Working Paper #415
Industrial Relations Section
Princeton University
January, 1999

Interpreting Instrumental Variables Estimates of the Returns to Schooling

Jeffrey Kling

Princeton University and NBER

I thank Josh Angrist, Matt Eichner, Jerry Hausman, Guido Imbens, Larry Kahn, Larry Katz, Whitney Newey, Steve Pischke and John Tyler for helpful comments, and David Card for providing data from his research. I gratefully acknowledge support from a National Science Foundation Graduate Fellowship.

Interpreting Instrumental Variables Estimates of the Returns to Schooling

ABSTRACT

An instrumental variable can be used to identify the labor market return to schooling by allowing comparisons between groups of individuals whose differences in schooling levels are uncorrelated with their underlying marginal benefit from schooling and with other aspects of unobserved ability. When the education decisions are based on individual-specific marginal benefits and costs, there is no single rate of return for everyone in the population. This paper demonstrates economic insights from methods interpreting instrumental variables estimates as weighted averages of individual-specific causal effects of schooling on wages by synthesizing existing theoretical and econometric work, and by using geographic variation in college proximity as an example of an instrumental variable.

Characterizing the groups affected by the college proximity instrument, I find the largest increase in schooling levels among individuals from more disadvantaged backgrounds. Although the data is insufficient to obtain useful estimates of group-specific rates of return, I directly compute the weight each group receives in the overall estimate. In analyzing the response function and showing the level of schooling at which individuals change their behavior in response to the instrument, I demonstrate that the instrument has the greatest impact on the transition from high school to college. This corresponds to the economic intuition that changes in the marginal cost of college should be concentrated at this transition and should not affect all levels of schooling equally. The results suggest that disadvantaged groups are most responsive to policies lowering college costs, and that increases in education for these groups may have high payoff.

How can we estimate the causal effect of an additional year of schooling on earnings? In answering this question, we need to consider the counterfactual: what would individuals have earned if they had acquired more education, in comparison to what they are earning now? To obtain an estimate of this causal effect we could consider analyzing the earnings of other individuals with similar characteristics -- say, their twin sisters -- who have more education. However, the earnings of these higher educated twins are unlikely to correspond to the counterfactual earnings of the less educated twins because these individuals *chose* their education levels.¹ Only random variation uncorrelated with the marginal benefit of schooling among these twins would identify the causal effect of schooling. This same reasoning holds if we were estimating the returns to schooling from observational data using traditional statistical methods to account for observable differences between individuals.

Where might we find such random variation in schooling? For the empirical examples in this paper, I utilize differences in schooling attributable to geographic variation in the location of colleges, which has been analyzed previously by both Kane and Rouse (1993) and Card (1995b). To understand this source of variation, think of two groups composed of observably similar individuals, except that some grew up in areas that happened to have colleges nearby and others

¹ The very fact that these twins chose to acquire different amounts of schooling suggests that they differ in unobserved ways that may be related to their potential earnings at various levels of education. Indeed, in a rational model of investment in schooling, two individuals with the same background characteristics (including "ability to borrow") would choose different levels of schooling precisely because the one with the greater benefit from an additional year of schooling would invest in more schooling, resulting in estimates of returns to schooling that would overstate the true effect. This would be consistent with the large 12-16% estimated returns to schooling for twins in work by Ashenfelter and Krueger (1994). In later work, Ashenfelter and Rouse (1998) argue explicitly that differences in schooling levels of twins are actually optimization errors which are uncorrelated with the marginal benefit from schooling. They also show that using larger sample of twins results in a estimate of about 9 percent for the return to schooling.

grew up in areas where colleges happened to be farther away. These authors argue that college proximity affects the amount of schooling acquired by reducing housing and transportation costs, and that proximity only affects earnings through schooling and is unrelated to differences in labor markets or other factors that may affect earnings. For individuals with the same demographic and socioeconomic characteristics, college proximity is argued to be as good as randomly assigned -- so if those who grew up farther from a college had instead been near one, they would on average have gotten as much schooling and earned just as much as those who did grow up near a college. In this way, the college proximity can be used as an instrumental variable to identify the causal effect of schooling.

The instrumental variables used to identify the returns to schooling typically influence people to acquire more schooling through some rule or incentive that only affects the schooling decisions of a small subgroup of the population.² The return to schooling for individuals in this subgroup of the population may well differ, for example, between those who could afford to attend an additional year of school because they happened to grow up near a college and those from a random sample of the population. The existence of these differences is predicated on the idea that the return to schooling varies among individuals; there is no single rate of return

² For example, Angrist and Krueger (1991) use the quarter of the year in which people are born as an instrumental variable, arguing that compulsory schooling laws generate a quasi-experiment that compels some individuals to stay in school longer than others. They then compare schooling and earnings of individuals born at different times of the year, since people born in the calendar year start school at the same time in most states -- so those born earlier in the year reach the minimum school leaving age at a lower grade than individuals born later in the year. Kane and Rouse (1993) argue that local college tuition level is an instrumental variable for differences in completed college education that are unrelated to the counterfactual "potential" earnings -- i.e. the hypothetical amount individuals would have earned if local college tuition had been different than it actually was.

applicable to everyone in the population. Drawing on previous theoretical work by Card (1995a) and Lang (1993), I formulate a model of returns to schooling in order to make explicit the idea that the instrumental variable generates an estimate of the average causal effect among individuals with different marginal benefits from schooling.

