SciencesPo
DEPARTMENT OF ECONOMICS

ScPoEconometrics Advanced

Diff in Diff Applications

Nikiforos Zampetakis
SciencesPo Paris
2024-04-03

Diff in Diff

Last Time

- Binary Dependent Variables
- LPM
- probit-logit models

Today

- What economic research looks like today
- Applications of the Diff-in-Diff methods

Flash Recap 2x2 DiD

- 2 groups: Treated and Untreated
- 2 periods: Before and after

Possible comparisons

- Post vs Pre: Time trend problem
- Treatment vs Control: Selection bias problem

Combine the two to get the DiD

$$\delta_{2x2}^{DD} = (\bar{y}_T^{post} - \bar{y}_T^{pre}) - (\bar{y}_C^{post} - \bar{y}_C^{pre})$$

Crucial Assumption **Parallel Trends**

- Differences between treatment and control are time invariant (don't affect the slope of the time trend).
- Absent treatment the outcome between treatment and control groups would follow the same time trend.

Regression Version DiD

The typical regression model is

$$Y_{it} = \beta_1 + \beta_2 Treat_i + \beta_3 Post_t + \beta_4 (Treat_i \times Post_t)_{it} + \epsilon_{it}$$

- Treat: dummy for treatment group.
- Post: post-treatment dummy
- β_1: pre-treatment mean in control group.
- β_2: Treatment vs. Control comparison - captures selection bias (assumed to be time-invariant).
- β_3: Pre vs. Post comparison, capturing time trend.
- β_4: is the DD effect, identifying **ATT** (w. homogeneous TE)

TWFE - DiD with Panel Data

When we observe treated and untreated units for multiple time periods and units potentially experience treatment at different times the model becomes:

$$Y_{it} = \alpha_i + \alpha_t + \delta_{DD}(Treat_i \times Post_t)_{it} + \epsilon_{it}$$

- α_i: unit fixed effects
- α_t: time fixed effects

- Cluster standard errors at the unit level to allow for serial correlation n (Bertrand, Duflo and Mullainathan 2004).

Not sure exactly what δ_{DD} is and how it compares mean outcomes across units. New expanding literature has dived in to this problem.

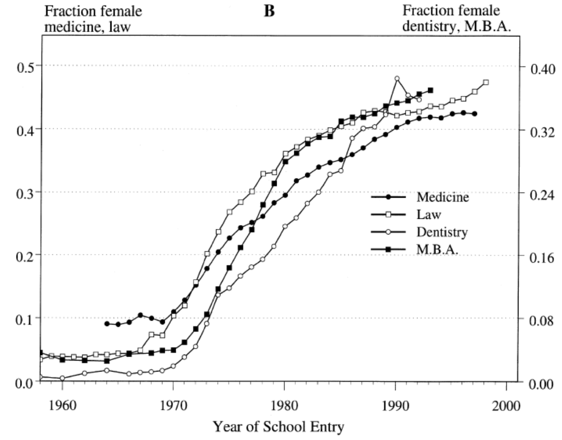
The Power of the Pill: Oral Contraceptives and Women's Career and Marriage Decisions, Goldin and Katz (2002)

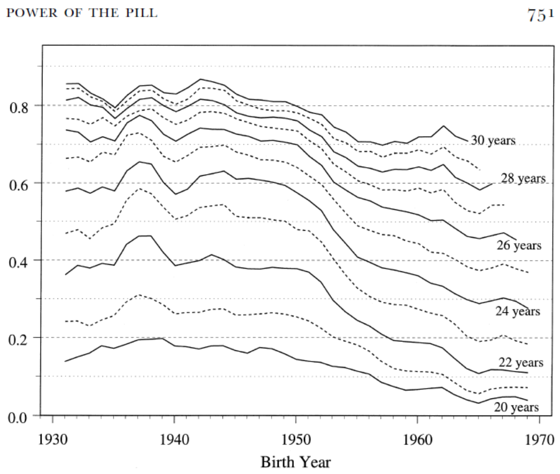
The Power of the Pill Goldin and Katz (2002)

