# SLAM

## MICHIGAN

# SYMPOSIUM ON LEARNING ANALYTICS AT MICHIGAN

# The Application of Quasi-Experimental Methods in Education Research

**Stephen L. DesJardins**
**Professor & Director**
**Center for the Study of Higher and Postsecondary Education (CSHPE)**

**Brian P. McCall**
**Professor**
**CSHPE, Department of Economics, & Ford School of Public Policy**

**Rob Bielby, Allyson Flaster, Jeongeun Kim, Alfredo Sosa**
**Doctoral Students, CSHPE**

**Presented to Symposium on Learning Analytics at Michigan (SLAM)**
**November 30, 2011**

# Objectives

- Discuss the State of Conducting Rigorous Evaluations of Educational Programs, Policies & Practices

- Demonstrate the Application of Some Methods Now Being Used to Conduct Evaluations in Higher Education

- Provide an Opportunity for Our Students to Present to University Community & to Showcase Their Expertise

# Importance of Rigor in Education Research

- Systematically Improving Education Policies, Programs, Practices Requires an Understanding of "What Works"
- Goal Should be to Make **Causal** Statements
  – Without doing so "it is difficult to accumulate a knowledge base that has value for practice or future study" (Schneider, 2007, p. 2).
- However, Education Research Has Lacked Rigor & Relevance Quote

# Determining Causal Effects

- Randomized Controlled Trials are the "Gold Standard" for Determining Causal Effects
- Pros: Reduce Bias & Spurious Findings of Causality, Thereby Improving Knowledge of What Works
- Cons: Ethics, External Validity, Cost, Possible Errors Inherent in Observational Studies (measurement problems; "spillover" effects, attrition from study…)
- Possibilities: Oversubscribed Programs (Living Learning Communities, UROP…)

# Quasi- or Non-Experimental Designs

- Compared to RCTs, No Randomization
- Many Quasi-Experimental Designs
  - Many are variation of pretest-posttest structure without randomization
  - Apply when non-experimental ("observational") data used, which is often case in ed. research
- Pros: When Properly Done May Be More Generalizable Than RCTs
- Cons: Internal Validity
  - Did the "treatment" really produce the effect?

# A Common Causal Scenario
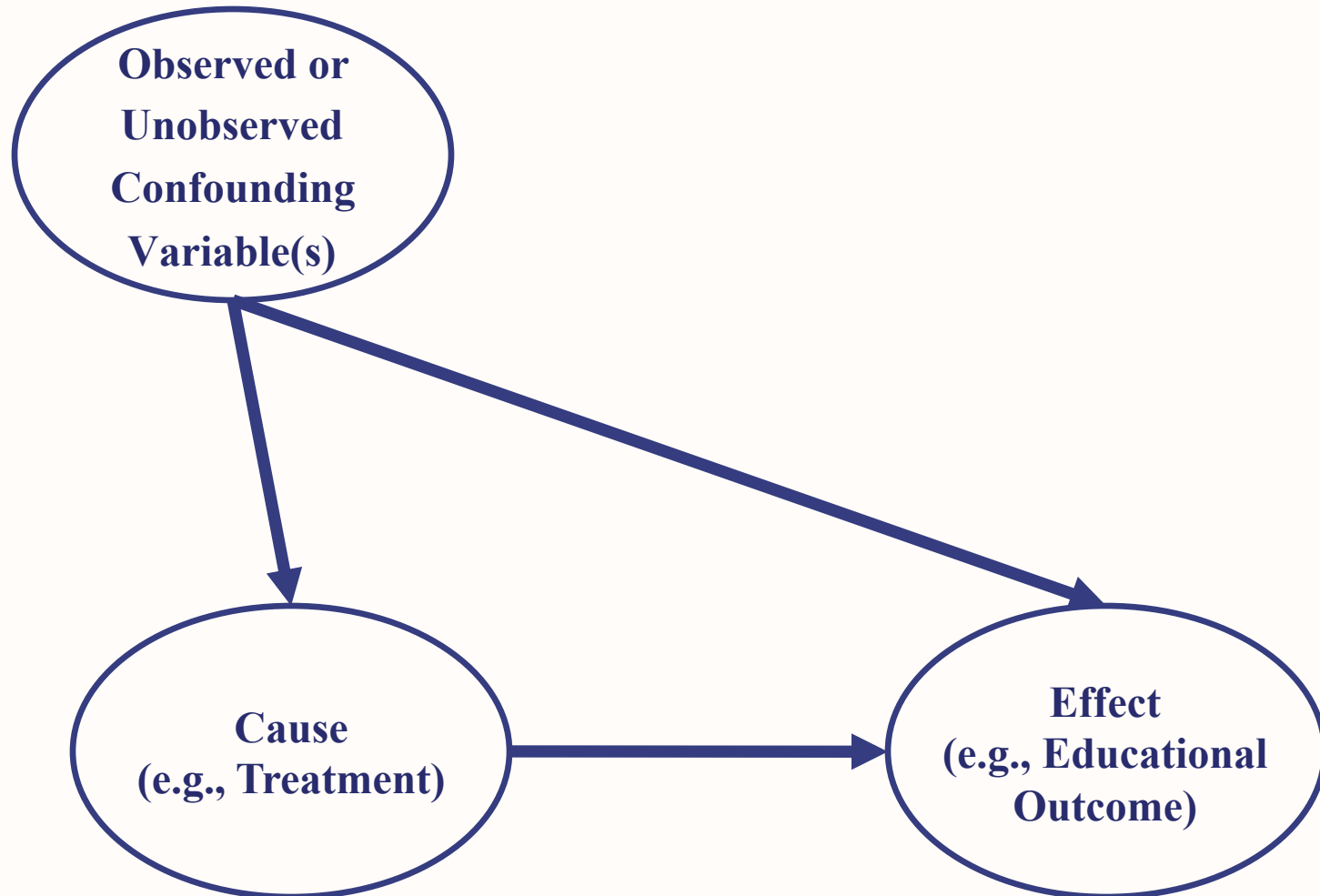
# Determining "Causal" Effects With Observational Data

- Often Difficult Because of Non-Random Assignment to "Treatment"
  - Example: Students often self-select into treatments (e.g., courses, interventions, programs…); may result in biased estimates when standard regression methods employed to determine effects of treatment