It is extremely useful to have information on the composition of the subgroup affected by an instrumental variable for the purposes of applying results in new contexts -- that is, to evaluate the external validity of the results. For instance, in assessing the likely impact of tuition subsidies (such as those discussed in the 1996 Presidential election campaign in the U.S.), results based on college proximity as an instrument are particularly relevant for at least two reasons. First, this instrument may have greater effects on educational attainment for individuals from different backgrounds. For instance, we might expect those with limited ability to borrow for college to be more sensitive to costs lowered by college proximity. Second, the instrument may induce changes at different levels of education. If college proximity is actually inducing individuals to go to college who would not have otherwise, then we would expect the largest effect of proximity on education to occur for individuals who would otherwise have had between twelve and fifteen years of schooling.

To inform our interpretation, this paper draws on recent theoretical econometric research by Angrist and Imbens (1995) in order to quantify the composition of the subgroup affected by the instrumental variable. Specifically, I examine differences in the impact of college proximity on schooling by covariates such as family background. As the overall estimate of the return to education is an average among individuals affected by the instrument, I compute estimates of covariate weights used in creating this average. These weights demonstrate the influence on the

overall estimate of subgroups defined by covariates such as low parental education. I also analyze the effect of college proximity at different levels of education, and trace out the response function that is used in computing the relative weights of these levels in the estimation of the overall average effect. While helping inform policy discussion, this characterization of the response function can also serve as a check on the validity of the instrument. If college proximity was only associated with ninth graders otherwise attending tenth grade, then plausibility of estimates would be seriously undermined.

In addition to characterizing the affected subgroups and the response function, examining the sensitivity of IV estimates is the second aspect of interpretation treated in this paper. This evaluation of internal validity is critical for interpretation because the basic assumptions underlying any instrumental variable are inherently untestable -- as they involve restrictions on counterfactuals that cannot be directly observed. For example, no statistical test can prove that college proximity only affects schooling but does not affect earnings directly. This assumption could be violated if college proximity were affecting earnings through some other channel such as residing in a high-wage urban labor market. Some necessary assumptions do have testable implications, however, that I can check in order to buttress the interpretation that the instrument is valid.

To summarize the organization of this paper, Section II reviews a simple model of schooling choice that provides the framework for interpretation of empirical results. Section III replicates recent work by Card (1995b) using college proximity to study the returns of schooling using data from the original young men's cohort of the National Longitudinal Survey (NLSYM). Section IV discusses interpretation of these estimates as a weighted average of causal effects for

particular subgroups of the population. Section V examines the impact of the instrument at different levels of schooling. Section VI discusses the sensitivity of results to possible violations of the identifying assumptions. Section VII concludes.

II. A Model of Schooling Choice

Much of the literature on returns to schooling has been concerned with "ability bias" (e.g. Griliches, 1977). The following discussion of a simple human capital model -- an adaptation from Card (1995a) of Becker's supply and demand model (1975) -- is an attempt to clarify what it is that I hope to estimate, and to categorize the types of biases that may complicate interpretation of results.

In a canonical model of schooling choice, individuals will invest in schooling until the point at which the marginal change in present discounted value of future income from an additional year of schooling is equal to the intertemporal rate of substitution. To simplify, I assume a person-specific discount rate r_i and zero earnings while in school. Individuals earnings vary with their level of schooling, $y = g(S)$. This generates the form of the utility function in (1).

$$U(y, S) = \log(y) - \phi(S) = \log(g(S)) - \phi(S) \quad (1)$$

The optimal level of schooling is determined implicitly by the first order condition in (2), when the marginal benefit of schooling is equal to the marginal cost.

$$\frac{g'(S)}{g(S)} = \phi'(S) \quad (2)$$

Equations (3) and (4) establish notation for the marginal benefits and costs that influence an individual's optimal schooling choice, parameterizing them as functions of observable characteristics X and Z , and unobservable components.

$$\frac{g'(S)}{g(S)} = b_i = X_i\beta + v_i \quad (3)$$

$$\phi'(S) = r_i + kS ; \quad k \geq 0 ; \quad r_i = X_i\delta - Z_i\gamma + \xi_i \quad (4)$$

The assumption in (4) of a discount rate that is increasing in the amount of schooling seems plausible, for example, when individuals can finance education first internally from family savings, then from federally subsidized sources, and finally from private sources. Substituting parameterizations from (3) and (4) into (2) results in explicit equation for the optimal schooling level in (5).

$$\frac{b_i - r_i}{k} = S_i^* = X_i\theta + Z_i\gamma + \eta_i \quad (5)$$

Integrating (3) results in the equation for earnings in (6).

$$\int_S \frac{g'(S)}{g(S)} = \log(y) = b_i S_i + a_i \quad (6)$$

There are two types of "ability" in this model. Individual differences in the marginal benefit to a year of schooling are captured in the education slope coefficient of the earnings equation in (6). Individual differences in earnings capacity that do not interact with education

are embodied in the individual specific earnings equation intercept a_i . I would like to identify the marginal benefit to an additional year of schooling; this model allows b_i , the marginal benefit of schooling, to vary across individuals, but I can still hope to estimate an average marginal return, $E[b_i]$.