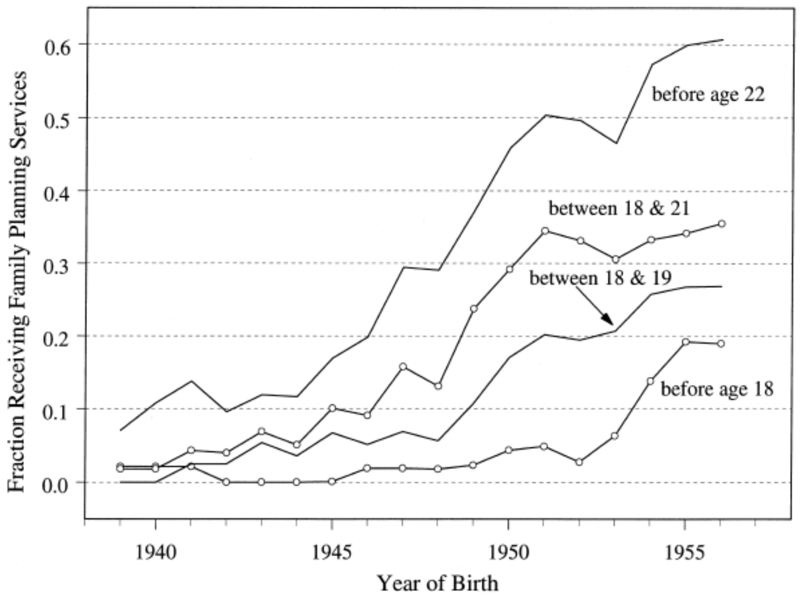
- Early 1970s in the US
- Fraction of women entering graduate professional programs increase
- Women's age at first marriage increases as well
- **Why?**

- Answer by Goldin and Katz:
- the diffusion of the **birth control pill** among the young women in late 1960s.
- They use TWFE to prove that.

What we are after







The Paper

- In the paper Goldin and Katz present different ways to support their argument.
- Descriptive evidence a simple theoretical model, and an econometric analysis based on a TWFE DiD.
- We will concentrate on the metrics part of their argument but the paper is of high standard in general.
- They concentrate on post-graduate education (e.g. Medicine, MBA) because it requires extensive up-front education.

Data

- Various subsets of the US Census data
- They use the US Census population census data from 1970, 1980, 1990.
- They observe measures of:

- Career and marital status outcomes
- Access to the pill
- Abortion when young
- for single year of birth cohorts of college women

The DiD Strategy for Age of the First Marriage

$$M_{isy} = \alpha_s + \delta_y + \beta X_{isy} + \gamma P_{sy} + \pi A_{sy} + \epsilon_{isy}$$

- i indexes individuals; s indexes state of birth; y indexes year of birth.
- M_{isy}: is a dummy variable equal to one if individual i was married before age 23;
- X_{isy}: contains controls for race.
- P_{sy}: is a dummy variable equal to one if i's state of birth had a non restrictive birth control law for minors at the time i was 18 years old.
- A_{sy}: is a dummy variable equal to one if abortion was legal in i's state of birth at the time i was 18 years old.
- α_s: state of birth dummies.
- δ_y: year of birth dummies.

Basic idea: exploit the substantial cross-state variation in the timing of the enactment of impact of state laws regarding birth control access for minors on the likelihood of getting married before age23, for college-educated women born in the United States from 1935 to 1957.

Regression Results

TABLE 4
STATE LAWS AND THE AGE AT FIRST MARRIAGE FOR COLLEGE WOMEN (U.S. Natives Born 1935–57)
Dependent Variable: 1 = Married before Age 23

	College Graduates							Some College or More	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mean of dependent variable	.41	.41	.41	.41	.41	.41	.41	.53	.53
Nonrestrictive birth control law at age 18*	−.0196	−.0162	−.0207	−.00986	−.0227			−.0124	
	(.00737)	(.00762)	(.00920)	(.00791)	(.00917)			(.00600)	
	[.0109]	[.0105]	[.00941]	[.0107]	[.00995]			[.0100]	
Pill access by age 17†						−.0262	−.0324		−.0240
						(.0115)	(.0131)		(.00872)
						[.0163]	[.0143]		[.00143]
Pill access by ages 18–20‡						−.00894	−.0126		−.0132
						(.00822)	(.00821)		(.00593)
						[.00922]	[.00920]		[.00676]
Legalized abortion at age 18§		−.0236	−.0114					−.00974	−.00904
		(.00992)	(.00956)					(.00777)	(.00761)
		[.0103]	[.0103]					[.00727]	[.00705]
Average abortion rate at ages 18–21∥				−.0653	.00523		.00280		
				(.0164)	(.0267)		(.0267)		
				[.0146]	[.0260]		[.0258]		
Race dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes
State of birth dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes
Year of birth dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes
State-specific linear trends	no	no	yes	no	yes	no	yes	no	no
Observations	60,714	60,714	60,714	60,714	60,714	60,714	60,714	130,335	130,335
R^2	.0458	.0458	.0469	.0459	.0469	.0459	.0469	.0434	.0434

SOURCE.—1980 *Census of Population*, IPUMS, 1 percent sample (Ruggles and Sobek 1997).

