)} &\approx \frac{E(W_jW_j)\,E(\beta_j\gamma_j)}{E(W_jW_j)\,E(\gamma_j\gamma_j)}\\ &= \frac{E(\beta_j\gamma_j)}{E(\gamma_j\gamma_j)}\\ &\approx \frac{cov(\widetilde{CH^*},\widetilde{Y})}{var(\widetilde{Y^*})}.\end{aligned}$$

The second example is if there exists some constant τ such that

$$E\left(\beta_j\gamma_{j-\ell}\right) = \tau E\left(\gamma_j\gamma_{j-\ell}\right).$$

In this case

$$\begin{aligned}\frac{cov(\widetilde{CH^*},\widetilde{Y})}{var(\widetilde{Y^*})} &= \frac{\sum_{\ell=-\infty}^{\infty} E\left(\widetilde{W_j}\widetilde{W_{j-\ell}}\right)\tau E\left(\gamma_j\gamma_{j-\ell}\right)}{\sum_{\ell=-\infty}^{\infty} E\left(\widetilde{W_j}\widetilde{W_{j-\ell}}\right) E\left(\gamma_j\gamma_{j-\ell}\right)}\\ &= \tau.\end{aligned}$$

Such a case can occur when γ_j and β_j have the same stationary ARMA process. To see this consider the MA(∞) process

$$\begin{aligned}\beta_j &= \omega_j^1 + \theta_1\omega_{j-1}^1 + \theta_2\omega_{j-2}^1 + \ldots\\ \gamma_j &= \omega_j^2 + \theta_1\omega_{j-1}^2 + \theta_2\omega_{j-2}^2 + \ldots\end{aligned}$$

where the joint distribution of (ω_j^1,ω_j^2) is serially uncorrelated with constant variance and $cov(\omega_j^1,\omega_k^2) = 0$ when $k \neq j$, then (defining $\theta_0 = 1$)

$$\begin{aligned}E\left(\beta_j\gamma_{j-\ell}\right) &= cov(\omega_j^1,\omega_j^2)\sum_{r=0}^{\infty}\theta_r\theta_{r+\ell}\\ E\left(\gamma_j\gamma_{j-\ell}\right) &= var(\omega_j^2)\sum_{r=0}^{\infty}\theta_r\theta_{r+\ell}\end{aligned}$$

so

$$E\left(\beta_j\gamma_{j-\ell}\right) = \frac{cov(\omega_j^1,\omega_j^2)}{var(\omega_j^2)}E\left(\gamma_j\gamma_{j-\ell}\right).$$

A.3 Proof of Theorem 2

Proof. Consiser any value $\alpha^* \neq \alpha$ that is consistent with the model and Condition 1′. Then define $(g^*(X),\varepsilon^*)$ as the analogues of $(g(X),\varepsilon)$ that accompany it. Then

$$\begin{aligned}E(Y \mid CH = 0, X) &= g^*(X) + E(\varepsilon^* \mid CH^* \leq 0, X)\\ &= g(X) + E(\varepsilon \mid CH^* \leq 0, X),\end{aligned}$$

$$\begin{aligned}E(Y \mid CH = 1, X) &= \alpha^* + g^*(X) + E(\varepsilon^* \mid CH^* > 0, X)\\ &= \alpha + g(X) + E(\varepsilon \mid CH^* > 0, X),\end{aligned}$$

and

$$E(\varepsilon^* \mid X) = 0.$$

Solving these equations for g^* yields

$$g^*(X) = g(X) + p(X)(\alpha - \alpha^*),$$

where $p(X)$ is the propensity score (i.e. $p(X) = \Pr(CH = 1 \mid X)$ and thus

$$\varepsilon^* = (\alpha - \alpha^*)(CH - p(X)) + \varepsilon.$$

If the alternative model satisfies (3.16) then

$$\frac{cov(CH^*, g^*(X))}{var(g^*(X))} = \frac{cov(CH^*, \varepsilon^*)}{var(\varepsilon^*)}$$

or

$$\frac{cov(CH^*, g(X)) + (\alpha - \alpha^*)\, cov(CH^*, p(X))}{var(g(X)) + 2(\alpha - \alpha^*)\, cov(g(X), p(X)) + (\alpha - \alpha^*)^2\, var(p(X))}$$
$$= \frac{cov(CH^*, \varepsilon) + (\alpha - \alpha^*)\, cov(CH^*, CH - p(X))}{var(\varepsilon) + 2(\alpha - \alpha^*)\, cov(\varepsilon, CH - p(X)) + (\alpha - \alpha^*)^2\, var(CH - p(X))}.$$

Defining

$$\phi \equiv \frac{cov(CH^*, g(X))}{var(g(X))} = \frac{cov(CH^*, \varepsilon)}{var(\varepsilon)},$$

and dividing top and bottom by $var(g(X))$ and $var(\varepsilon)$, we get

$$\frac{\phi + (\alpha - \alpha^*)\frac{cov(CH^*, p(X))}{var(g(X))}}{1 + 2(\alpha - \alpha^*)\frac{cov(g(X), p(X))}{var(g(X))} + (\alpha - \alpha^*)^2 \frac{var(p(X))}{var(g(X))}}$$
$$= \frac{\phi + (\alpha - \alpha^*)\frac{cov(CH^*, CH - p(X))}{var(\varepsilon)}}{1 + 2(\alpha - \alpha^*)\frac{cov(\varepsilon, CH - p(X))}{var(\varepsilon)} + (\alpha - \alpha^*)^2 \frac{var(CH - p(X))}{var(\varepsilon)}}.$$

