Estimating the Returns to Education: A Methodological Review of Card (1993)

YITING TANG Student ID: 530756125 ECMT3997 Final Report

June 2, 2025

1 Introduction

Estimating the causal effect of education on earnings remains a fundamental challenge in labor economics. While human capital theory posits that additional years of schooling enhance productivity and hence income [6, 17], empirically isolating this effect is far from straightforward. The core difficulty lies in the endogeneity of educational attainment: individuals with greater innate ability or family support are more likely to pursue higher education and to earn more, leading to upward bias in OLS estimates due to omitted variable confounding [12].

In response, a growing literature has adopted quasi-experimental methods to obtain credible causal estimates. This "credibility revolution" [3] emphasizes research designs that generate plausibly exogenous variation in the treatment of interest. Among these, the instrumental variables (IV) approach plays a central role. However, its success hinges on finding instruments that are both relevant and valid—conditions that are difficult to satisfy in practice, especially in the context of education.

Card (1993) [8] offers a seminal solution by proposing college proximity as an instrument for years of schooling. The key identifying assumption is that individuals who grew up near a college are more likely to attend one, yet conditional on covariates, geographic proximity is assumed uncorrelated with unobserved earnings determinants. This identification strategy is illustrated in Figure 1, which presents a directed acyclic graph (DAG) summarizing the assumed causal structure of Card's IV design. Building on this structure, Card applies a two-stage least squares (2SLS) strategy using data from the National Longitudinal Survey and finds that the IV estimate of the return to schooling (13.2%) substantially exceeds the OLS estimate (7.3%), indicating the presence of ability bias. Card's methodological innovation has since influenced numerous subsequent studies across various contexts, such as labor markets and educational policy.

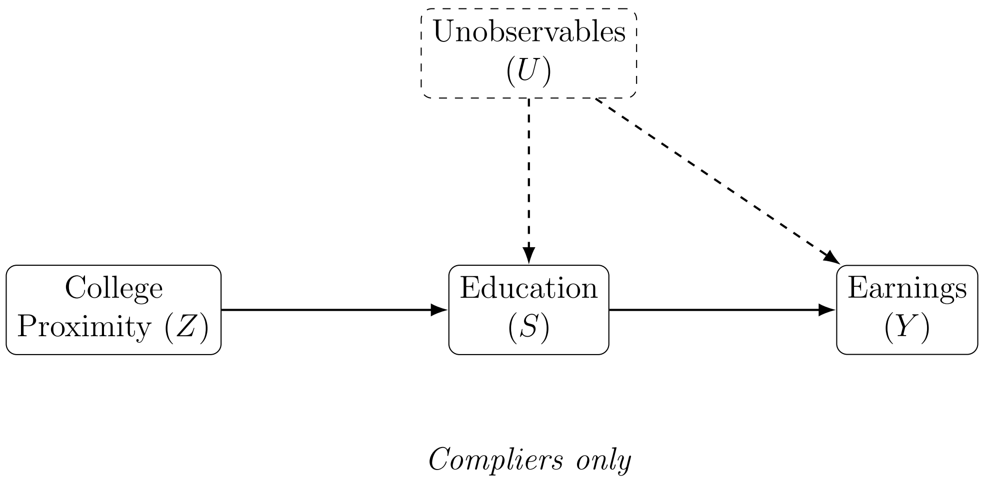

Figure 1: **Directed Acyclic Graph (DAG) of Card's IV Identification Strategy.** The instrument (Z) affects education (S), which in turn affects earnings (Y). Unobserved confounders (U) influence both S and Y, biasing OLS estimates. The exclusion restriction requires that Z affects Y only through S. The causal effect identified by 2SLS is Local Average Treatment Effect (LATE), relevant for compliers only.

Nevertheless, concerns have emerged regarding the external validity and identifying assumptions of this approach. Recent work has highlighted the potential endogeneity of residential sorting [9][16], the economic spillovers of university towns [24], and the evolving relevance of geographic proximity in the age of digital and hybrid education. These developments call for a reassessment of the Card identification strategy, especially under shifting urban structures and technological change.

This report revisits Card (1993)[8] methodologically, aiming to unpack the internal logic of the IV design, evaluate the robustness of its identifying assumptions, and explore its implications for contemporary education policy. This paper is organized into five sections. Section 1 introduces the background and central questions of Card (1993).

Section 2 formalizes the methodological framework, focusing on causal identification via the instrumental variable (IV) approach and the Local Average Treatment Effect (LATE). Section 3 presents the empirical regression results and evaluates the underlying identification assumptions. Section 4 offers a critical assessment from four perspectives: the exogeneity of the instrument, the policy relevance of LATE, the structural bias of 2SLS, and the historical constraints of the identification strategy. Section 5 concludes by summarizing the contributions and suggesting future directions for identification strategies under evolving spatial and institutional conditions.