The DiD strategy for Career and Marital Status Outcomes

- Aggregate cohort analysis based on data from the 1970, 1980 and 1990 U.S. population censuses.
- Unit of observation in the analysis is an age/year cell (or a year of birth /calendar year cell). 20 age groups (ages30–49) across 3 census years 70, 80 and 90.

- Actual pill usage $\rightarrow$ proxied by the fraction of college women in a cohort taking the pill before age 21 among those with no births before age 23.
- Pill access $\rightarrow$ proxied by the fraction in a cohort born in a state with a non-restrictive birth control access law when they were younger than 21years

The DiD strategy for Career and Marital Status Outcomes

$$Y_{at} = \alpha_a + \delta_t + \beta X_{at} + \gamma P_{at} + \pi A_{at} + \epsilon_{at}$$

- Y_{at}: the share of age group a experiencing a particular career or marital status outcome in year t;
- X_{at}: contains controls for race.
- P_{at}: measure of access to or usage of birth control for cohort members as young women.
- A_{at}: a measure of access to or usage of abortion for cohort members as young women.

Basic idea: to observe successive cohorts at the same age to examine whether between-cohort changes in career and marital status outcomes are related to between-cohort changes in access to the pill and abortion for young single women, controlling for preexisting trends in these outcomes across cohorts.

Regression Results

TABLE 5
IMPACT OF PILL ACCESS AND ABORTION LEGALIZATION ON CAREER AND MARITAL STATUS FOR COLLEGE WOMEN, 30–49 YEARS OLD
(U.S. Natives Born 1921–60)

	Professional Occupation, Excluding Teachers, Nurses		Lawyer, doctor		Never Married	Currently Married	Currently Divorced	Currently Divorced/Ever Married	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mean of dependent variable	.127	.127	.0141	.0141	.126	.749	.0916	.104	.104
Fraction using pill before age 21*	.0480		.0352					−.0325	
	(.0275)		(.00539)					(.0257)	
Fraction with pill access law before age 21†		.00410		.0159	.0608	−.00813	−.0596		−.0558
		(.0142)		(.00353)	(.0189)	(.0230)	(.0107)		(.0121)
Average abortion rate from ages 18 to 21‡	.0457		.0306					−.149	
	(.0230)		(.00451)					(.0215)	
Fraction with legalized abortion at age 18§		.0236		.00255	.0431	−.0299	−.0127		−.0135
		(.0146)		(.00362)	(.0194)	(.0236)	(.0110)		(.0124)
Census year = 1980	.0153	.0202	.00136	.00361	−.00305	−.0521	.0561	.0622	.0627
	(.00531)	(.00493)	(.00104)	(.00123)	(.00655)	(.00799)	(.00371)	(.00498)	(.00420)
Census year = 1990	.0381	.0496	.00568	.0104	−.0118	−.0847	.105	.112	.116
	(.00904)	(.00773)	(.00177)	(.00192)	(.0103)	(.0125)	(.00582)	(.00848)	(.00660)
Age dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes
Observations	60	60	60	60	60	60	60	60	60
R^2	.864	.859	.962	.936	.924	.886	.968	.959	.965

SOURCE.—1970, 1980, and 1990 U.S. *Census of Population*, IPUMS, 1 percent samples (Ruggles and Sobek 1997).