Algebraic manipulation yields

$$\begin{aligned}
0 = {} & (\alpha - \alpha^*)^3 \left[\frac{var(CH - p(X))}{var(\varepsilon)} \frac{cov(CH^*, p(X))}{var(g(X))} - \frac{var(p(X))}{var(g(X))} \frac{cov(CH^*, CH - p(X))}{var(\varepsilon)}\right] \\
& + (\alpha - \alpha^*)^2 \left[\phi \frac{var(CH - p(X))}{var(\varepsilon)} + 2\frac{cov(CH^*, CH - p(X))}{var(\varepsilon)} \frac{cov(CH^*, p(X))}{var(g(X))} \right. \\
& \qquad \left. - \phi \frac{var(p(X))}{var(g(X))} - 2\frac{cov(g(X), p(X))}{var(g(X))} \frac{cov(CH^*, CH - p(X))}{var(\varepsilon)}\right] \\
& + (\alpha - \alpha^*) \left[\frac{cov(CH^*, p(X))}{var(g(X))} + 2\phi \frac{cov(\varepsilon, CH - p(X))}{var(\varepsilon)} \right. \\
& \qquad \left. - \frac{cov(CH^*, CH - p(X))}{var(\varepsilon)} - 2\phi \frac{cov(g(X), p(X))}{var(g(X))}\right].
\end{aligned}$$

Thus the only values of α^* that are consistent with the observed data and the assumptions of the model are members of the set $\mathcal{A}$, so it must be identified. Furthermore note that $\alpha^* = \alpha$ is a member of the set.

■

A.4 Proof of Theorem 3

Proof. Follow similar logic to Theorem 2. Consider any value $\alpha^* \neq \alpha$ that is consistent with the model and Condition 1$'$. Then define $(g^*(X), \varepsilon^*)$ as the analogues of $(g(X), \varepsilon)$ that accompany it. Then

$$Y = \alpha^* CH + g^*(X) + \varepsilon^*.$$

Since ε^* must be mean zero conditional on X,

$$g^*(X) = g(X) + (\alpha - \alpha^*)\, b(X),$$

and

$$\varepsilon^* = \varepsilon + (\alpha - \alpha^*)\, u.$$

If the alternative model satisfies the conditions in the theorem then

$$\frac{cov(CH^*, g^*(X))}{var(g^*(X))} = \frac{cov(CH^*, \varepsilon^*)}{var(\varepsilon^*)}.$$

Substituting in for g^* and ε^* leads to

$$\text{(A-4)} \qquad \begin{aligned} & \frac{cov(b(x), g(X)) + (\alpha - \alpha^*)\, var(b(X))}{var(g(X)) + 2\,(\alpha - \alpha^*)\, cov(g(X), b(X)) + (\alpha - \alpha^*)^2\, var(b(X))} \\ = \ & \frac{cov(u, \varepsilon) + (\alpha - \alpha^*)\, var\,(u)}{var(\varepsilon) + 2\,(\alpha - \alpha^*)\, cov(\varepsilon, u) + (\alpha - \alpha^*)^2\, var(u)}. \end{aligned}$$

As above, defining

$$\phi \equiv \frac{cov(b(X), g(X))}{var(g(X))} = \frac{cov(u, \varepsilon)}{var(\varepsilon)}$$

and dividing top and bottom of the left hand side of A-4 by $var(g(X))$ and the right hand side by $var(\varepsilon)$ (respectively), one finds that

$$\frac{\phi + (\alpha - \alpha^*)\, \frac{var(b(X))}{var(g(X))}}{1 + 2\,(\alpha - \alpha^*)\,\phi + (\alpha - \alpha^*)^2\, \frac{var(b(X))}{var(g(X))}} = \frac{\phi + (\alpha - \alpha^*)\, \frac{var(u)}{var(\varepsilon)}}{1 + 2\,(\alpha - \alpha^*)\,\phi + (\alpha - \alpha^*)^2\, \frac{var(u)}{var(\varepsilon)}}.$$

Algebraic manipulation leads to

$$0 = (\alpha - \alpha^*)\left[\frac{var(b(X))}{var(g(X))} - \frac{var(u)}{var(\varepsilon)}\right] + (\alpha - \alpha^*)^2\,\phi\left[\frac{var(b(X))}{var(g(X))} - \frac{var(u)}{var(\varepsilon)}\right].$$

This gives two roots

$$\begin{aligned} \alpha^* &= \alpha \\ \alpha^* &= \alpha + \frac{1}{\phi} = \alpha + \frac{var(\varepsilon)}{cov(u, \varepsilon)} \end{aligned}$$

Thus the only values of α^* that are consistent with the observed data and the assumptions of the model are members of the set $\mathcal{A}$, so it must be identified. Furthermore note that $\alpha^* = \alpha$ is a member of the set.