2 Methodology and Identification Strategy

To identify the causal effect of education on earnings, Card employs an instrumental variable (IV) approach based on geographic proximity. Specifically, whether an individual lived near a four-year college during adolescence serves as an instrument for educational attainment, aiming to isolate the exogenous variation in schooling. From a statistical standpoint, years of schooling are potentially endogenous, as they may correlate with unobserved factors such as innate ability or family support. Direct estimation using OLS would therefore yield biased results.

Card proposes a structural framework consisting of the following equations:

$$S_i = X_i\gamma + v_i \tag{1}$$

$$y_i = X_i\alpha + S_i\beta + u_i \tag{2}$$

where X_i denotes a vector of observed control variables, including gender, race, region, and family background.

Given that $\mathrm{Cov}(S_i, u_i) \neq 0$, OLS estimation of β is biased. Card thus advocates the use of IV estimation. The instrument is defined as a binary variable $Z_i \in \{0, 1\}$, where $Z_i = 1$ if the individual lived near a four-year college during youth and 0 otherwise.

The validity of this IV strategy hinges on two key assumptions:

1. **Instrument Relevance:**

$$\text{Cov}(Z_i, S_i) \neq 0$$

   Geographic proximity must be significantly correlated with educational attainment.

2. **Exclusion Restriction:**

$$\text{Cov}(Z_i, u_i) = 0$$

   Conditional on observed covariates, proximity should not directly affect earnings except through its effect on education.

Under these assumptions, the causal effect of education can be identified via Two-Stage Least Squares (2SLS):

First Stage:

$$S_i = \pi_0 + \pi_1 Z_i + \gamma' X_i + v_i \tag{3}$$

Second Stage:

$$y_i = \alpha_0 + \alpha_1 \hat{S}_i + \delta' X_i + u_i \tag{4}$$

where $\hat{S}_i$ is the predicted value of schooling from the first-stage regression.

However, when the impact of education varies across individuals—such as by socioeconomic background or ability—OLS and IV estimators may identify different causal parameters. Imbens and Angrist (1994) further note that when treatment effects are heterogeneous[2], 2SLS identifies the *Local Average Treatment Effect (LATE)*:

$$\text{LATE} = \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i(1) > S_i(0)] \tag{5}$$

That is, it represents the average causal effect for compliers, defined as individuals whose educational decisions are affected by the instrument.

This identification also relies on the *monotonicity assumption*, which rules out "defiers":

$$S_i(1) \geq S_i(0), \quad \forall i \tag{6}$$

While monotonicity cannot be directly tested, it is often considered plausible in contexts where proximity lowers educational costs.

Another crucial concern is the strength of the instrument. Staiger and Stock (1997) warn that if the first-stage F-statistic is below 10, weak instrument bias may arise. In Card's analysis, the F-statistic exceeds 20, suggesting that the proximity instrument (nearc4) is sufficiently strong for causal inference[20].

In summary, Card's approach offers a clear and operational IV strategy: leveraging plausibly exogenous variation in college proximity to estimate the causal impact of education on earnings under standard identification assumptions.

3 Empirical Results and Evaluation of Identification Strategy

This section presents the empirical findings of Card (1993) and evaluates the credibility of his IV strategy[8]. The analysis focuses on three aspects: (1) comparing OLS and IV estimates of the returns to education; (2) testing the strength of the instrument; and (3) assessing the plausibility of the exclusion restriction.

3.1 OLS Estimates: Baseline Results

Card begins with a standard OLS regression to estimate the impact of years of schooling on weekly earnings. Controlling for race, experience (and its square), region, and urban residence, the results are summarized in Table 1.

Table 1: OLS Regression Results (Dependent Variable: Log Weekly Earnings)

Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**p-value**
Years of Education	0.0747	0.0035	21.35	0.000
Black Dummy	-0.1990	0.0182	-10.91	0.000
Work Experience	0.0848	0.0066	12.81	0.000
Experience Squared	-0.0023	0.0003	-7.22	0.000

The coefficient on years of education implies that each additional year of schooling increases weekly earnings by approximately 7.47%. This estimate is statistically significant and consistent with previous literature. However, as Card notes, this result may suffer from ability bias, since educational attainment is likely correlated with unobserved factors such as cognitive ability or family resources.

3.2 IV Strategy and First-Stage Regression

To address the potential endogeneity of schooling, Card employs an instrumental variable: whether an individual lived near a four-year college during adolescence (nearc4). The first-stage regression evaluates whether nearc4 significantly predicts years of education, the results are summarized in Table 2.