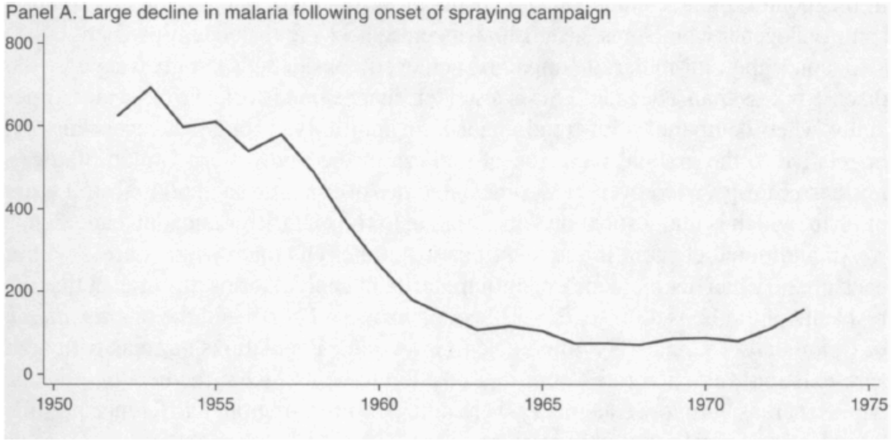
Malaria Eradication in the Americas: A Retrospective Analysis of Childhood Exposure, Bleakley (2010)

Bleakley (2010)

- Malaria persists in tropical regions up to the present day.
- Same areas lower level of development compared to countries in more tempered climates.
- Question naturally arising: Does Malaria hold back economic progress?

- Computing correlations not enough to answer the question.
- Malaria might depress productivity but the failure to eradicate Malaria might be a symptom of underdevelopment (reverse causality).
- Solution via metrics: exogenous variation to Malaria.
- Setting: two major attempts of Malaria eradication South US (1920) and Brazil, Colombia and Mexico (1950s).

What we are after



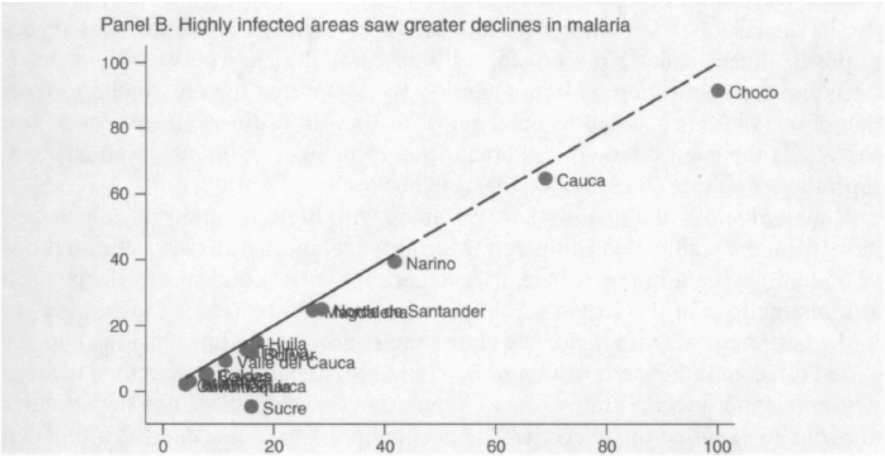


Figure 1. Malaria Incidence Before and After the Eradication Campaign, Colombia

The Paper

- Malaria has acute symptoms (fever, headache and nausea) and chronic symptoms anemia.
- This paper: long-term effects of childhood exposure to Malaria eradication campaigns on subsequent labor productivity as an adult.
- **Key factor** for the analysis: the eradication **campaigns were exogenous to growth potential** of the affected regions.
- **Data**: Integrated Public Use Micro Sample (IPUMS)
- Variation in the Malaria intensity between regions in a given state (southern Mexico - norther Mexico, US south - US north, south Brazil - north Brazil and tierra caliente Colombia).
- Areas with high infection rates benefited more from the eradication campaign. 50 to 80 percent decrease after the campaigns.
- Compare outcomes between different age groups at he time of the campaign (Adults - kids).

Dif-in-Diff Strategy

- Concentrate on males: consistent and higher labor-force participation.
- Labor productivity measured by: average by occupation of all reported labor earnings weighted average of earnings and education among males within each occupation
- Combine data to construct panels by cohort and aggregate to year of birth $\times$ census year $\times$ place of birth.

$$Y_{j,Post} - Y_{j,pre} = \beta M_{j,pre} + X_{j,pre}\Gamma + \alpha + \epsilon_{j,post}$$

- Exactly the same as a TWFE with two periods.

Results US

- Areas with higher malaria burdens prior to the eradication efforts saw larger cross-cohort growth rates in income.