I have set up the model to suggest that instrumental variables in Z can be available from an instrumental variable that affects schooling through the discount rate (e.g. capacity to pay for schooling) and not the marginal benefit. That is, including Z in (4) but excluding Z from (3) generates variation in schooling in (5) uncorrelated with the benefit b_i . If this capacity to pay does not directly determine earnings -- i.e. Z is uncorrelated with a_i and can be excluded from the earnings equation (6) -- then Z serves as a legitimate instrumental variable. With E representing potential experience, this model provides theoretical motivation for two stage least squares estimation of the returns to schooling, in which the first stage is estimated by (5) and the second stage is based on (7).

$$Y_{it} = S_{it}\rho + E_{it}\pi_1 + X_i\pi_2 + \alpha_i + \varepsilon_{it} \quad (7)$$

I make an additional assumption that the instrument affects behavior monotonically. That is, the presence of a college nearby never causes individuals to obtain less schooling than they would in its absence. This allows instrumental variables estimates of the schooling coefficient in the earnings equation to be interpreted as an average causal response (Angrist and Imbens 1995). In this case, the estimated schooling coefficient ρ is the weighted average of the heterogeneous marginal effects for those whose schooling choice is affected by the instrument.

In order for proximity to a college to serve as an instrument, the distribution of ability

(both b_i and a_i) must be independent of the presence of a college nearby at the time of the choice to attend college, conditional on background characteristics. For example, I assume that the geographic proximity to college of an individual whose parents both dropped out of high school is independent of his marginal benefit from an additional year of schooling and that college proximity has no direct effect on wages except through the additional schooling. College proximity may be more important for some subgroups of the population, such as those from less wealthy families that are more liquidity constrained. In Section IV, I consider the interaction of the instruments (Z) with individual characteristics (X).

There are of course several reasons why college proximity may not be a good instrument. Observably equivalent families may have an unobserved taste for education, and choose to live closer to a college. Also, wages in labor markets for those grew up near a college may simply reflect unobserved geographic wage premia, rather than difference in wages due to schooling. I discuss an estimation strategy to control for geographic wage differences in Section VI.

III. Replication of Previous Results

In this section I replicate estimates reported by Card (1995b) of the overall average causal effect of schooling on earnings for those affected by college proximity. Card utilizes data from the original young men's cohort of the National Longitudinal Survey, which began with youth aged 14-24 in 1966. For direct comparability, I use data and variable definitions provided by Card (1995c). For college proximity, I use an indicator for whether or not there was a four year college in the local labor market (defined as the local county) in 1966. Card notes that proximity to two year colleges seems to have little impact on educational attainment, so I focus this

analysis on proximity to four year colleges. Data on hourly wages at current or past job and highest grade completed are taken from the 1976 wave. The resulting sample size is 3010.³

Card (1995b) uses a variety of specifications to estimate the returns to schooling. In Table I, I present a sample of these results which I replicate exactly. The relation between years of schooling and college proximity is reasonably strong; in the first stage regression (ages 14-24 in 1966) for years of schooling, the coefficient of college proximity is 0.37 with a standard error of 0.10. Instrumental variables estimates of the return to schooling range from .094 to .132 in Table I, and have larger point estimates of the return to education than ordinary least squares in these and all specifications reported by Card, although the coefficients are imprecisely estimated and I cannot reject the hypothesis that the difference is due to sampling error.

IV. Characterizing the Affected Groups

With these replicated estimates as a frame of reference, I present new quantitative evidence characterizing the group affected by this instrumental variable. These estimates cannot be generalized to the larger population without additional assumptions--such as a constant marginal benefit of education across individuals. Without these assumptions, the external validity of the estimates depends upon the precision with which the affected subgroup can be

³ The sample described in Card's paper (1995c) includes all observations with reporting highest grade completed in 1976 and a wage between \$1 and \$30 per hour, a sample size of 3059. Card's (1995c) computer programs show that 49 observations were excluded for whom the interview month was missing in 1966. This appears to have been an attempt to omit those not interviewed, but those 49 individuals were actually interviewed even though the interview month is not recorded in the data. I use this same sample creation rule for greater comparability. Fortunately, inclusion of these additional observations does not substantially change any results in Card's original paper.

characterized and on the policy interest generated by that group.