Table 2: First-Stage Regression Results (Dependent Variable: ed76)

Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**p-value**
College Proximity (nearc4)	0.3199	0.0879	3.64	0.000
Urban Residence	0.4022	0.1048	3.84	0.000
Regional FE (example)	0.5239	0.2675	1.96	0.050

The first-stage F-statistic is 182.13 (F(15, 2994), $p < 0.001$), far surpassing the conventional threshold of 10 recommended by Staiger and Stock (1997), thus confirming that the instrument is strongly correlated with education and satisfies the relevance condition[20].

3.3 Reduced-Form Regression and Exclusion Restriction

To preliminarily test the exclusion restriction, Card estimates the reduced-form regression of college proximity on earnings, the results are summarized in Table 3.

Table 3: Reduced-Form Regression (Dependent Variable: Log Earnings)

Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**p-value**
College Proximity (nearc4)	0.0421	0.0181	2.33	0.020

The coefficient on nearc4 is statistically significant but economically modest, suggesting that college proximity has only a weak direct effect on earnings after conditioning on education—supporting the exclusion restriction at least partially.

3.4 2SLS Estimates and Robustness

Card then implements Two-Stage Least Squares (2SLS) estimation using nearc4 as the instrument, the results are: summarized in table 4.

Table 4: IV Estimation Results (Dependent Variable: Log Earnings)

Model	**Coefficient (ed76)**	**Std. Error**	**z-Statistic**	**p-value**
2SLS	0.1315	0.0548	2.40	0.016
OLS	0.0747	0.0035	21.35	0.000

The IV estimate (13.15%) exceeds the OLS estimate (7.47%) by approximately 76%, consistent with OLS underestimating the true causal return due to ability bias, as illustrated in Figure 2.

Additionally, Card constructs an interaction term, *nearc4* $\times$ parental education, as an alternative instrument to better isolate the effect among marginal students from disadvantaged backgrounds. This modification yields an estimated return of 10.06% with a smaller standard error, suggesting some robustness to instrument specification.

3.5 Summary and Preliminary Evaluation of Identification Assumptions

In short, Card's IV strategy appears strong in terms of instrument relevance. While the exclusion restriction cannot be definitively tested, the modest direct effect of nearc4, along with multiple robustness checks, lends credibility to the identification strategy. The design largely conforms to the classical IV assumptions and provides a tractable method to estimate the causal effect of education on earnings among marginal populations.

4 Critical Assessment

4.1 The Exogeneity of the Instrument

Card (1993) proposes using college proximity during adolescence as an instrumental variable to identify the causal effect of education on earnings[8]. This strategy has been widely cited