- Goal: Mimic Desirable Properties of RCTs

- Solution? Employ Designs/Methods That Account for Non-Random Assignment
  - We will demonstrate some of them today

# The "Counterfactual" Framework

- Provides Conceptual & Statistical Frame for Studying Causal Relationships
  - Owing to Donald Rubin (1974), Paul Holland (1986), and others
- Simple Definition of "Counterfactual": What Would Have Happened to the "Treated" Absent the Treatment?
- However, Individuals Only Observed in One State (treatment or control)
  - Therefore, problem is in est. "counterfactual" (comparison group) for the treated

# Applications

- Effect of Starting in CC vs. 4-Year on Subsequent Educational Outcomes (PSM)
- Does HS Course Selection Affect Subsequent Educational Outcomes? (IV)
- Effect of Gates Millennium Scholars Program on Ed. Choices & Outcomes (RD)
- In each we use methods to adjust for non-random assignment, thereby providing more rigorous statements of effects (relative to naïve statistical model)

# Symposium on Learning Analytics at Michigan

**Alfredo Sosa PhD Student**

**Center for the Study of Higher and Postsecondary Education**
**University of Michigan**

**November 30, 2011**
**Ann Arbor, MI**

# Matching Methods in Educational Research

Application:

"Answering Whether Attendance at a Two-Year Institution Results in Differences in Educational Attainment" (Reynolds and DesJardins, 2009).

*Higher Education: Handbook of Theory and Research, Volume XXIII*

# Does Where You Start College Affect Your Educational Attainment?

- Some Start in CCs, Other in 4-Year IHEs

- Inferential Problem: Self-Selection
  - Students who begin in CCs ("treated") may be very different (on observed & unobserved factors) than those starting in 4-years
  - Correlation between Prob(where one starts) & educational outcomes makes parsing causal effects from the observed & unobserved differences in students very difficult

# A Possible Remedy: Matching

- Intuition: Find Controls w/ Pre-treatment Characteristics the Same as the Treated

- Could Use Direct Matching: Problems

- Solution: Propensity Score Matching
  - Control for pre-treatment differences by balancing each group's observable characteristics using a single number, the "propensity score" (PS)

- Goal: Estimate Effect of Treatment for Individuals with Similar Characteristics

# Estimating the Propensity Score

- 1st: Estimate Probability of Treatment (PS) for Treated and Untreated
  - Typically done using logistic regression
- 2nd: Match Treated Cases to Untreated Cases with "Same" Propensity Score
  - Establishes counterfactual ("control" group)
- Then Estimate Differences in Outcomes Between Treated & Control Groups
  - Typically done using regression methods

# Goal of Matching

- Balancing the Treated and Control Groups on Observable Characteristics

- When Done Correctly, the Probability That Treated Observation Has Trait $X=x$ will be the Same as Probability that Untreated Case Has Trait $X=x$

# Matching vs. OLS Regression

- Matching weights the observation differently than does OLS in calculating the expected counterfactual for each treated observation. In OLS all the untreated units play a role in determining the expected counterfactual for any given treated unit. In matching, only untreated units similar to each treated unit have positive weight in determining the expected counterfactual.
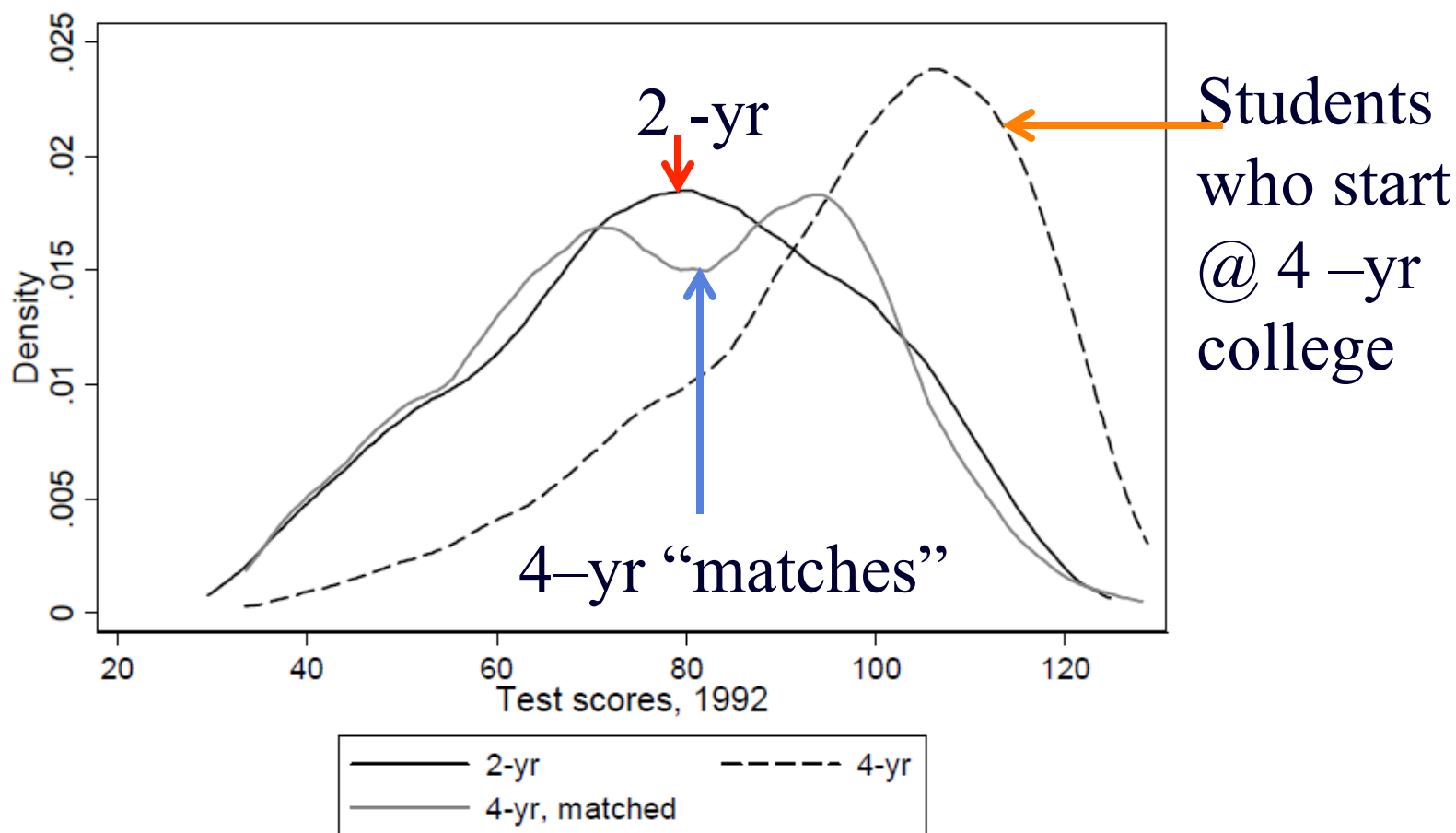
# Matching vs. OLS Regression (cont)

- Matching does not make the linear functional form assumption that OLS regression does.