■

A.5 Cubic Solution from Instrumental Variable

Following the text above, the question is whether the assumptions allow us to pin down the bias. Suppose it cannot. Then there would exist alternative values α^*, γ^*, and ε^* with $\alpha^* \neq \alpha$ so that for the same $\widehat{\alpha}$ as in the text

$$\widehat{\alpha} = \alpha^* + \frac{cov\,(v, \varepsilon^*)}{\lambda var(v)}.$$

Under these conditions note that

$$\begin{aligned} Y - \alpha^* CH &= (\alpha - \alpha^*)\, CH + X'\gamma + \varepsilon \\ &= (\alpha - \alpha^*)\left[X'\beta + u + \lambda\,(X'\pi + v)\right] + X'\gamma + \varepsilon, \end{aligned}$$

and thus

$$\begin{aligned} \gamma^* &= \gamma + (\alpha - \alpha^*)\,(\beta + \lambda\pi) \\ \varepsilon^* &= \varepsilon + (\alpha - \alpha^*)\,(u + \lambda v)\,. \end{aligned}$$

But if this model satisfies the assumptions we know that

$$\frac{cov(X'\pi, X'\gamma^*)}{var(X'\gamma^*)} = \frac{cov(v, \varepsilon^*)}{var(\varepsilon^*)},$$

which is equivalent to

$$\begin{aligned} &\frac{cov\,(X'\pi, X'\gamma) + (\alpha - \alpha^*)\, cov\,(X'\pi, (X'\beta + \lambda X'\pi))}{var\,(X'\gamma) + 2\,(\alpha - \alpha^*)\, cov(X'\gamma, (X'\beta + \lambda X'\pi)) + (\alpha - \alpha^*)^2\, var(X'\beta + \lambda X'\pi)} \\ &= \frac{cov\,(v, \varepsilon) + (\alpha - \alpha^*)\, cov\,(v, (u + \lambda v))}{var(\varepsilon) + 2\,(\alpha - \alpha^*)\, cov(\varepsilon, (u + \lambda v)) + (\alpha - \alpha^*)^2\, var(u + \lambda v)}. \end{aligned}$$

Imposing the restriction from the true model

$$\phi \equiv \frac{cov(X'\pi, X'\gamma)}{var(X'\gamma)} = \frac{cov(v, \varepsilon)}{var(\varepsilon)},$$

yields

$$\begin{aligned} &\frac{\phi + (\alpha - \alpha^*)\,\frac{cov(X'\pi,(X'\beta+\lambda X'\pi))}{var(X'\gamma)}}{1 + 2\,(\alpha - \alpha^*)\,\frac{cov(X'\gamma,(X'\beta+\lambda X'\pi))}{var(X'\gamma)} + (\alpha - \alpha^*)^2\,\frac{var(X'\beta+\lambda X'\pi)}{var(X'\gamma)}} \\ &= \frac{\phi + (\alpha - \alpha^*)\,\frac{cov(v,(u+\lambda v))}{var(\varepsilon)}}{1 + 2\,(\alpha - \alpha^*)\,\frac{cov(\varepsilon,(u+\lambda v))}{var(\varepsilon)} + (\alpha - \alpha^*)^2\,\frac{var(u+\lambda v)}{var(\varepsilon)}}. \end{aligned}$$

Solving out yields

$$\begin{aligned} 0 = (\alpha - \alpha^*)^3 &\left[\frac{cov\,(v, (u + \lambda v))}{var(\varepsilon)}\frac{var(X'\beta + \lambda X'\pi)}{var(X'\gamma)} - \frac{cov\,(X'\pi, (X'\beta + \lambda X'\pi))}{var(X'\gamma)}\frac{var(u + \lambda v)}{var(\varepsilon)}\right] \\ + (\alpha - \alpha^*)^2 &\left[\phi\frac{var(X'\beta + \lambda X'\pi)}{var(X'\gamma)} + 2\frac{cov\,(v, (u + \lambda v))}{var(\varepsilon)}\frac{cov(X'\gamma, (X'\beta + \lambda X'\pi))}{var(X'\gamma)}\right. \\ &\left. -\phi\frac{var(u + \lambda v)}{var(\varepsilon)} - 2\frac{cov\,(X'\pi, (X'\beta + \lambda X'\pi))}{var(X'\gamma)}\frac{cov(\varepsilon, (u + \lambda v))}{var(\varepsilon)}\right] \\ + (\alpha - \alpha^*) &\left[\frac{cov\,(v, (u + \lambda v))}{var(\varepsilon)} + 2\phi\frac{cov(X'\gamma, (X'\beta + \lambda X'\pi))}{var(X'\gamma)}\right. \\ &\left. -\frac{cov\,(X'\pi, (X'\beta + \lambda X'\pi))}{var(X'\gamma)} - 2\phi\frac{cov(\varepsilon, (u + \lambda v))}{var(\varepsilon)}\right]. \end{aligned}$$