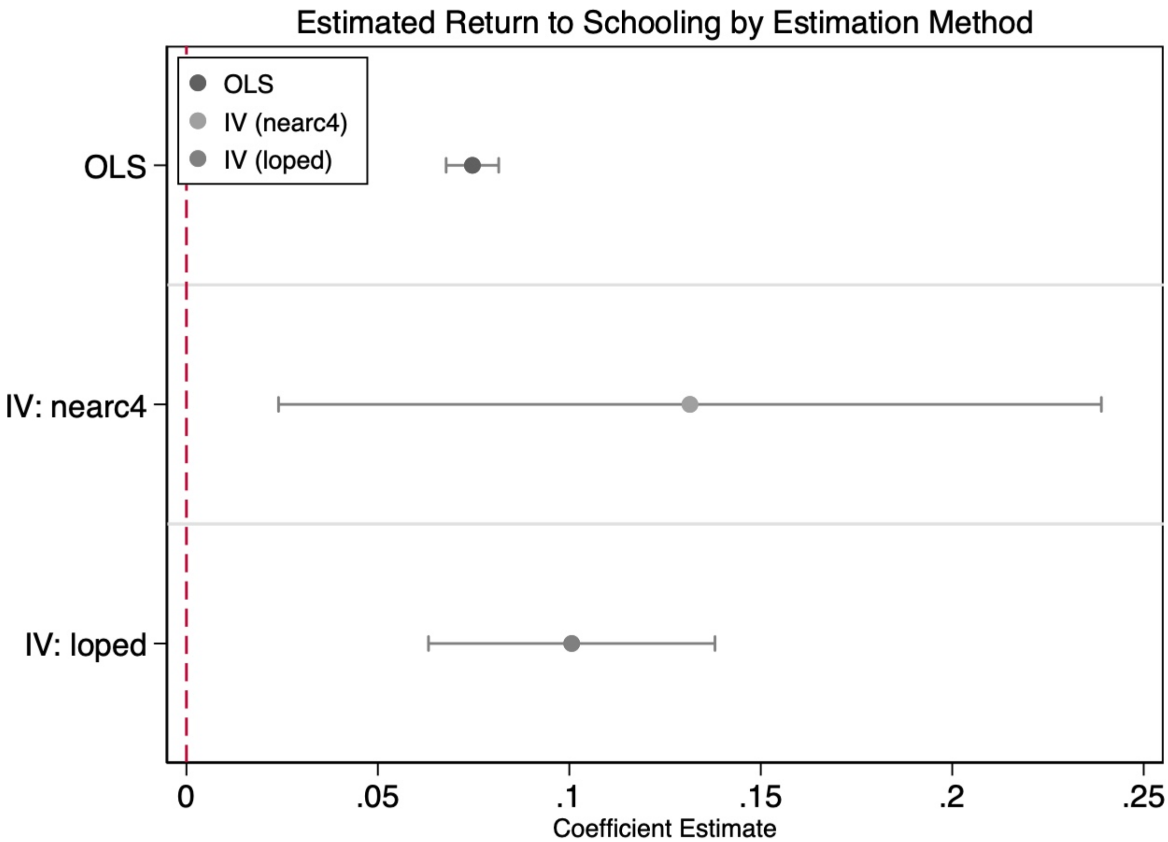

Figure 2: **Estimated Returns to Schooling by Estimation Method.** This figure compares the estimated causal effect of education on log wages using OLS and two IV strategies. The IV estimates, using college proximity (nearc4) and low parental education (loped) as instruments, are substantially higher than the OLS estimate, consistent with negative selection bias in OLS. Error bars represent 95% confidence intervals.

as a textbook example of IV identification in empirical microeconomics. However, from the perspective of modern identification theory, the exclusion restriction underlying this approach is open to challenge.

Card assumes that living near a college affects earnings only through its effect on educational attainment, and not via other channels. Yet this assumption overlooks two major concerns: (1) the endogeneity of residential choice by families; and (2) the broader regional spillovers associated with college locations.

First, Bayer and McMillan (2005) document that higher-income families are more likely to reside near colleges and in areas rich in educational resources[5]. This residential sorting implies that college proximity may reflect unobserved family preferences and social capital, rather than random assignment. Similar findings are reported in Black (1999) and Kane et al. (2006), particularly in high-demand urban or suburban neighborhoods, casting doubt on the exogeneity of the instrument[7][15].

Second, Valero and Van Reenen (2019) show that universities generate positive externalities for local economies, including increased employment, innovation, health services, and cultural amenities[24]. These regional spillovers could directly influence long-term earnings outcomes, violating the exclusion restriction.

Moreover, Card's sample is drawn from the 1976 wave of the National Longitudinal Survey of Young Men, focusing primarily on urban, middle- to high-income males. Urban residency (smsa76r) is strongly correlated with the instrument (nearc4): over 80% of those living near a college resided in urban areas, compared to just 46% among those who did not, as shown in Appendix Figure 3. This suggests that "college proximity" may proxy for urbanization and associated infrastructure advantages, rather than isolated geographic variation.

Parental education data reinforces this concern. Individuals living near colleges had significantly higher parental education: fathers averaged 10.26 years compared to 9.40, and mothers 10.50 years compared to 9.98, as shown in Appendix Figures 4 and 5. These differences reflect deeper disparities in socioeconomic background and social capital, which may not be fully accounted for by observable covariates, as shown in Appendix Figure 6.

While Card attempts to control for these differences through additional covariates and interactions (e.g., *nearc4* $\times$ parental education), these adjustments do not fundamentally resolve the structural selection bias stemming from non-random college location and residential choice. As Owens (2016) argues, access to colleges in the U.S. is spatially unequal, with high-quality institutions disproportionately clustered in affluent, predominantly white communities[19]. The assumption that *nearc4* approximates "as-if random" assignment is therefore questionable.

In sum, *nearc4* is unlikely to be a purely exogenous instrument. Its value is systematically related to family background, urbanization, and educational infrastructure. Without stronger controls or design innovations, the validity of the IV strategy may be compromised, leading to upward-biased estimates of the return to education.

4.2 The Policy Limitations of LATE

Even if one accepts the validity of the instrument, the treatment effect estimated via 2SLS reflects only a Local Average Treatment Effect (LATE), that is, the return to education for the subset of individuals (compliers) whose schooling decision is affected by the instrument [2].

In Card's context, these compliers are "marginal students", those who are on the fence about attending college and whose decisions are nudged by geographic accessibility. While Card's estimate may be informative for this group, it is not generalizable to always-takers or never-takers, nor to individuals living under different institutional or geographic conditions.

Heckman and Vytlacil (2005) warn that LATE estimates can vary substantially depending on the instrument used and may be highly sensitive to sample composition[13]. Deaton (2010) further critiques LATE for lacking clear policy relevance: policymakers often care about the Average Treatment Effect (ATE) or subgroup-specific effects, rather than effects for an unobserved, instrument-defined subgroup[11].

Thus, although Card's study represents a methodological breakthrough in applying quasi-experimental tools to observational data, the policy interpretability of the estimated return is limited. LATE provides a credible estimate of the effect of education on a specific population, but it should not be extrapolated without caution to broader contexts or used as a sole guide for national policy design.