- Matching helps identifying problems of lack of support (lack of balance on observable characteristics).

# Empirical Strategy to Study Effect of Starting at CC vs. 4-Year College

- Example of dependent variables examined:
  - Second and third year college retention rates
  - Completion of a bachelor's degree
- Variable of interest (treatment variable)
  - $T = 1$ if student attended a CC, $T = 0$ if attended a 4-year institution; estimate of this is of primary interest
- Data: National Education Longitudinal Study 1988 (NELS:88)

# Results From CC/4 Year Study

Figure 4: Distributions of Standardized Test Scores by Treatment Status

# Results From CC/4 Year Study

Table 6: ATT for attending a two-year instead of a four-year institution

| | OLS | | NN | | Kernel | | LLR | | RW | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Men** | | | | | | | | | | |
| Second year | -0.174 | * | -0.146 | * | -0.163 | * | -0.152 | * | -0.093 | * |
| | (0.019) | | (0.029) | | (0.025) | | (0.026) | | (0.047) | |
| Third year | -0.246 | * | -0.200 | * | -0.219 | * | -0.199 | * | -0.111 | * |
| | (0.020) | | (0.029) | | (0.024) | | (0.024) | | (0.039) | |
| AA | 0.182 | * | 0.128 | * | 0.118 | * | 0.111 | * | 0.106 | * |
| | (0.017) | | (0.022) | | (0.020) | | (0.022) | | (0.027) | |
| BA | -0.250 | * | -0.189 | * | -0.219 | * | -0.203 | * | -0.135 | * |
| | (0.019) | | (0.027) | | (0.022) | | (0.022) | | (0.032) | |
| Total credits | -23.320 | * | -20.563 | * | -22.940 | * | -21.077 | * | -11.658 | * |
| | (2.159) | | (3.380) | | (2.856) | | (2.891) | | (5.083) | |

Simple OLS that does not account for non-random assignment overestimates the effect of attending CC

# Results From CC/4-Year Study

- In this Sample, Starting at CC Lowers Attainment Compared to Starting at 4-Yr College

- CC Students Have:
  - Lower Retention Probs to 2$^{nd}$ & 3$^{rd}$ Yr
  - Lower Prob(BA Completion)

- Heterogeneous Treatment Effects:
  - Results similar for women/men, with women exhibiting slightly larger treatment effects

# Results (cont'd)

- OLS Estimates Overstate Treatment Effects Relative to Matching Estimates

- In this case, the "naïve" (in a statistical sense) OLS regression method performs poorly because of large differences in underlying characteristics of students who start in CC and those who start in 4-year institutions

# Thank you

# Using Instrumental Variables in Educational Policy Analysis

**Rob Bielby, PhD Student**

**Center for the Study of Higher and Postsecondary Education**
**University of Michigan**

**November 30, 2011**
**Ann Arbor, MI**

# Common Applications of IV Models

- When randomized controlled trials (RCT) are not available/possible

- When problems with observational data:
  - Omitted variables
  - "Reverse" causation

# Using Instrumental Variables in Educational Policy Analysis

- Application:
  - High school coursetaking and its impact on educational outcomes