TABLE 1—CROSS-COHORT DIFFERENCES AND MALARIA: UNITED STATES

	Malaria mortality (fraction of total), 1890		Malaria ecology (Hong)	
Dependent variable:				
Occupational Income Score	X		X	
Duncan's Socioeconomic Index		X		X
Panel A. Alternative control sets				
Additional controls:				
Basic specification only	0.112***	0.134**	0.236***	0.219***
	(0.039)	(0.065)	(0.032)	(0.053)
Health	0.100***	0.144**	0.225***	0.280***
	(0.038)	(0.067)	(0.031)	(0.048)
Education	0.136***	0.131**	0.219***	0.206***
	(0.041)	(0.062)	(0.027)	(0.055)
Other	0.094**	0.115*	0.204***	0.178***
	(0.044)	(0.063)	(0.029)	(0.068)
Full controls	0.110**	0.172*	0.215***	0.265***
	(0.049)	(0.094)	(0.049)	(0.096)
Panel B. Estimates using two-stage least squares				
Instrumental variables:				
The other malaria proxy	0.142***	0.175**	0.207***	0.244**
	(0.054)	(0.088)	(0.060)	(0.106)
Average temperature and altitude	0.154*	0.209**	0.138**	0.174**
	(0.083)	(0.104)	(0.059)	(0.075)
All of the above instruments	0.149***	0.192**	0.164***	0.185***
	(0.054)	(0.095)	(0.052)	(0.071)

Results Mexico and Brazil

TABLE 3—CROSS-COHORT DIFFERENCES AND MALARIA: COLOMBIA

Dependent variables: Differences across cohorts in…	Malaria ecology (Poveda) Literacy	Malaria ecology (Poveda) Years of schooling	Malaria ecology (Poveda) Income index	Malaria ecology (Mellinger) Literacy	Malaria ecology (Mellinger) Years of schooling	Malaria ecology (Mellinger) Income index
Panel A. Alternative controls						
Additional controls:						
None (basic specification)	0.035***	0.168*	0.065***	0.071***	0.064	0.048***
	(0.013)	(0.088)	(0.011)	(0.016)	(0.108)	(0.014)
Conflict	0.032***	0.175*	0.063***	0.068***	0.068	0.046***
	(0.012)	(0.090)	(0.011)	(0.016)	(0.110)	(0.014)
Economic activity	0.008	0.194**	0.057***	0.043***	0.156	0.039***
	(0.010)	(0.089)	(0.012)	(0.013)	(0.110)	(0.014)
Other diseases	0.024*	0.180**	0.065***	0.058***	0.057	0.042***
	(0.013)	(0.089)	(0.012)	(0.016)	(0.114)	(0.015)
Full controls	0.006	0.165*	0.064***	0.046***	0.076	0.034**
	(0.011)	(0.095)	(0.013)	(0.015)	(0.117)	(0.015)
Panel B. Alternative instrument sets						
Instrumental variables:						
Temperature, altitude, and their interaction	0.037**	0.372***	0.092***	0.067*	0.766***	0.170***
	(0.018)	(0.136)	(0.017)	(0.036)	(0.268)	(0.037)
The other two malaria proxies	0.126***	0.113	0.084***	0.082***	0.390*	0.149***
	(0.032)	(0.190)	(0.026)	(0.029)	(0.203)	(0.028)
Holdridge climate zone	0.045**	0.303*	0.102***	0.082**	0.593**	0.124***
	(0.021)	(0.159)	(0.020)	(0.037)	(0.248)	(0.035)
All of the above instruments	0.049***	0.323***	0.092***	0.074***	0.516***	0.120***
	(0.017)	(0.122)	(0.016)	(0.026)	(0.184)	(0.025)

TABLE 2—CROSS-COHORT DIFFERENCES AND MALARIA: BRAZIL AND MEXICO

Dependent variables: Differences across cohorts in…	Brazilian states ($N = 24$) Literacy	Brazilian states Education	Brazilian states Log total income	Brazilian states Log earned income	Mexican states ($N = 32$) Literacy	Mexican states Education	Mexican states Log earned income
Panel A. Estimates using ordinary least squares							
Specification:							
Basic	0.063	0.555	0.351**	0.267**	0.116***	0.058	0.292***
	(0.063)	(0.607)	(0.173)	(0.131)	(0.032)	(0.298)	(0.112)
Include infant mortality	0.063	0.576	0.366**	0.262*	0.119***	0.138	0.286**
	(0.063)	(0.581)	(0.147)	(0.136)	(0.032)	(0.237)	(0.112)
Include sectorial shares	0.131***	1.288**	0.434**	0.283***	0.032	−0.234	0.196
	(0.042)	(0.597)	(0.183)	(0.094)	(0.039)	(0.247)	(0.135)
Full controls	0.147***	0.995**	0.393**	0.283*	0.035	−0.247	0.254*
	(0.042)	(0.487)	(0.178)	(0.147)	(0.035)	(0.260)	(0.148)
Panel B. Estimates using two-stage least squares (temperature and altitude instruments)							
Specification:							
Basic	0.225	−1.356	0.649*	0.434	0.128**	0.112	0.494**
	(0.215)	(2.162)	(0.335)	(0.335)	(0.058)	(0.648)	(0.196)
Full controls	0.215*	0.257	0.785*	0.497	0.048	−0.234	0.398**
	(0.120)	(0.979)	(0.414)	(0.330)	(0.042)	(0.510)	(0.176)

