The effect of college expansion on college attainment: evidence from historical US censuses

Collin J Wardius

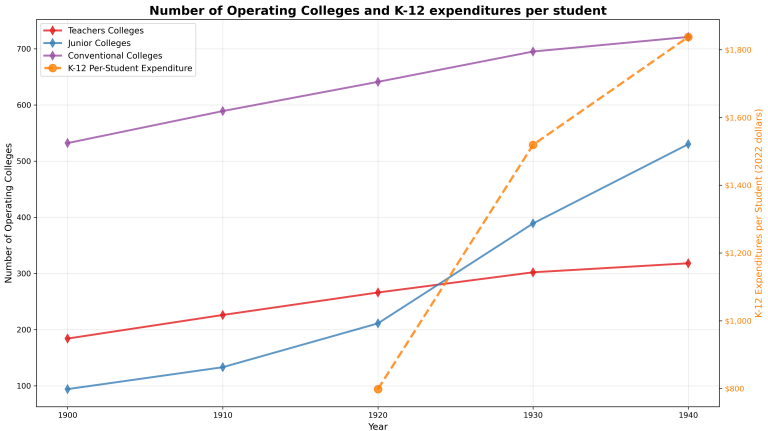
Department of Economics, UC San Diego
Approved by

October 30, 2025

Education in the US experienced a major transformation in the early 1900s

- Many more students completed HS and college
- Massive increase in capacity and spending at all levels of education
- **My focus: dramatic expansion in college openings and enrollment**

The great expansion of educational resources



Research Question

- Do supply-side expansions of colleges drive increased educational attainment?
- Specifically: How did new college openings affect local college attainment?

Preview of identification approach

- **Identifying variation**: quasi-random founding date of a university
- Some people are lucky as they are born just late enough to access a new university
- Some people are unlucky as they are born too early to access a new university

Literature

- **History of US higher education (1900-1940)**
  - → *My contribution:* Quantify the causal effect of university expansion on education access
  - Goldin (1998), Goldin and Katz (1998), Goldin (2001)
- **Effects of school building in non-US countries**
  - → *My contribution:* US university foundings and variation in public vs private control
  - Duflo (2001), Nimier-David (2023)
- **How proximity to college affects attainment and earnings**
  - → *My contribution:* Examine extensive margin of college access via new university foundings
  - Card (1993), Acton et al. (2025)
- **Historical census analysis to answer current questions in economics**
  - → *My contribution:* Create a dataset of university expansions and link them to the census data
  - Abramitzky, Boustan, and Eriksson (2014), Derenoncourt (2022), Bleemer and Quincy (2025)

Data

- **1900-1940 Decennial, Linked Full-Count US Censuses** Ruggles et al. (2025): occupation, income, education, pre- and post-college-age location
- **1947 College Blue Book** H.W. Hurt, H.J. Hurt, and Burckel (1947): college founding year, enrollment, student capacity, state or private control
- **Biennial Surveys of Education and Commissioner's on US Education**: college-level data on enrollment, finances, faculty, and programs (novel data in the process of being digitized by me)