# The Effect of High School Curriculum on College Completion
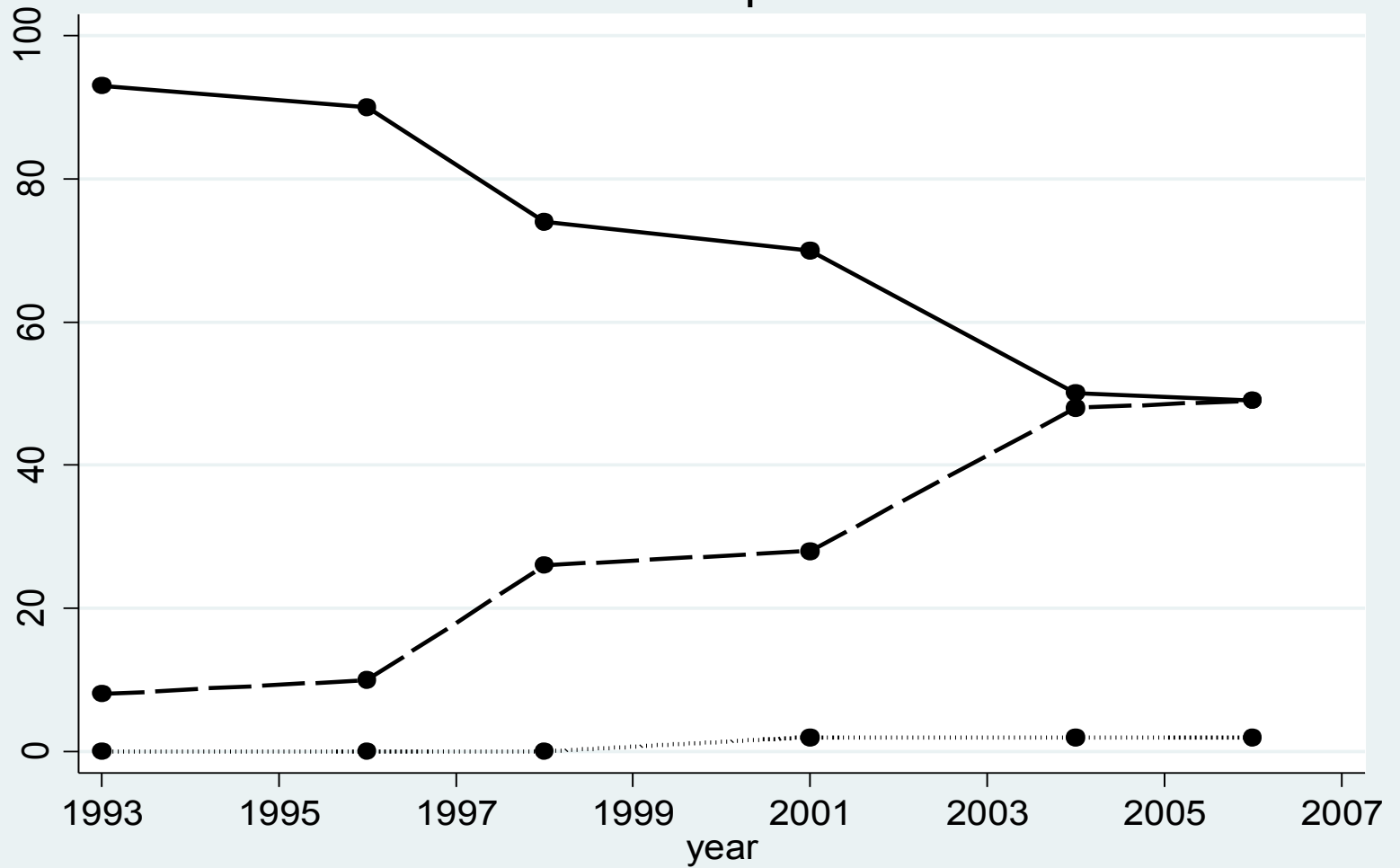
Policy Issue:

- In an attempt to increase college enrollment and graduation rates, states are increasingly requiring college-prep curricula for all high school students

- Research has found a positive association between taking rigorous coursework and college matriculation and success (Adelman, 2006)
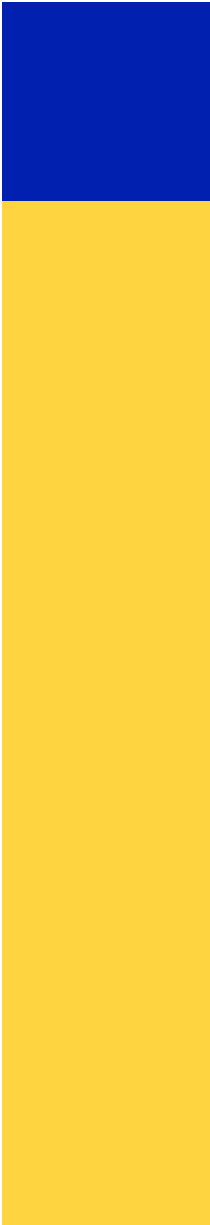
Math Requirements
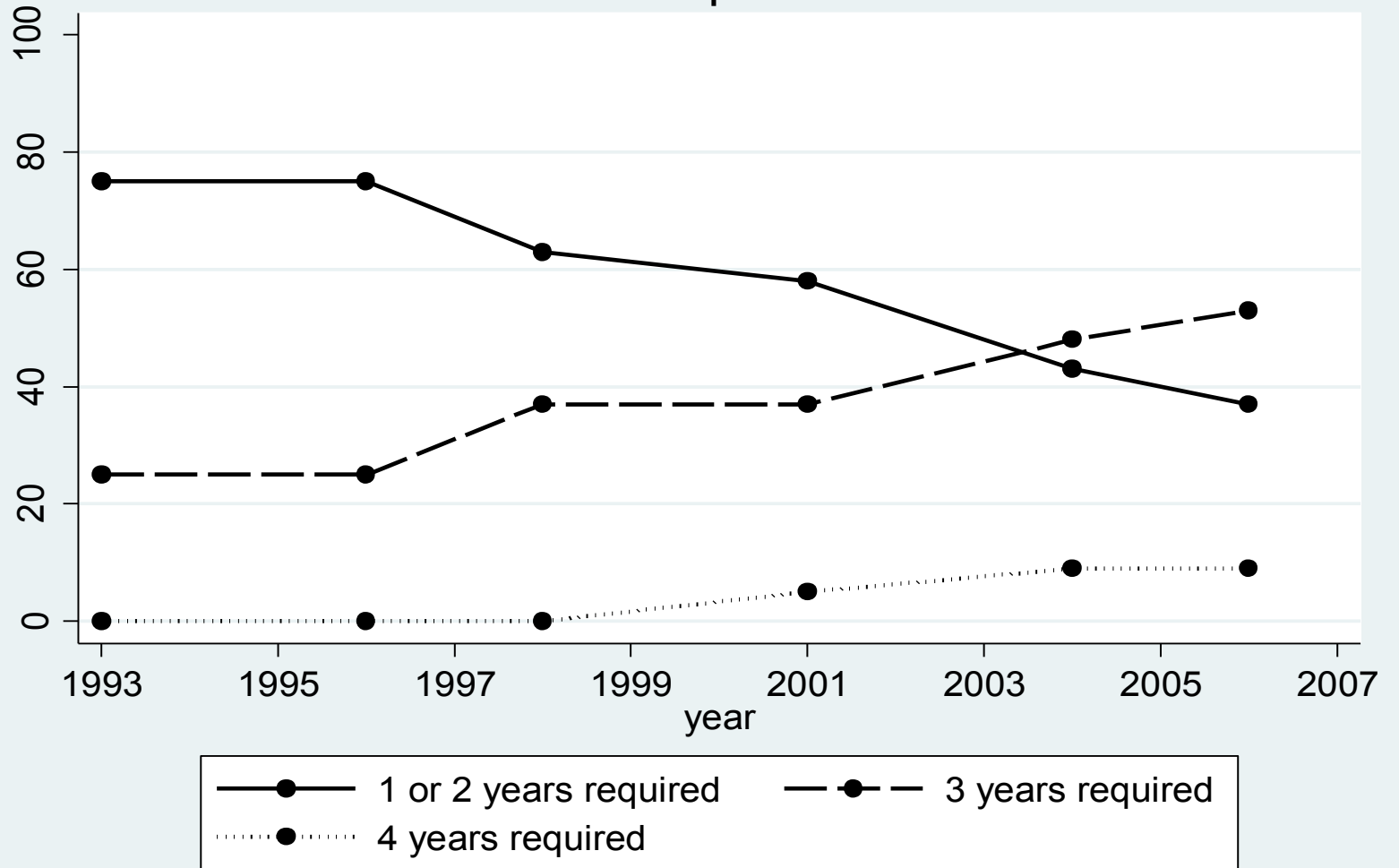
Science Requirements

# The Effect of High School Curriculum on College Completion

<u>Empirical Question</u>:

What is the **<u>causal</u>** effect of high school curricular choices on high school graduation, college access, college completion, and other outcomes related to higher education?