4.3 Structural Estimation Bias in 2SLS

Since Angrist and Imbens formalized the LATE framework, Two-Stage Least Squares (2SLS) has become a default method in causal inference. Its popularity owes much to its simplicity and interpretability. However, as recent literature emphasizes, 2SLS may yield structurally biased estimates under treatment effect heterogeneity.

Sloczynski (2022) shows that in the presence of heterogeneity, 2SLS estimates a variance-weighted average treatment effect, rather than a simple average[21]. Specifically, the asymptotic 2SLS estimand converges to:

$$\operatorname{plim} \hat{\beta}_{2SLS} = \frac{\mathbb{E}\left[\delta_Y(X_i)\delta_D(X_i)\operatorname{Var}(Z_i \mid X_i)\right]}{\mathbb{E}\left[\delta_D(X_i)\operatorname{Var}(Z_i \mid X_i)\right]} \tag{7}$$

This structure places disproportionate weight on observations where the instrument's conditional variance is highest, typically around $p(Z_i = 1 \mid X_i) = 0.5$. In Card's study, this may elevate the influence of marginal compliers and bias the overall estimate upward.

Tübbicke (2023) re-analyzes Card's data using semi-parametric methods like propensity score matching, inverse probability weighting (IPW), and efficient covariate balancing (ECB). These estimates are 50–100% lower than the 2SLS estimate, suggesting that 2SLS may significantly overstate the return to education in this context[22].

In contrast, alternative estimators such as Abadie's (2003) κ-weighting and Heckman's Marginal Treatment Effect (MTE) framework offer more robust ways to estimate heterogeneous effects without structural weighting biases[5].

Therefore, while 2SLS remains a valuable benchmark, it should not be blindly trusted. Researchers should evaluate its implicit weighting structure, consider the distributional representativeness of the compliers, and apply semi-parametric or machine learning methods for sensitivity analysis to improve robustness and interpretability.

4.4 Historical Constraints of Card's Identification Strategy

Card (1993) candidly acknowledges that his sample, white males born in the 1960s and surveyed in the 1980s, has limited generalizability. Later studies (e.g., Chetty et al., 2014; Zimmerman, 2014) show that returns to education vary significantly by race, gender, geography, and parental income, suggesting that Card's findings may not extend to more diverse or contemporary populations[10].[18]

More fundamentally, his identification strategy relies on the structural assumption that geographic proximity to colleges determines educational opportunities. While this may have been plausible in the 1970s and 1980s, modern dynamics, such as online education , educational marketization, increased social mobility, and suburban expansion, have greatly weakened geographic constraints[23].

As such, Card's instrument (nearc4) may no longer meet the relevance or exclusion assumptions in present-day settings. Treating it as a timeless identification device risks ignoring the institutional and technological shifts that reshape access to education.

In sum, while Card's framework was methodologically pioneering for its time, its assumptions are embedded in a specific historical context. Modern causal inference must move beyond geographic instruments and develop new strategies that reflect the evolving nature of education delivery and inequality.

5 Conclusion

This report revisits Card's (1993) influential study on the causal returns to education, focusing on its identification strategy and empirical findings. By exploiting geographic proximity to four-year colleges as an exogenous instrument, Card addresses endogeneity concerns inherent in OLS estimates and finds that the IV-estimated return to schooling (13.2%) is substantially higher than the OLS estimate (7.3%).

The strength of Card's approach lies in isolating exogenous variation among a policy-relevant subpopulation, namely marginal college attendees from disadvantaged backgrounds, highlighting the broader implications for education access and mobility. At the same time, external validity concerns remain, given the historical specificity of the sample and evolving patterns in education and labor markets.

Despite these limitations, Card's study remains a methodological benchmark for applied microeconometrics and continues to inform debates on education policy and inequality. Future research should extend this framework by exploring alternative instruments, broader populations, and contemporary education contexts.

Overall, Card's core insight endures: expanding access to education generates substantial private and social returns, particularly for those facing the greatest barriers.

References

[1] Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

[2] Joshua D Angrist and Guido W Imbens. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

[3] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press, 2009.

[4] R. L. Basmann. A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica*, 25(1):77–83, 1957.

[5] Patrick Bayer, Robert McMillan, and Kim Rueben. Residential segregation in general equilibrium. *NBER Working Paper No. 11095*, 2005.

[6] Gary S. Becker. *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education, First Edition.* University of Chicago Press, 1964.

[7] Sandra E Black. Do better schools matter? parental valuation of elementary education. *Quarterly Journal of Economics*, 114(2):577–599, 1999.

[8] David Card. Using geographic variation in college proximity to estimate the return to schooling. *NBER Working Paper No. 4483*, 1993.

[9] Raj Chetty, John N Friedman, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. The opportunity atlas: Mapping the childhood roots of social mobility. *NBER Working Paper No. 25147*, 2018.