BA Completion: 1900 vs 1936 Birth Cohorts

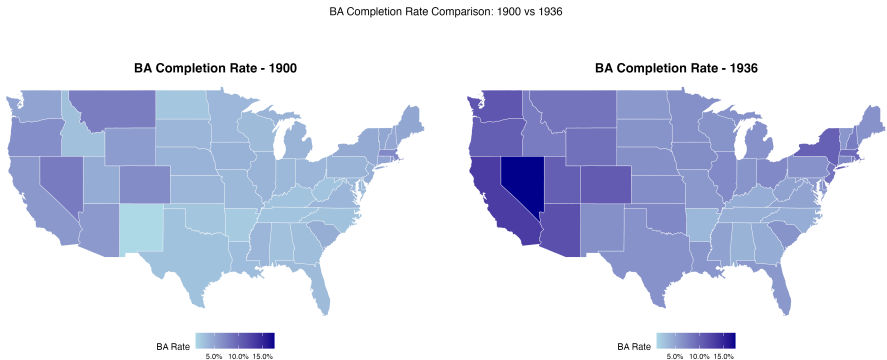


Figure: BA Completion: 1900 vs 1936 Birth Cohorts

College Founding Years by Region

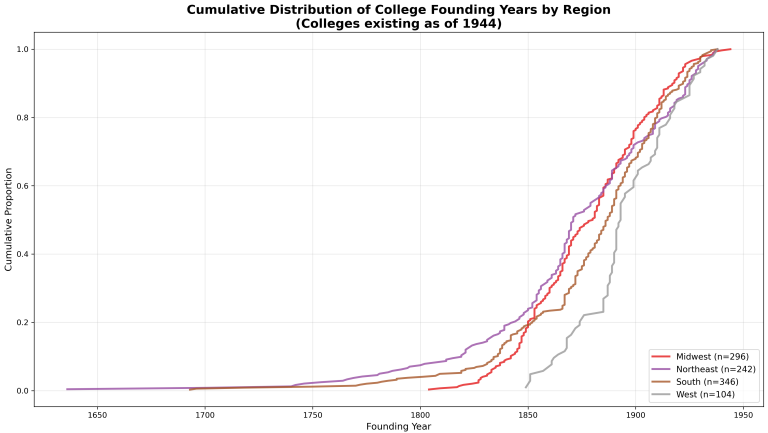


Figure: Regional Distribution of College Founding Years

College Founding Years by Control

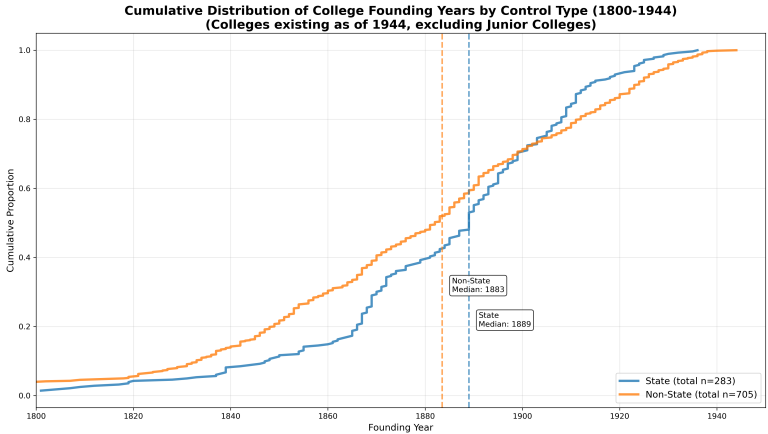


Figure: Regional Distribution of College Founding Years (1800+)

Estimating the effect of a university founding on college attainment: cohort DD approach

Cross sectional regression, identifying variation is at the age cohort-by-county level.

$$y_{ick} = \alpha_c + \lambda_k + \beta \text{New college}_{ck} \times \lambda_k + \xi \boldsymbol{X}_{ick} + \epsilon_{ick} \quad (1)$$

- c: county, k: age cohort, i: person
- $\text{New college}_{ck} = \mathbb{1}\{\text{There is a college founded in } c \text{ before } k \text{ graduates high school}\}$

The cohort DD identification assumption

- Compare the gap in attainment between older cohorts and younger cohorts in counties that have a new college versus those that do not

$$\tau_{DD} = \underbrace{[\mathbb{E}[y_{ick} \mid \text{New college}_{ck} = 1, k = \text{young}] - \mathbb{E}[y_{ick} \mid \text{New college}_{ck} = 1, k = \text{old}]]}_{\text{Cohort difference in treated counties}} - \underbrace{[\mathbb{E}[y_{ick} \mid \text{New college}_{ck} = 0, k = \text{young}] - \mathbb{E}[y_{ick} \mid \text{New college}_{ck} = 0, k = \text{old}]]}_{\text{Cohort difference in control counties}} \quad (2)$$

- **Identifying assumption**: Conditional on controls, counties that gained a college would have experienced parallel trends in attainment across cohorts absent the new college

Alternative approach using county-specific trends

$$y_{ick} = \alpha_c + \gamma_c \cdot k + \beta \text{New college}_{ck} + \xi \mathbf{X}_{ick} + \epsilon_{ick} \quad (3)$$

where:

- c: county, k: age cohort, i: person
- $\gamma_c \cdot k$: county-specific linear trends in cohort outcomes

Key feature: Identification uses only within-county variation, comparing young (exposed) vs. old (unexposed) cohorts in the same county

The county-specific trend identifying assumption

For each county c that gets a college, the treatment effect is deviations from the county-specific linear trend:

$$\tau_c = [\mathbb{E}[y_{ick} \mid c, k = \text{young}] - \mathbb{E}[y_{ick} \mid c, k = \text{old}]] - \gamma_c \cdot \Delta k \tag{4}$$

β averages these within-county effects across all treated counties.

Identifying assumption: Absent the college, outcomes for cohorts within county c would follow a linear trend.