<u>Data</u>:

Education Longitudinal Study (ELS) of 2002

National Longitudinal Study of Youth (NLSY) of 1997

National Education Longitudinal Study (NELS) of 1988

# Inferential Problem: Self Selection

- Students choose to enroll in certain high school courses

- Motivators for making these choices may also be related to educational outcomes, which limits our ability to estimate causal relationships (omitted variables)

# Logic of Instrumental Variable Approach

- Variation in the independent variable of interest (course selection or program participation) has 2 components:
  - Selection (endogenous)
  - Other factors (exogenous)

# The Instrumental Variable Approach

- IV Selection
    - Seek out variable(s) that are strongly related to our independent variable of interest (course selection), but are unrelated to the eventual outcome variable (e.g., ed. outcomes)

- Goal
    - Use instruments to remove unwanted variation in predictor variables and allow for the estimation of a causal relationships

# The Instrumental Variable Approach

- Application: 2 Stage Least Squares
  - Stage 1: Instruments and other covariates are used to produce predicted levels of high school coursetaking
  - Stage 2: Predicted values of coursetaking are used as independent variables when modeling educational outcomes

# Our Project

- Examine Effect of High School Course Taking on Educational Choices and Outcomes (e.g., higher ed. aspirations; application to college; enrollment in college; college course performance and completion)

- IVs Used:
  - High school level maximum course offerings
  - Local labor market conditions

# Thank you

# Symposium of Learning Analytics at Michigan

**Jeongeun Kim & Allyson Flaster, PhD Students**

**Center for the Study of Higher and Postsecondary Education
University of Michigan**

**November 30, 2011
Ann Arbor, MI**

# Regression Discontinuity (RD)

- Use when subjects are assigned to either a "treatment" or "control" group based on a pre-specified cut score (e.g., standardized test)

- Assumption is that students are very similar close to the cut score used to determine treatment/not

- Calculate the average difference in the outcome between the treated and untreated groups using data only for students who are close to the cut score

# Regression Discontinuity (RD)

Application:

- How is the Gates Millennium Scholars (GMS) Program related to college students' time use and activities? (DesJardins, McCall, et al., 2010)
  - Effect of scholarship program on student success

# The Gates Millennium Scholars (GMS) Program

- Administered by Bill & Melinda Gates Foundation, provides $1 billion in scholarships over 20 year period
- Goal: Improve access for high achieving, low-income students of color
- Provides a scholarship that covers unmet need
- Selection criteria: Must be Pell-eligible, 3.3 high school GPA, **score on non-cognitive test**

# Estimating Effects of the Program
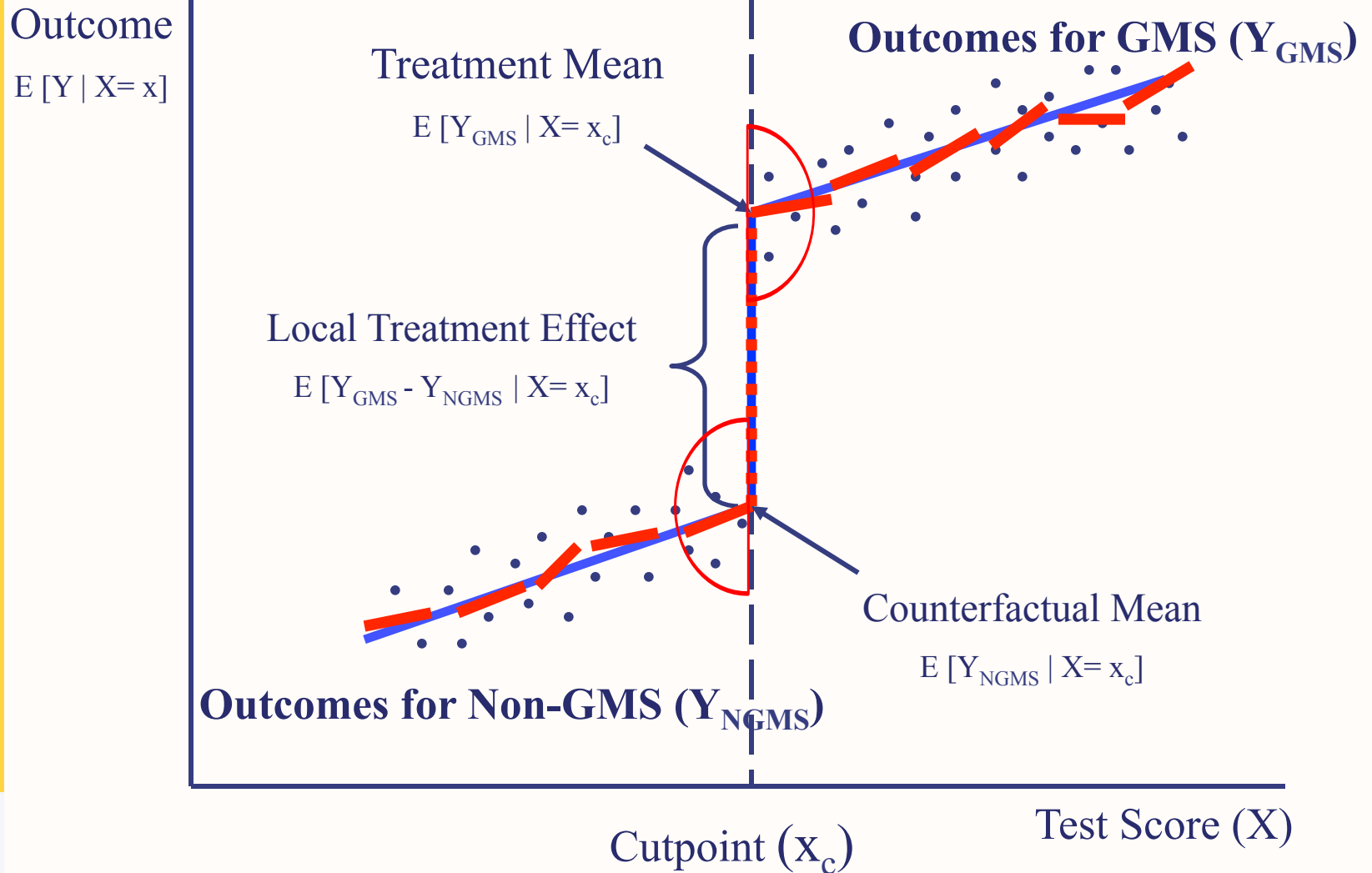
- Expected outcomes:
  - Students will respond to relaxed credit constraints by reallocating time devoted to various activities
  - May reduce time devoted to **working for pay**
  - May increase time for extra-curricular activities such as **community service**