[10] Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. Where is the land of opportunity? the geography of intergenerational mobility in the united states. *Quarterly Journal of Economics*, 129(4):1553–1623, 2014.

[11] Angus Deaton. Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48(2):424–455, 2010.

[12] Zvi Griliches. Estimating the returns to schooling: Some econometric problems. *Econometrica*, 45(1):1–22, 1977.

[13] James J Heckman and Edward J Vytlacil. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738, 2005.

[14] Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

[15] Thomas J. Kane, Douglas O. Staiger, and Stephanie K. Reigg. School quality, neighborhoods and housing prices: The impacts of school desegregation. *NBER Working Paper No. 11347*, 88(9-10):1819–1842, 2005.

[16] Jens Ludwig, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, and Lisa Sanbonmatsu. Neighborhood effects on the long-term well-being of low-income adults. *Science*, 337(6101):1505–1510, 2012.

[17] Jacob A. Mincer. Schooling and earnings. In *Schooling, Experience, and Earnings*, chapter 3, pages 41–63. National Bureau of Economic Research, 1974.

[18] Pablo A. Mitnik, Victoria Bryant, Michael Weber, and David B. Grusky. New estimates of intergenerational mobility using administrative data. Working paper in Stanford University and U.S. Internal Revenue Service, 2015.

[19] Ann Owens. Inequality in children's contexts: Income segregation of households with and without children. *American Sociological Review*, 81(3):549–574, 2016.

[20] Douglas Staiger and James H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997.

[21] Tymon Słoczyński. When should we (not) interpret linear iv estimands as late? IZA Discussion Paper 14349, IZA Institute of Labor Economics, May 2021.

[22] Stefan Tübbicke. When to use matching and weighting or regression in instrumental variable estimation? evidence from college proximity and returns to college. *Empirical Economics*, 65(6):2979–2999, 2023.

[23] Miguel Urquiola. *Markets, Minds, and Money: Why America Leads the World in University Research*. Harvard University Press, Cambridge, MA, 2020.

[24] Anna Valero and John Van Reenen. The economic impact of universities: Evidence from across the globe. *Economics of Education Review*, 68:53–67, 2019.

Data and Code Availability

Figure 1 is generated using the `TikZ` package in LaTeX.

Figures 2, 3, 4, 5, and 6, as well as Tables 1, 2, 3, and 4, are produced using Stata 18.0.

The complete set of Stata do-files used for data cleaning, estimation, and figure generation is attached to this submission.