Can Openness Mitigate the Effects of Weather Shocks? Evidence from India's Famine Era, Burgess and Donaldson (2010)

Burgess and Donaldson (2010)

- Rural citizens in **developing countries** today are highly exposures to **weather fluctuations**.
- This exposure can lead to **famines**
  - Acutely low nominal agricultural income
  - Acutely high food prices.
- Debate whether **trade openness** mitigates or exacerbates the weather socks.
  - Since Adam Smith openness leads to less volatility in real incomes.
  - Mahatma Gandhi and others support that openness increases the frequency of famines.

- **Fundamental ambiguity**: income more volatile - consumer prices less volatile
  - What is the net effect on income?
- In the paper concentrate on colonial era (1875 - 1919) India.
- Worst string of famines in history
- Also the period where the bulk of the railroad network was built.
- **Research Question**: What is the effect of the railroad expansion on the mitigation of the weather shocks on famines?

The Paper

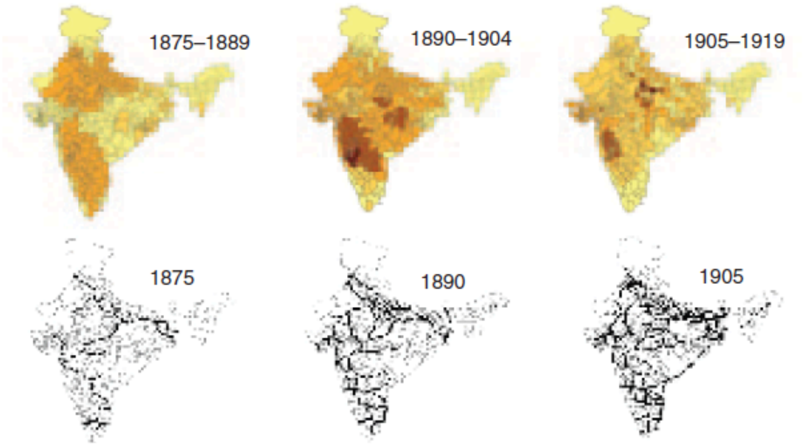


Figure 1. Famines and Railroads in India, 1875–1919

- $\frac{2}{3}$s of the population farmers - almost no irrigation system at the period ("gambling on monsoons).
- They Gather data on famines at the district level between 1875 - 1919 - make an order variable (0-3).
- Rainfall data and railroad data at the district level.

The Diff-in-Diff Strategy

$$F^*_{dt} = \alpha_d + \beta_t + \gamma_1 RAIN_{dt} + \gamma_2 RAIL_{dt} + \gamma_3 RAIN_{dt} \times RAIL_{dt} + \epsilon_{dt}$$

- α_d: district fixed effect
- β_t: year fixed effect
- $RAIN_{dt}$: amount of local rainfall enjoyed by district d in year t
- $RAIL_{dt}$: dummy variable indicating whether district d has a railroad line in the district in year t or not
- F^*_{dt}: famine intensity. Not observed use the ordered qualitative index instead.

- Estimate through fixed effect ordered logit model.

- did railroads mitigate ($\gamma_3>0$) or exacerbate ($\gamma_3<0$) the ill effects of a given rainfall shortage on famine intensity?