- Other outcomes were examined

- Causal problem: Observed & unobserved student characteristics may influence outcomes

# Regression Discontinuity Design

- Forcing variable:
  - "Scoring" rule to assign the intervention to study units

- Use cutoff value for assignment
  - Below the cutoff = control group (Non-recipients)
  - Above the cutoff = treatment group (Recipients)

- Units just above and below the cut point: Distributed in an approximately random fashion, mimicking randomized trial

# The RD Estimation Strategy



Outcome
$E[Y | X = x]$

Treatment Mean
$E[Y_{GMS} | X = x_c]$

**Outcomes for GMS ($Y_{GMS}$)**

Local Treatment Effect
$E[Y_{GMS} - Y_{NGMS} | X = x_c]$

Counterfactual Mean
$E[Y_{NGMS} | X = x_c]$

**Outcomes for Non-GMS ($Y_{NGMS}$)**

Cutpoint ($x_c$)

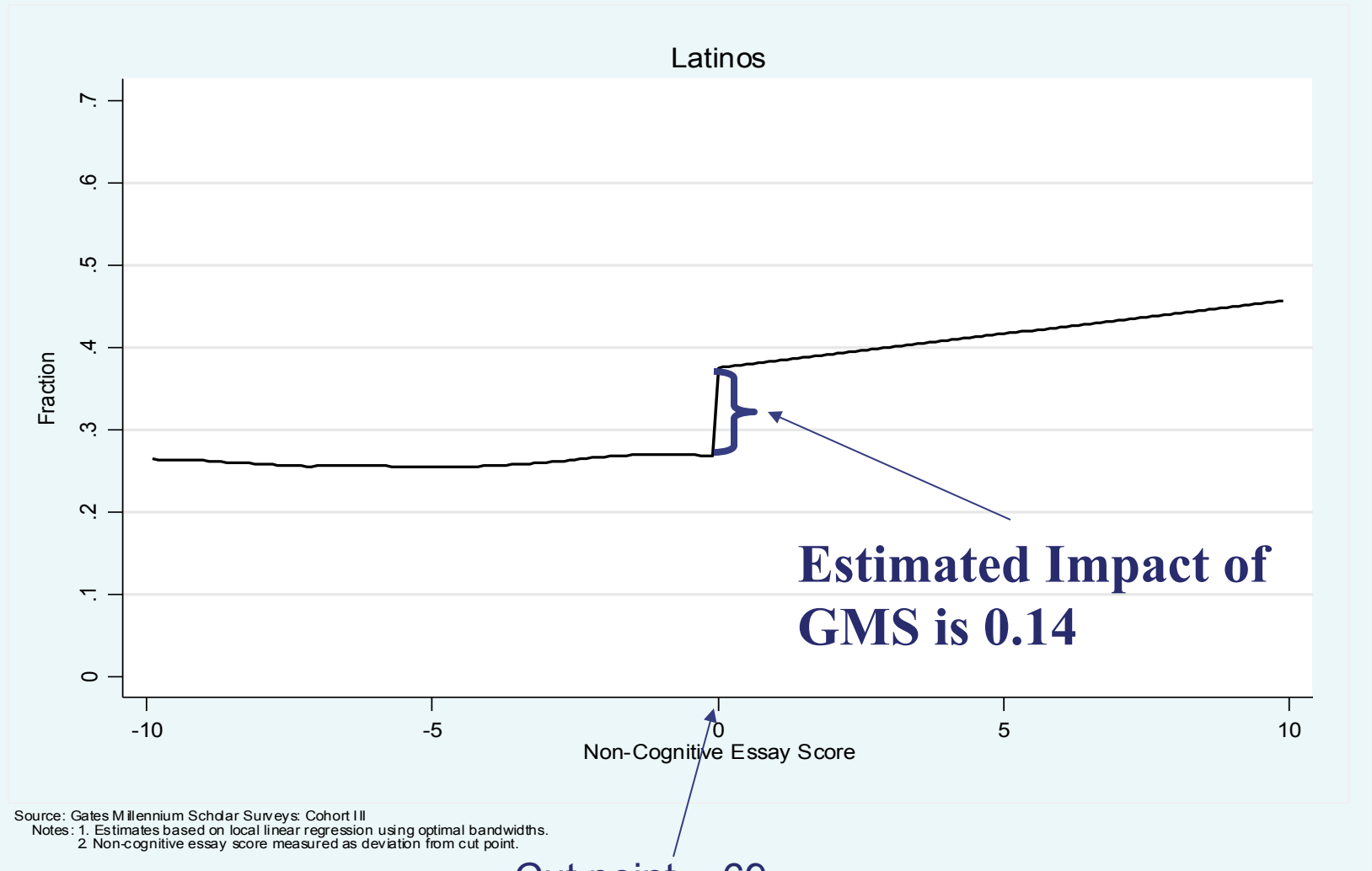Test Score (X)

# Empirical Strategy

- Selection mechanism
- Forcing variable: Non-cognitive scores
- Use cutoff value for assignment
    - varies by race/ethnicity and cohort
    - Below the cutoff = control group (Non-recipients)
    - Above the cutoff = treatment group (GMS)

# Empirical Strategy

- • Sharp RD

  - Assignment solely determined by a single index variable (i.e., non-cognitive test score)

- • Fuzzy RD

  - Not all students with scores above the cut point receive GMS: other eligibility criteria (i.e. Pell eligibility, high school GPA requirement)

  - Instrument Variable approach (IV)

    Stage1: non-cognitive score > Receipt of GMS

    Stage2: fitted value from stage 1 > outcomes

# Results: Higher Community Service



Participates in Community Service Often or Very Often: Junior Year

Cohort III

Latinos

**Estimated Impact of GMS is 0.14**

Cut point = 69

Source: Gates Millennium Scholar Surveys: Cohort III
Notes: 1. Estimates based on local linear regression using optimal bandwidths.
2 Non-cognitive essay score measured as deviation from cut point.