Appendix

Appendix A: Supporting Evidence on Instrument Validity

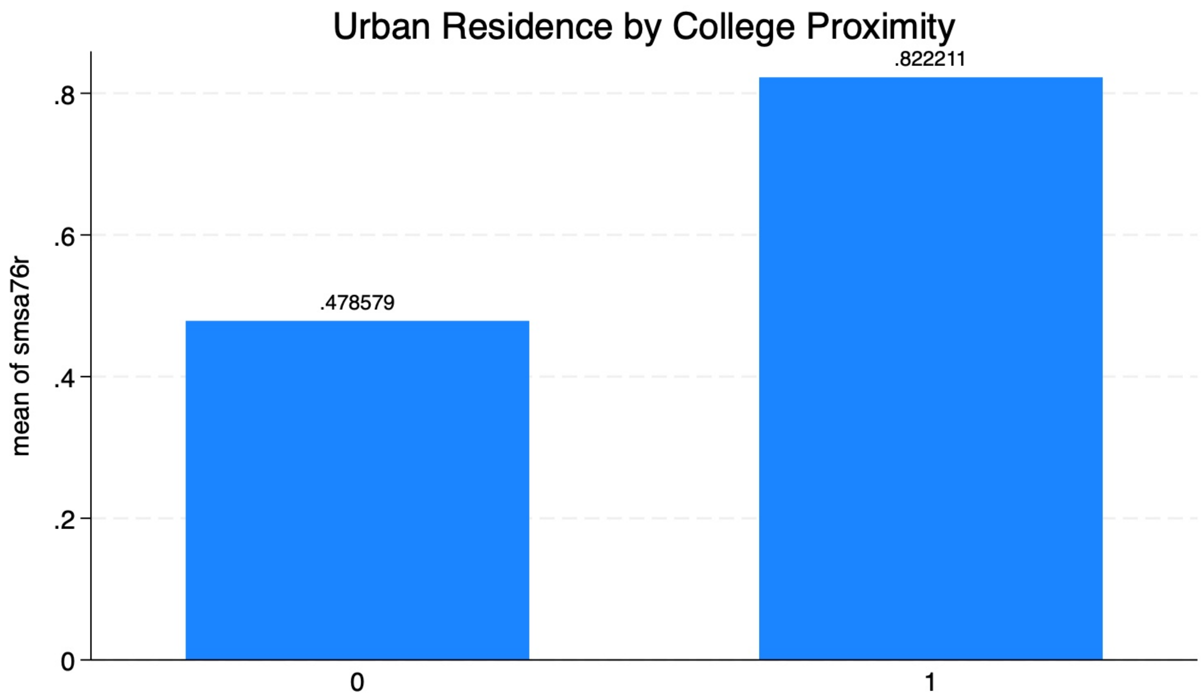

Figure 3: **Urban Residency Rates by College Proximity.** Individuals living near a four-year college (nearc4 = 1) are significantly more likely to reside in urban areas compared to those who do not. This suggests that the instrument may partially proxy for urban infrastructure and social capital, challenging the assumption that college proximity is as-good-as-random.

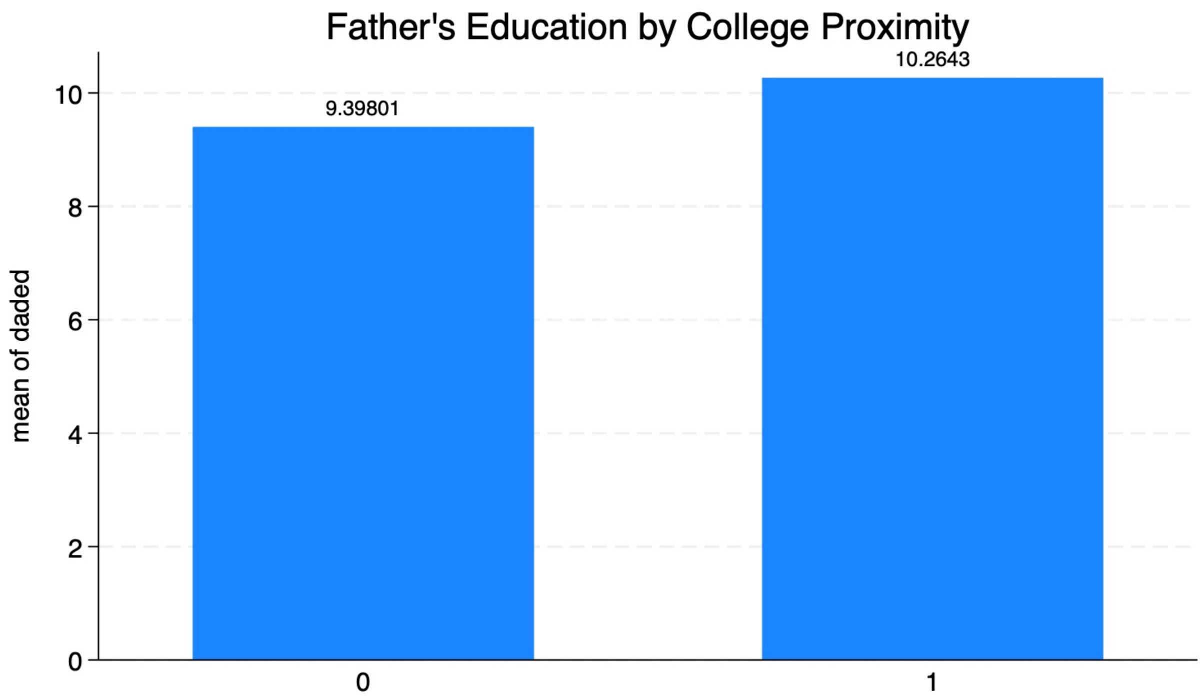

Figure 4: **Father's Education by College Proximity.** Respondents who lived near a four-year college (nearc4 = 1) had fathers with significantly higher average years of education, indicating that the instrument may be correlated with parental human capital. This raises concerns about residual confounding due to family background.