As a point of departure in attempting to characterize the affected group, Card (1995b) speculates that the effect of college proximity is more important for children of less wealthy households. He constructs an index of background (predicted education from a linear regression of years of schooling on region and background characteristics, estimated on the sample that did not have a college nearby). I define four quartiles based on this index.

To see the differences in the underlying components used to create the background quartiles, I show the probability of selected characteristics in Table II. The most important predictor of family background quartile is parental education. In the lowest quartile, none of individuals report that either their mother or father graduated from high school. (Note that in this table, those with missing parental education are coded as not reporting that parent was a high school graduate.) In the highest quartile, 97 percent of mothers are high school graduates, as are 84% of fathers. The presence of college in the 1966 county does vary somewhat across quartiles, from 59 percent for the lowest to 73 percent for the highest quartile. While this difference is much less dramatic than that of other characteristics of these individuals, I focus this analysis on differences in education and wages by college proximity conditional on having the same background characteristics.

Card (1995b) shows that actual average education difference by college proximity is much larger for the lowest background quartile than the highest background quartile. The IV estimate of the return to schooling can be interpreted as a weighted average across the four background quartiles (Card 1995a), as shown in (8)--where growing up with a college nearby is denoted as $Z=1$. Here, the weight given to each quartile q is the product of the proportion of the

population in that subgroup (w_q) and the impact on schooling for that subgroup (ΔS_q).

$$\rho = \frac{E[Y | Z=1, X] - E[Y | Z=0, X]}{E[S | Z=1, X] - E[S | Z=0, X]} = \frac{\Delta Y}{\Delta S} = \frac{\sum_{q=1}^4 w_q \Delta S_q \rho_q}{\Delta S} \quad (8)$$

The impact of the instrument on schooling for each quartile is presented column two of Table III. After partialling out age, location, and background characteristics for men ages 24-34 in 1976, having a college nearby is associated with 0.92 more years of schooling for the lowest background quartile, and 0.13 more years of schooling for the highest background quartile. The standard errors of these estimates (in parentheses) show the coefficients to be significant for the lower two quartiles. This suggests that the discount rates of disadvantaged individuals are indeed more sensitive to college proximity than those of more well-off families -- perhaps due to borrowing constraints for their families.

In Table III, I also compute the relative weight each background quartile receives in the IV estimate using a single instrument. Column three contains the weight that would be used to form a weighted average of the marginal return to schooling for that group (ρ_q) for those affected by the instrument, across the four quartiles.⁴ (Unfortunately, the sample size is too small to obtain reliable estimates of each ρ_q itself). This calculation provides direct evidence that the

⁴ The background quartiles are constructed so that each includes one quarter of the observations with no college in their 1966 county, 239-240 observations per quartile. Since background quartile is positively correlated with college in county, this implies that there is a larger share of data in the upper quartiles with a college in their county in 1966, as is reflected in the larger overall weights for the upper quartiles in Table III. This rule was adopted to improve estimation efficiency by ensuring that the upper quartiles do not include only a small number of observations with no college in county. There are 339 observations that do have a college in county in the lowest quartile, 483 in the third quartile, 593 in the second quartile, and 638 in the highest quartile.

lowest background quartile receives 50 percent of the weight in the overall IV estimate, and the highest two groups receive only 26 percent of the weight. If the average marginal benefit of schooling were higher for the lowest background quartile, this would be also consistent with the large estimates of return to schooling using distance as an instrument.

These IV estimates will represent the average marginal benefit from an additional year of education for the subgroup affected by the college proximity instrument . This subgroup turns out to be of policy interest both because it is composed largely of children from disadvantaged families, and because it represents individuals "on the margin" whose schooling decisions are affected by small changes in education costs.

V. Characterizing the Response Function

In addition to affecting a particular group as in Section III, an instrumental variable may affect different transitions between levels of schooling. For example, college proximity may have the greatest impact on the transition from twelfth grade to the first year of college if it is affecting liquidity-constrained individuals. The response function traces out the impact of the instrument at each level of schooling, providing a richer description of the process through which the instrument influences behavior which can be useful for assessing external validity and relating IV estimates to proposed policy initiatives.

This response function can be estimated from the cumulative distribution functions (CDFs) of schooling at different values of the instrument. The difference in the CDFs is equivalent to the fraction of the population who received at least one more year of schooling due to the instrument as shown in (9), where "j" is the number of years of schooling, X represents

background characteristics, and $Z=1$ when a college is nearby.

$$\begin{aligned}
 \Delta S &= \sum_{j=1}^J Pr(S < j | Z=0, X) - Pr(S < j | Z=1, X) \\
 &= \sum_{j=1}^J Pr(S_1 \geq j > S_0)
 \end{aligned} \tag{9}$$

This difference integrates to the average difference in schooling with and without a college nearby. The response function is shown by Angrist and Imbens (1995) to be proportional to the Average Causal Response (ACR) weighting function used by IV estimation, where the average marginal benefit of a year of schooling for the group affected by the instrument can be interpreted as the weighted average of per-unit changes in schooling.