# RD Estimated Impact of GMS on Participation in Work Hours

|  | **Freshmen Year** | **Junior Year** |
|---|---|---|
| Combined | -4.14 | -5.28 |
|  | (0.000) | (0.000) |
| African Americans | -5.42 | -5.08 |
|  | (0.000) | (0.000) |
| Asian Americans | -5.93 | -4.98 |
|  | (0.000) | (0.000) |
| Latinos | -2.55 | -6.07 |
|  | (0.004) | (0.000) |

# Thank you

cshpe

A Legacy of Leaders · Making a Difference · 1957 - 2007

# Conclusions

- RCTs are Desirable in Terms of Making Causal Statements, But Often Difficult to Employ

- In Education We Often Only Have Observational Data But Methods Often Used to Make Statements About Treatment Effects are Typically Deficient

- Ultimate Goal: Make Strong ("Causal") Statements so as to Improve Knowledge of Mechanisms That Determine Whether Programs, Practices, & Interventions are Effective

- We Need to Employ These Methods More Broadly at Michigan to Ascertain "What Works"

# References

- Andrews, R., DesJardins, S. L., & Ranchhood, V. (2010). Estimating the Effects of the Kalamazoo Promise on Student College Choice. *Economics of Education Review,* 29(5): 722-737.

- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association, 91*(434), 444-455.

- Angrist, J. & Pischke, J. (2009). Mostly Harmless Econometrics. Princeton University Press.

- Cameron, A. & Trivedi, P. (2005). Microeconometrics, Methods and Applications. Cambridge University Press.

- DesJardins, S. L., McCall, B. P., Ott, M. & Kim, J. (2010). A Quasi-Experimental Investigation of How the Gates Millennium Scholars Program is Related to College Students' Time Use and Activities. *Educational Evaluation and Policy Analysis,* 32(4): 456-475.

- DesJardins & McCall: Numerous Reports to the Bill and Melinda Gates Foundation on the Effects of the GMS and WSA Programs (available from the authors).

# References

- Guo, S. and Fraser, M. W. (2010). *Propensity Score Analysis: Statistical Methods and Applications*. Thousand Oaks, CA: Sages Publications.

- Heckman, J. & Vytlacil, E. (2007). Econometric evaluation of social programs, part II. Handbook of Econometrics, Volume 6B 2007 Elsevier B.V.

- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81(396), 945–960.*

- Lesik, S. A. (2008). Studying the Effectiveness of Programs and Initiatives in Higher Education Using the Regression-Discontinuity Design. In John Smart (Ed.), *Higher Education: Handbook of Theory and Research* XXIII.

- McCall, B. P. and Bielby, R. (Forthcoming, 2012). Regression Discontinuity Design: Recent Developments and a Guide to Practice for Researchers in Higher Education.  In John Smart and Michael Paulsen (Eds.), *Higher Education: Handbook of Theory and Research* XXVII.

# References

- Morgan, S. L. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge, UK: Cambridge University Press.

- Reynolds, C. L., & DesJardins, S. L. (2009). The Use of Matching Methods in Higher Education Research: Answering Whether Attendance at a Two-Year Institution Results in Differences in Educational Attainment.  In John Smart (Ed.), *Higher Education: Handbook of Theory and Research* XXIII: 47-104.

- Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician, 39*(1), 33-38.

- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66(5), 688–701.*

# References

- Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating Causal Effects Using Experimental and Observational Designs.* Washington, DC: American Educational Research Association.

- Shadish, W. R., Cook, T. D., Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton-Mifflin

- Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment. *Journal of Educational Psychology*, 51: 309-317.

- Wooldridge, J. (2002). Econometric Analysis of Cross Section and Panel Data. The MIT Press.

- Zhang, L. (2010). The Use of Panel Data Models in Higher Education Policy Studies.  In John Smart (Ed.), *Higher Education: Handbook of Theory and Research* XXV: 307-349.

# Background Material

# Recent AERA Report on the Issue

- "Recently, questions of causality have been at the forefront of educational debates and discussions, in part because of dissatisfaction with the quality of education research…".   A common concern "revolves around the design of and methods used in education research, which many claim have resulted in fragmented and often unreliable findings" (Schneider, et al., 2007)

# The "Naïve" Statistical Model

- A typical program evaluation model is:

$$Y = \alpha + B\,X + \delta T + \varepsilon \quad (1)$$

- But often the "treatment" (T) is function of some of the same factors (Xs) & others (Zs)

$$T = \alpha + B\,X + \theta Z + \varpi \quad (2)$$

- Failure to account for this structural relationship often leads to biased estimates of the treatment effect ($\delta$)

# Many Methods to Do the Matching

- Interval or Cell Matching
  - Stratify by PS

- Nearest Neighbor
  - Treated units matched with control cases with similar PS; latter used as counterfactual for the former
  - Typically one-to-one matching
  - Question: Matching done with or w/o replacement?

# Methods to Do Matching (cont'd)

- Caliper Matching
  - NN matching within a range of PS
  - "Bandwidth" (range) chosen by researcher & is max. interval in which match is made

- Radius Matching
  - One-to many caliper matching
  - All matches within bandwidth equally weighted to construct the counterfactual.