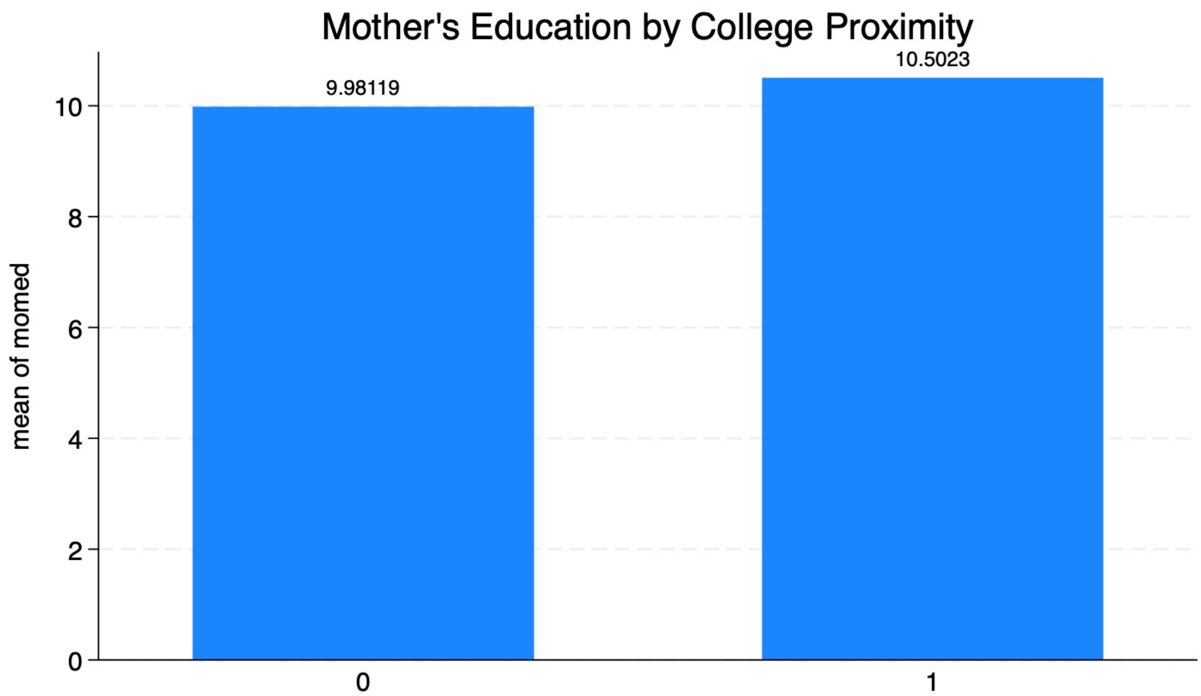

Figure 5: **Mother's Education by College Proximity.** A similar pattern is observed in maternal education: individuals living near a college had mothers with more years of schooling on average. This reinforces concerns that nearc4 may proxy for unobserved socioeconomic advantage.

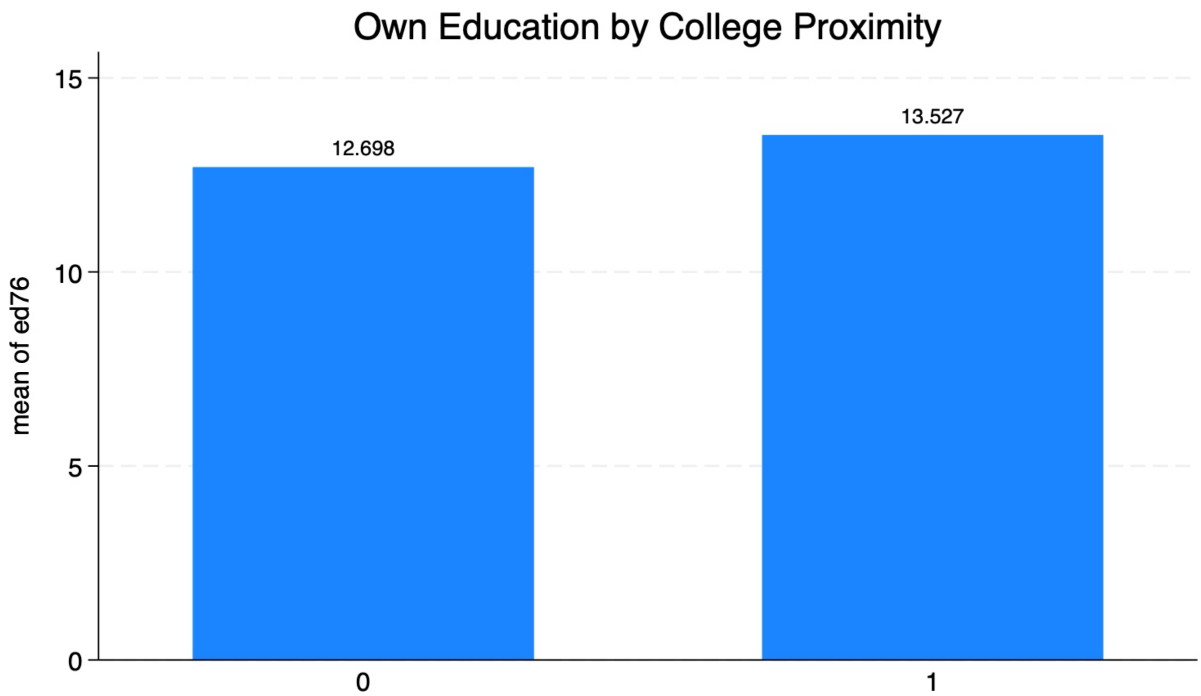

Figure 6: **Respondent's Education by College Proximity.** Individuals near a college exhibit substantially higher educational attainment, consistent with the relevance condition of the instrument. However, this may also reflect pre-existing differences tied to parental background or neighborhood characteristics.