Results

TABLE 1—FAMINES, RAINFALL, AND RAILROADS

Dependent variable: famine severity index	(1)	(2)	(3)	(4)	(5)
Railroad in district [0, 1]	0.194	–1.625***	0.309	–2.178***	–2.136***
	(0.374)	(0.572)	(0.390)	(0.690)	(0.754)
Rainfall in district, year t [meters]	–0.855***	–2.218***	–0.860***	–2.316***	–17.35
	(0.208)	(0.532)	(0.204)	(0.518)	(20.40)
(Railroad in district) × (rainfall in district, year t)		1.858***		1.848***	1.729***
		(0.541)		(0.521)	(0.565)
Rainfall in district, year $t-1$ [meters]			–0.699***	–1.171***	9.316
			(0.215)	(0.395)	(21.51)
(Railroad in district) × (rainfall in district, year $t-1$)				0.692*	0.758*
				(0.404)	(0.458)
(Rainfall in district) × (trend)	No	No	No	No	Yes
Observations	3,809	3,809	3,551	3,551	3,551
Pseudo R^2	0.248	0.260	0.255	0.271	0.271

Backlash: The Unintended Effects of Language Prohibition in U.S. Schools after World War I, Fouka (2020)

Fouka (2020)

- Multiple countries strive to **integrate ethnic minorities**.
- Theories of nation building: imposing national language or **repressing minority culture increases homogeneity**.
- Integration policies can lead to a backlash $\rightarrow$ **strengthen the ethnic identity**.
- No empirical evidence whether this is true or not.
- In this paper concentrate on the **prohibition of German in US schools after WWI**.
- How did affect the **first name choice** and the **intermarriage decisions** among the German community in the US?
- How did affect their **patriotism** measured by their voluntary enlistment in the US Army during WWII?

The Paper

Focus on Indiana and Ohio $\rightarrow$ legislation against the German language.

- Control group all the bordering states (Michigan, Kentucky, Illinois, W.Virginia and Pennsylvania)
    - No banning or English as the mandatory language in public schools.

Create index of German name distinctiveness (German Name Index) (0 - 100)

- How much more frequent a name is among the population of German origin compared with the rest of the population.

Intermarriage difficult to measure from the censuses.

- In the 40s census only 5% asked their parental birthplaces.
- In the 30s treatment group to young to be married (14 - 27)/
- Still unbiased estimates but not representative to the whole population.

WWII enlistment

- Match the enlistment data to the census data for volunteers between 1940 - 1942

The Diff-in-Diff Strategy

Main strategy is based on comparing cohorts of school age with older cohorts between states with and without a language Law.

$$Y_{isc} = \alpha + \beta T_{cs} + \lambda_c + \theta_s + \delta \mathbf{Z}_{isc} + \epsilon_{isc}$$

- T_{cs} is an indicator for individuals living in a state with a law and who were within the age range for compulsory schooling (CSL) at the time that law was in place.
- λ_c and θ_s: cohort and state fixed effects
- $\mathbf{Z}_{isc}$: vector of name properties that affect the probability of a record being matched in a later census