Figure 1 shows the difference in the CDFs conditional on age and background, with a 95% pointwise confidence interval. Note that the ACR weighting function is derived for nonparametric conditioning; I approximate using linear probability models for each integer schooling level conditioning on age, location and background.

The response function indicates that the college proximity instrument does indeed have larger effects for individuals who were caused to attend at least some college. At 13 years of schooling, I interpret the estimates in Figure 1 to indicate that seven percent of individuals with similar demographics were induced to obtain 13 or more years of schooling due to college proximity, when they would have obtained 12 or fewer years of schooling without a college nearby. Note that the CDF difference captures the number of person-years of additional schooling, so an individual who was induced to complete 13 years of schooling by college proximity but who would have otherwise only finished tenth grade contributes to the CDF difference at 11, 12 and

13 years of schooling. This multi-year effect of the instrument rationalizes at least part of the positive CDF difference over the range of 9 to 12 years of schooling. The significant difference in CDFs at grade 10 implies that some individuals were induced to receive four additional years of schooling by college proximity, or that the anticipation of college kept potential dropouts in high school for one or two extra years even if they did not eventually attend college.

To identify an average causal effect using IV when marginal benefits of schooling vary across individuals, Angrist and Imbens (1995) show that an instrument must affect all individuals in a monotonic manner. In this case, college proximity should not decrease educational attainment for any individuals. This assumption has the testable implication that the difference in the cumulative distribution functions of schooling for the two values of a binary instrument should not be negative, which is satisfied in Figure 1.

The model of schooling choice in Section II has an additional simple prediction that the density of the weighting function should be greater over lower years of schooling for individuals from families with higher discount rates. This prediction is derived from (10), which shows that the covariance of discount rates and optimal schooling is nonpositive in this model.

$$Cov(r_i, S_i^*) = \frac{\sigma_{br} - \sigma_r^2}{k} \leq 0 \quad (10)$$

For higher discount rates, schooling will be lower regardless of the distribution of ability, unless discount rates and ability are perfectly positively correlated. Thus, on average I can expect that a group with a high discount rate will be induced by college proximity to increase their education from a lower initial level (under the model's assumption that proximity affects all discount rates equally). For example, proximity to college will tend to cause high discount rate individuals to

graduate from high school and attend some college when they might otherwise have dropped out of high school, while proximity will tend to cause low discount rate individuals who would have attended college anyway to increase the number of years they attend college.

The prediction that lower quartiles will be effected at lower ranges of the response function is verified in Figure 2, where I compute the CDF difference by background quartile. There appears to be substantial response in the two lower background quartiles from those who otherwise would not have attended college. The response for the upper two quartiles is concentrated among those who would have attended at least some college, but attend additional years due to college proximity. In assessing the statistical variability, a simple rule of thumb for Figure 2 is that point estimates greater than .05 can be distinguished from zero at a p-value of .05.

Within the lowest background quartile, 13 percent received 13 or more years of schooling in Figure 2 when a college was nearby in 1966, while comparable members of that quartile who did not have a college nearby completed twelve years of schooling or less. There also appears to be some effect of inducing 9th-12th graders to graduate high school and attend some college. One interpretation is that the schooling levels of those in the lowest quartile differ for reasons that have nothing to do with college proximity itself, which would cast doubt on the validity of the instrument. This behavior could be consistent, however, with the role of college proximity in affecting schooling choices. For instance, individuals may be forward looking but have imperfect foresight (e.g. college proximity caused a potential tenth grade dropout to complete eleventh grade in anticipation of being able to afford to go to college, but this individual had to quit high school and start working after an unexpected death in the family). Alternatively,

college proximity may simply have had large effects on these individuals (e.g. causing a potential tenth grade dropout to attend at least some college).

Overall, these patterns suggest that for the least wealthy, being near a college helps motivate individuals to stay in school longer and makes it more affordable to at least try out college. For individuals from moderately wealthy families, college proximity allows individuals to complete additional years of college.

VI. Sensitivity of Estimation

While sections four and five focused largely on the interpretation of the external validity of existing estimates, an important aspect in interpreting IV estimates is investigation of internal validity -- examining the sensitivity of estimates to violations of key identifying assumptions. First, I present evidence that college proximity is not as good as randomly assigned among older individuals in the data; it appears to be related to schooling choices because the variable it is based upon used residential location gathered after some college may have been attended. Second, I show that strategies that attempt to properly account for geographic wage differences related to college proximity suggest that there does not appear to be a substantial geographic wage premium associated with growing up in county that had a college. Third, I find that using alternative data sources does not seem to reduce the variability of the estimates.

In a model in which college proximity affects schooling choices, the appropriate measure of proximity would be based on information about residential location while this decision was being made, say at age 18. In this data however, college proximity is constructed based on residential location at the time of first interview in 1966. At that time, older sample members

(say, ages 20-24) may have already have completed their postsecondary education. If those who went to college were more likely to live in near a college, then using college proximity determined by post-college location would not identify an appropriate comparison group, since college attendees who should have been grouped with those who grew up further from a college would been have misclassified as having grown up near a college. Thus, proximity cannot be thought of as being as good as randomly assigned, since it would be related to the outcome of educational choices of some sample members. Card notes this possibility, and presents estimates based only on sample members ages 14-19 in 1966, shown in the fourth column of Table I. For these younger sample members, the estimated return to schooling is somewhat lower than the estimate for the full sample, although it is imprecisely estimated. This is consistent with an upward bias in the coefficient induced by endogenous location of older sample members.