# Methods to Do Matching

- Kernel/LLR Matching
  - Weight each untreated observation according to how *close the match is*
  - As match becomes worse the weight placed on the untreated unit decreases
- New: Optimal Matching
  - Instead of min. pair wise distance in PS, min. *total sample* distance of PS

# Results From CC/4 Year Study

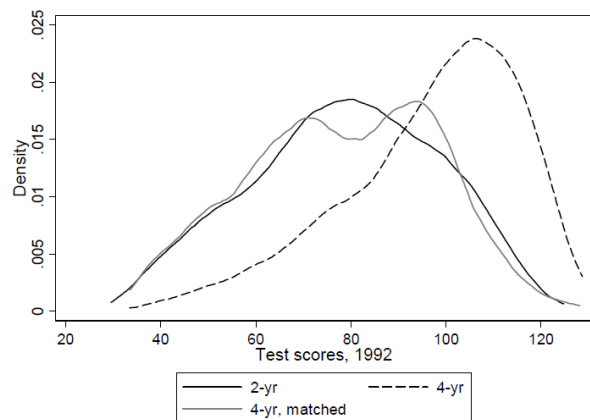Figure 4: Distributions of Standardized Test Scores by Treatment Status

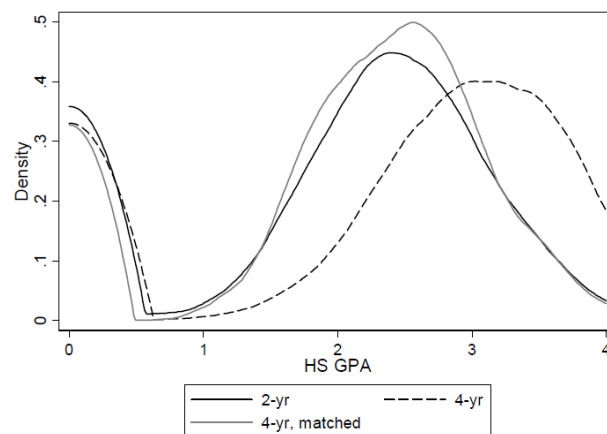Figure 5: Distributions of High School Grade Point Average by Treatment Status

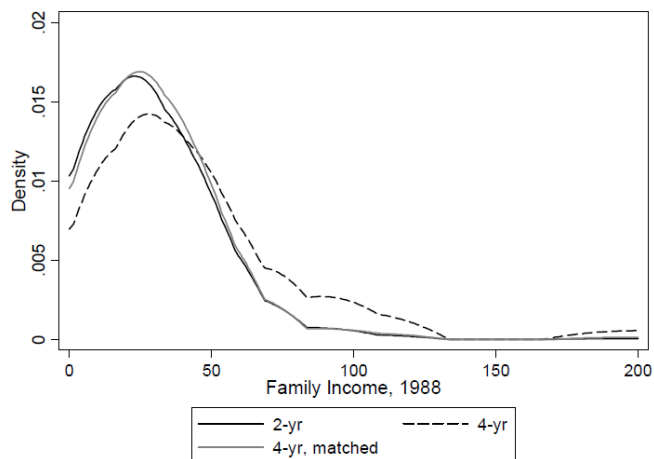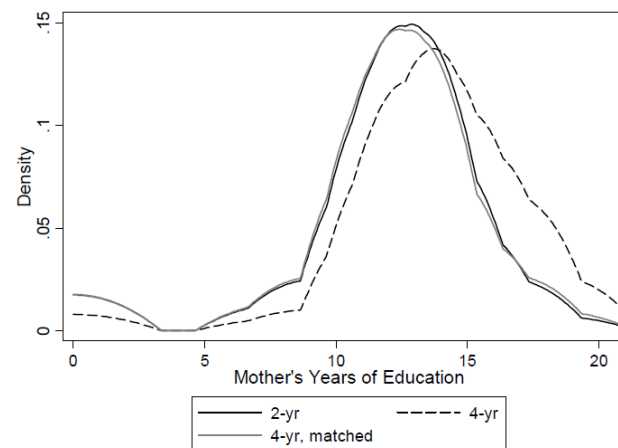Figure 6: Distributions of Family Income in 1988 by Treatment Status

Figure 7: Distributions of Mother's Years of Schooling by Treatment Status

# Results From CC/4 Year Study

Table 6: ATT for attending a two-year instead of a four-year institution

|  | OLS |  | NN |  | Kernel |  | LLR |  | RW |  |
|---|---|---|---|---|---|---|---|---|---|---|
| **Men** |  |  |  |  |  |  |  |  |  |  |
| Second year | -0.174 | * | → -0.146 | * | -0.163 | * | -0.152 | * | -0.093 | * |
|  | (0.019) |  | (0.029) |  | (0.025) |  | (0.026) |  | (0.047) |  |
| Third year | -0.246 | * | → -0.200 | * | -0.219 | * | -0.199 | * | -0.111 | * |
|  | (0.020) |  | (0.029) |  | (0.024) |  | (0.024) |  | (0.039) |  |
| AA | 0.182 | * | → 0.128 | * | 0.118 | * | 0.111 | * | 0.106 | * |
|  | (0.017) |  | (0.022) |  | (0.020) |  | (0.022) |  | (0.027) |  |
| BA | -0.250 | * | → -0.189 | * | -0.219 | * | -0.203 | * | -0.135 | * |
|  | (0.019) |  | (0.027) |  | (0.022) |  | (0.022) |  | (0.032) |  |
| Total credits | -23.320 | * | → -20.563 | * | -22.940 | * | -21.077 | * | -11.658 | * |
|  | (2.159) |  | (3.380) |  | (2.856) |  | (2.891) |  | (5.083) |  |
|  |  |  |  |  |  |  |  |  |  |  |
| **Women** |  |  |  |  |  |  |  |  |  |  |
| Second year | -0.168 | * | -0.156 | * | -0.153 | * | -0.126 | * | -0.088 | * |
|  | (0.017) |  | (0.031) |  | (0.024) |  | (0.026) |  | (0.039) |  |
| Third year | -0.263 | * | -0.227 | * | -0.239 | * | -0.209 | * | -0.16 | * |
|  | (0.018) |  | (0.028) |  | (0.023) |  | (0.024) |  | (0.037) |  |
| AA | 0.228 | * | 0.151 | * | 0.172 | * | 0.171 | * | 0.157 | * |
|  | (0.016) |  | (0.024) |  | (0.019) |  | (0.024) |  | (0.028) |  |
| BA | -0.284 | * | -0.227 | * | -0.232 | * | -0.209 | * | -0.166 | * |
|  | (0.018) |  | (0.024) |  | (0.020) |  | (0.021) |  | (0.030) |  |
| Total credits | -24.277 | * | -21.620 | * | -22.360 | * | -18.972 | * | -13.988 | * |
|  | (2.055) |  | (3.279) |  | (2.623) |  | (2.773) |  | (3.877) |  |

Standard errors are jointly bootstrapped using 1000 replications. Asterisks denote statistical significance at the 10% level or better.