Results

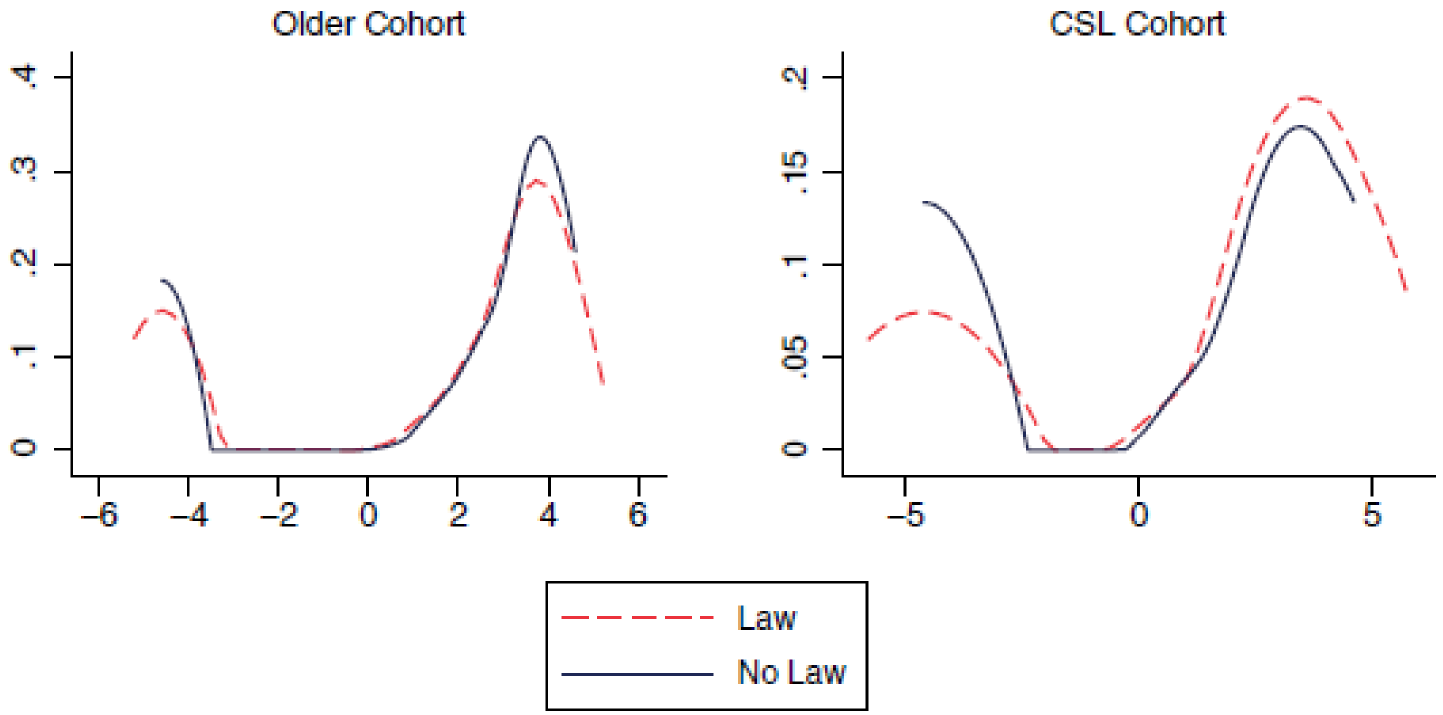


FIGURE 3

Densities of log GNI of first son by cohort

Notes: The figure illustrates, for the linked border data set, the kernel density of the logarithm of the GNI of the first son. The panel on the left plots this density for the cohort too old to have been in school (by compulsory law) at the time German was banned; the right panel plots the density for the treated cohort.

Results

TABLE 5
Baseline results: border data set

	[1]	[2]	[3]	[4]
Panel A: Dep. variable is Log average GNI of children				
Law × CSL age	0.473	0.884**	0.920**	0.794**
	(0.279)	(0.309)	(0.308)	(0.334)
	[0.1191]	[0.0801]	[0.0631]	[0.1291]
Observations	26334	26334	26334	26334
R-squared	0.0494	0.0515	0.0547	0.0553
Panel B: Dep. variable is Log GNI of first son				
Law × CSL age	1.231***	1.615***	1.638***	1.310***
	(0.280)	(0.300)	(0.299)	(0.241)
	[0.00200]	[0.0230]	[0.0230]	[0.0240]
Observations	18459	18459	18459	18459
R-squared	0.0162	0.0194	0.0290	0.0295
Panel C: Dep. variable is Spouse German				
Law × CSL age	0.0472**	0.0573***	0.0365**	0.0389***
	(0.0164)	(0.0183)	(0.0148)	(0.0105)
	[0.0781]	[0.0691]	[0.0941]	[0.0290]
Observations	24925	24921	24921	24921
R-squared	0.0510	0.0523	0.0634	0.0636
Additional controls	N	Y	Y	Y
County FE	N	N	Y	Y
State trends	N	N	N	Y

Results

TABLE 6

Baseline results: WWII enlistments

	[1]	[2]	[3]	[4]	[5]	[6]
Dep. variable:	Volunteer					
Law × CSL age	−0.0256**	−0.0235**	−0.0243**	−0.0370***	0.00595	−0.00389
	(0.00696)	(0.00928)	(0.00986)	(0.00785)	(0.0414)	(0.00784)
	[0.004]	[0.004]	[0.004]	[0.460]	[0.903]	[0.768]
Law × CSL age × German parents						−0.0313*
						(0.0149)
						[0.116]
Observations	2679	2667	2667	2667	5443	160246
R-squared	0.0261	0.0641	0.0643	0.0698	0.0777	0.0746
Additional controls	N	Y	Y	Y	Y	Y
Share German in state in 1910 × Cohort FE	N	N	Y	Y	N	N
State trends	N	N	N	Y	N	N

SciencesPo
DEPARTMENT OF ECONOMICS

END

	nikiforos.zampetakis@sciencespo.fr
	Slides
	Book
	@ScPoEcon
	@ScPoEcon