A more direct test of differences in the effect of the distance instrument by age examines the additional schooling obtained by members of different age groups. Older individuals might be expected to have attained more schooling simply because have had more time since high school to go to college, so I focus on highest grade attained at age 25. If the relationship between college proximity and educational attainment is not affected by location choice of older individuals, I would expect that the effect should be essentially the same across ages within the cohort of young men in the NLSYM.

The results in Table IV show that individuals ages 14-19 in 1966 received about 0.26 additional years of schooling when a college was nearby in 1966 (conditional on age, past location, and background), while those ages 20-24 in 1966 appear to have received 0.66 additional years of schooling when a college is nearby. The significantly higher schooling level

of the older group suggests that this group contains some individuals with higher education who moved to a local area that has a college.⁵ If so, I cannot identify the causal effect of college proximity on years of schooling for those ages 20-24 in 1966 without additional information about college proximity in the area in which they actually grew up. Although Card's basic specification uses data for all ages, this evidence suggests that results including those aged 20-24 are not based on a legitimate instrument.

Wage differences due to geographic location (such as local labor market conditions or urban wage premia) could also confound interpretation of the estimate of returns to schooling. Card's first attempt to account for these potential factors was to control for current region of residence in the regression. After controlling for residence in an SMSA in 1966, an additional indicator for residence in an SMSA in 1976 is effectively a indicator of moving. Similarly the indicator for current residence in the South is also an indicator for moving after controlling for region in 1966. Since moving is a choice, simply using it as a regressor does not control for the

⁵ The 1994 NLS User's Guide notes that variables depending on address information have been created by the Census in an inconsistent manner. In particular, it is not clear whether address information was based on location of household at screening or address at first interview. "One critical universe that was affected, apparently, was that of college students temporarily away from their permanent residence at the time of the interview." (p. 228). The majority of geographic variables were revised in the mid-1970s to correct for known discrepancies in permanent versus temporary address data, but the "Presence and Type of College in Labor Market of Current Residence" was not updated. This suggests that some members of the cohort who were old enough to enter college in 1966 (say at least 18) may appear to have lived within geographic proximity of a college because they were attending college at the time, and not because their permanent residence was near a college.

The number of cases affected by this coding procedure is unknown. If the residence of students in college was systematically coded as their permanent residence, then I would expect to find large difference in the correlation between college proximity and educational attainment between those ages 14-17 in 1966 and those ages 18-19. The results in Table 2 suggest that any effect of this phenomenon is not overwhelmingly strong, as the effect of the distance instrument for these groups is similar and well within sampling error.

causal effect of moving on wages (or the effect of living in a certain region on wages) -- for the same reasons that educational choices complicate estimation of the causal effect of schooling on wages. Since the data shows that the endogenous variable for moving is correlated with college proximity and other variables in the econometric specification, it may result in a bias of indeterminate sign in estimated returns to schooling -- even if there is no true geographic wage premium.

In another attempt to account for geographic wage differences, Card includes a main effect for presence of a college nearby, and instruments for education with an interaction between college proximity and an indicator of disadvantaged family background. The motivation for this specification is drawn from analysis similar to that in Section IV, where I document that individuals from more advantaged backgrounds seem to complete the same amount of schooling regardless of college proximity. This argument is strengthened once the sample is restricted to those ages 14-19 in 1966. Figure 3 replicates the analysis in Figure 2, except the sample is restricted to ages 14-19 in 1966 and current location is not conditioned upon. The substantial increases in the CDF difference just after high school graduation for the lowest and the second quartiles provide support for interpretation of college proximity as an instrumental variable that is causing additional education to be attained. Furthermore, the highest background quartile has essentially no difference in schooling by college proximity.

If the geographic wage premium is the same for the highest background quartile and the other quartiles, but only the three lower quartiles are induced to obtain extra schooling, then the main effect captures the geographic premium and the education coefficient is no longer biased by geographic wage differences. Note that once this premium has been captured, there is no longer

any theoretical justification for including the endogenous indicator for living in an SMSA in 1976 in the specification.

With these considerations in mind, Table V presents some additional results. As a benchmark, Column 1 presents the estimate of the effect of schooling on wages conditional on demographics for the sample ages 14-19 in 1966, using college proximity as the instrumental variable. This is the same specification as reported in column 4 of Table 1, except that main effects for background quartile are included for comparability with columns 2 and 3. The schooling coefficient is .094. Using the idea that the highest background quartile can be used to identify the geographic wage premium, a main effect of college proximity is included in column 2. The instrumental variable is formed as the interaction of college proximity and an indicator for background quartile 2, 3, or 4. The schooling coefficient rises slightly to .123, and the main effect of college proximity is estimated to be negative but insignificant. Dropping the potentially endogenous indicators for current location on the premise that the main effect for college proximity is capturing the geographic wage premium, column 3 shows that the schooling coefficient is .110, and the geographic wage premium is essentially zero. Given the available data, the specification in column three offers a superior method to control for geographic wage differences than the current location variables used in Table I, and these geographic wage differences appear to be fairly inconsequential.