Isolating treated and control counties

- Restrict attention to “conventional” colleges: exclude junior colleges, normal schools, teachers colleges, and colleges with capacity ≤ 100.
- Treated counties gain exactly one college over this period
- Three natural control groups:
  - Counties that never get a college (never-treated)
  - Counties that get a college later in the period (not yet treated)
  - Counties that received a college before 1900 and do not receive a college during this period (already treated)

County spatial stability over 1900-1940 County crosswalk construction

Quantifying the treated and control counties

Table: County Classification for College Analysis (1900-1940)

County Group	Count	Role in Analysis
Had college before 1900	278	—
Did not gain college 1900-1940	220	Potential Control
Gained college(s) 1900-1940	58	—
No college before 1900	2830	—
Gained exactly 1 college 1900-1940	55	**Treated**
Gained 2+ colleges 1900-1940	2	—
Never gained college by 1940	2773	Potential Control

Notes: Analysis excludes junior colleges, normal schools, teachers colleges, and colleges with capacity ≤ 100. Treated group consists of counties that had no college before 1900 and gained exactly one college 1900-1940. Potential control groups consist of (1) counties that had a college before 1900 but did not gain additional colleges 1900-1940, and (2) counties that never had a college by 1940.

The treated counties

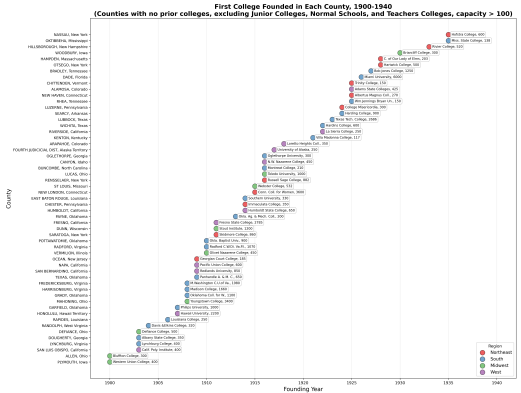


Figure: The treated counties

Determining which individuals experiences a college expansion

We only observe education and location in 1940, after individuals either received or did not receive a college education

1. Identify adults (age 25+) in the 1940 census
2. Link back to the censuses for which they are below the age of 18 using Ruggles et al. (2025) longitudinal linkage
3. If an individual is observed twice before 18, take the latest observation
4. Assign the individual that county of residence for the purposes of treatment assignment

Comparing linked versus unlinked individuals in the census

Table: Comparison of 1940 Characteristics: Linked vs Unlinked Individuals

	Linked Mean	Unlinked Mean	Difference
Female (%)	23.3	63.3	-40.0
Age	37.2	49.7	-12.5
College (%)	13.4	7.9	5.5
Married (%)	70.5	83.7	-13.1
White (%)	95.1	93.1	2.0
N	18,521,950	26,557,936	
% of Total	41.1%	58.9%	

Note: This table compares mean characteristics in 1940 for individuals age between 25 and 70 who were successfully linked to pre-age 18 observations versus those who were not linked.

Testing parallel trends: Event study specification

To test for pre-trends and trace out dynamic effects, estimate:

$$y_{ick} = \alpha_c + \lambda_k + \sum_{j \neq -1} \beta_j \mathbb{1}\{\text{Cohort } k \text{ born } j \text{ years relative to college founding in } c\} + \xi \boldsymbol{X}_{ick} + \epsilon_{ick} \tag{5}$$

- $j < 0$: Cohorts born *before* college founding (test for pre-trends)
- $j \geq 0$: Cohorts born *after* college founding (treatment effects)
- Omit $j = -1$ as reference category
- Null hypothesis: $\beta_j = 0$ for all $j < 0$ (no pre-trends)

County spatial stability over 1900-1940

Table: County Boundary Stability Between 1900 and 1940

	Reference Period	
Overlap Threshold	1940 Counties	1900 Counties
Total Counties	3108	2848
99% or more overlap	2852 (91.8%)	2538 (89.1%)
95% or more overlap	2941 (94.6%)	2616 (91.9%)
90% or more overlap	2976 (95.8%)	2647 (92.9%)
80% or more overlap	3005 (96.7%)	2681 (94.1%)

Notes: The 1940 Counties column shows the percentage of 1940 counties that overlap with a single 1900 county at the specified threshold. The 1900 Counties column shows the percentage of 1900 counties that overlap with a single 1940 county.

Back to Isolating treated and control counties

Creating a county crosswalk

We need consistent county boundaries to accurately assign which people experience a college creation versus which do not.

Approach:

1. Use 1940 as the reference year
2. Spatially intersect 1900, 1910, 1920, and 1930 boundaries with 1940 boundaries
3. Match counties where the intersection exceeds 70% overlap
4. Retain only counties that appear consistently across all census years

Back to Isolating treated and control counties

My reference on how the spatial join is performed

Consider A and B from 1900 (the base year) and 1940 (the target year) respectively. Then get the area of $A \cap B$. We then compare this to the area of the target to calculate

$$\frac{A \cap B}{B} \tag{6}$$

and we map B to A if this is above some threshold. I am using 70% as of right now but this could be modified. In practice, almost 90% of the counties have close to 100% overlap.