In an attempt to obtain more precise estimates, I also utilized some new data. First, I incorporated additional years of panel data from 1976-1980 for individuals in the NLSYM, but find that the effect on the precision of estimates of the return to schooling is negligible after accounting for the covariance structure of individual wages over time. Second, I created a

dataset of nearly identical variables for another cohort of individuals, 4125 young men ages 14-19 in 1979 surveyed in the National Longitudinal Survey of Youth (NLSY), by matching data on the location of four-year colleges (National Center for Education Statistics, 1978) to county-level NLSY geocode data. Again, the IV estimates of returns to schooling are very imprecise, with a standard error of 0.11 and a point estimate of 0.28 using 1990 hourly wage data in the NLSY. In the first stage estimation, the indicator for college-in-county is only associated with 0.18 years of additional schooling in the NLSY, which is about half the magnitude of the first stage relationship reported in for the NLSYM cohort in Section III. Thus it appears that impact of college proximity on educational attainment has weakened over the thirteen years separating these cohorts, as college education has become more commonplace in the population.

VII. Conclusion.

In this paper I have offered an interpretation of an instrumental variables estimate of the return to schooling as a weighted average of causal effects, using the example of college proximity as an instrumental variable. I replicate earlier work in which Card (1995b) concludes that the returns to education are about 10 to 14 percent for individuals whose schooling is affected by the instrumental variable of distance to college, as opposed to a typical OLS estimate of 8 percent.

One of the most important questions raised by this instrumental variable is its external validity: what subgroups of the population do these estimates tell us about? I find that most of the individuals affected were from more disadvantaged family backgrounds, particularly with lower parental education. I quantify the weights implicit in the instrumental variables estimates,

and show that 50 percent of the weight is on the lowest family background quartile, and only 26 percent is on the top two quartiles. Also, using family background quartiles as a proxy for discount rate differences, my results show that proximity to college tends to cause high discount rate individuals to graduate from high school and attend some college, while proximity tends to cause low discount rate individuals to increase the number of years they attend college.

A second important question for interpretation is the internal validity of this instrumental variable: does it identify a causal effect? I discuss several specification issues in this estimation, most notably the importance of focusing analysis on younger individuals and the use of the group from the most advantaged background (whose schooling is largely unaffected by college proximity) to account for potential geographic wage premia. Point estimates based on specifications that address these considerations range from .094 to .123, but they are imprecisely estimated. There does not appear to be a wage premium associated with growing up in county that had a college. The empirical evidence does cast some doubt on the validity of college proximity as a legitimate instrument: presence of a college nearby appears to affect the educational attainment of a few individuals who would otherwise have only completed ninth or tenth grade. Since the magnitude of the estimates of the return to education appears to be at least 8%, I cautiously conclude that the group affected by college proximity (mainly individuals from disadvantaged family backgrounds) has a marginal benefit from education at least as high as that thought to be typical of the population based on OLS estimates.

Although analysis of the internal validity of this particular application suggests that the reported point estimates of the returns to education should be viewed with some caution, the synthesis of modeling and methodology used to examine the external validity is generally

applicable to future studies of schooling using instrumental variables to identify causal effects. Those affected by this instrumental variable are of particular interest, for instance, because they are the ones responsive to changes in costs of education that are often considered as policy initiatives. Contrary to some conventional beliefs, interventions that target assistance so that disadvantaged individuals can obtain more schooling may have substantial benefits.

References

- Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings." *Quarterly Journal of Economics*, 106, 979-1014.
- Angrist, Joshua D., and Guido Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association*, 90, 431-442.
- Ashenfelter, Orley, and Alan B. Krueger. 1994. "Estimates of the Economic Returns to Schooling from a New Sample of Twins." *American Economic Review*, 84, 1157-1173.
- Ashenfelter, Orley, and Cecilia E. Rouse. 1998. "Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins." *Quarterly Journal of Economics*, 113, 253-284.
- Becker, Gary S. 1975. *Human Capital*. Chicago: University of Chicago Press.
- Center for Human Resources Research. 1994. "NLS User's Guide." Columbus: Ohio State University.
- Card, David E. 1995a. "Earnings, Schooling, and Ability Revisited," *Research in Labor Economics*, 14, 23-48.
- 1995b. "Using Geographic Variation in College Proximity to Estimate the Returns to Schooling." In *Aspects of Labour Market Behaviour: Essays in Honor of John Vanderkamp*,

eds. L. N. Christofides *et al.*, Toronto: University of Toronto Press, pp. 201-221.

----- 1995c. Data and variable definitions. Personal communication, April 20.

Griliches, Zvi. 1977. "Estimating the Returns to Schooling: Some Econometric Problems."
Econometrica, 45, 1-22.

Kane, Thomas J., and Cecilia E. Rouse. 1993. "Labor Market Returns to Two and Four Year
Colleges: Is a Credit a Credit and Do Degrees Matter?" NBER Working Paper No. 4268.

Lang, Kevin. 1993. "Ability Bias, Discount Rate Bias, and the Return to Education."
Unpublished paper, Boston University.

National Center for Education Statistics. 1978. *ICPSR 7647: Higher Education Institutional
Characteristics, 1977*, Ann Arbor, MI: Inter-university Consortium for Political and Social
Research.

TABLE I

Regression of Log Hourly Wage in 1976 on Schooling:
Instrumental Variables -- Proximity to Four Year College, Age, Age²

Estimation technique	OLS	IV	IV	IV
Card: Table-column/row	(T2-c2)	(T3-c5B)	(T3-c6B)	(T4-r7)
Schooling coefficient (std. error)	0.075 (.003)	0.122 (.046)	0.132 (.049)	0.094 (.064)
Quadratic Experience, Race, Original Location Current Location	yes	yes	yes	yes
Background	no	no	yes	yes
Sample (Age in 1966)	14-24	14-24	14-24	14-19
N = number of observations	3010	3010	3010	2037

Note: Experience = Age - Schooling - 6. Schooling, experience, and experience squared are treated as endogenous, with an indicator for college in county, age, and age-squared as excluded instruments. "Background" variables are highest grade attained by mother and by father and 12 indicators: residence with both natural parents at age 14, residence with mother only at age 14, missing mother's and missing father's education, and eight interactions of parental education. "Original Location" variables are indicators for nine regions of residence in 1966 and an indicator of residence in an SMSA in 1966. "Current location" variables are an indicator for living in an SMSA, and for living the Southern U.S.

TABLE II

Probability of Characteristics by Family Background Quartile

Quartile	Lowest	3rd	2nd	Highest
Mother Reported High School Graduate	0	.05	.55	.97
Father Reported High School Graduate	0	.06	.31	.84
Lived with Both Parents at Age 14	.54	.73	.86	.93
Race is Not Black	.41	.67	.88	.96
Lived in SMSA at Age 14	.46	.64	.70	.74
College in County in 1966	.59	.67	.71	.73

Note: Background Quartile is computed as follows. Following Card (1995b), a predicted value is estimated from a regression of schooling on indicators for age and race as well as original location and background variables defined in Table I for the sample of 1163 observations with no college in their county in 1966. The 25th, 50th, and 75th percentiles of the predicted values from this sample were used to group all 3010 observations with wage and education data into four quartiles.

TABLE III

Decomposition of IV weighting by Family Background Quartile

Family Background Quartile	w_q	ΔS_q	$(w_q \Delta S_q) / \Delta S$
Lowest	.19	0.92 (.18)	.50
3rd	.24	0.35 (.18)	.24
2nd	.28	0.19 (.19)	.15
Highest	.29	0.13 (.20)	.11

Note: w_q is the fraction in each quartile. ΔS_q in column two is the coefficient on quartile interactions from a regression of highest grade completed on age, region and background characteristics (defined in Table I) as well as main effects and interactions of background quartile with the college proximity indicator -- as in the first stage of an IV estimate based on equation (5) with the interactions serving as Z. ΔS is the average change in schooling over the four quartiles.

TABLE IV

Difference in Highest Grade Completed at age 25
by College Proximity in 1966 for Five Age groups

Age in 1966:	14-15	16-17	18-19	20-21	22-24
Interaction of Age and College Proximity	.25 (.18)	.24 (.18)	.29 (.21)	.66 (.25)	.65 (.21)

Note: Standard errors in parentheses. Additional regressors are 11 indicators for age, as well as race, original location and background as defined in Table I.

TABLE V

Alternative Specifications for Instrumental Variables Estimates
of the Effect of Schooling on Log Hourly Wage in 1976

Schooling	.094 (.065)	.123 (.058)	.110 (.060)
College Proximity main effect	—	-.010 (.029)	.005 (.031)
Background Quartiles forming Proximity Instrument	1-4	2-4	2-4
Current region	yes	yes	no
Quadratic Experience, Race, Original Location	yes	yes	yes
Family Background and Quartile	yes	yes	yes
N	2037	2037	2037

Note: Standard errors are in parentheses. Column 1 presents results from the same data and specification from column 4 in Table I using college proximity, age and age² as instruments – except that three indicators for quartile an indicator for background quartile are also included as main effects. In Column 2, college in county is included as a main effect, and the excluded instrument is an interaction of college in county and background quartile 2-4, effectively using the top quartile to identify the main effect of proximity. Column 3 is the same as Column 2, except that indicators for current region are not used.

Figure 2: CDF Difference by Family Background Quartile

Figure 3: CDF Difference by Family Background Quartile