

Causal Inference II

MIXTAPE SESSION



Roadmap

Introduction

Managing expectations

Introducing difference-in-differences

Potential outcomes

Estimation

Inference

Parallel trends violations

How parallel trends can get violated

DiD in Court

Falsifications

Interesting Applications using staggered adoption

Facebook and Mental Health

Generative AI Randomized Experiment

ChatBot Rollout

Conclusion

Introduction

- Welcome to Mixtape Sessions workshop on difference-in-differences and synthetic control (“Causal Inference II”)
- 8:00am to 5:00pm CST, 15 min breaks every hour, 1 hour lunch at noon CST
- Lecture, discussion, exercises, application

What my pedagogy is like

- Long days that don't feel long because it's high energy, with regular breaks including lunch
- Move between the econometrics, history of thought, videos, applications, code, spreadsheets, exercises
- Ask questions at any point; I'll do my best to answer them

Class goals

Pedagogical goal is to break down the procedures into plain English, rebuilding it into something you can and want to use, but also:

1. **Confidence:** You will feel like you have a good enough understanding of diff-in-diff and synthetic control, both in its basics and some more contemporary issues, so that by the end of the week it a very intuitive, friendly, and useful tool
2. **Comprehension:** You will have learned a lot both conceptually and in the specifics, particularly with regards to issues around identification and estimation in the diff-in-diff and synth context
3. **Competency:** You will have more knowledge of programming syntax in Stata and R so that later you can apply this in your own work

Day 1 outline

Introduction to DiD basics

- Potential outcomes review and the ATT parameter
- DiD equation (“four averages and three differences”), parallel trends and estimation with OLS
- Evaluating parallel trends with falsifications, event studies
- Compositional changes, triple differences and covariates

Day 2 outline

Differential timing

- TWFE Pathologies in static and dynamic specifications ("event study")
- Solution 1: Aggregating group-time ATTs
- Solution 2: Imputation estimators
- Solution 3: Stacked regression

Day 3 outline

- Canonical synth (Abadie papers)
- Augmented synth (Ben-Michael, et al)
- Matrix completion with nuclear norm regularization (Athey, et al.)
- Exercises and estimation

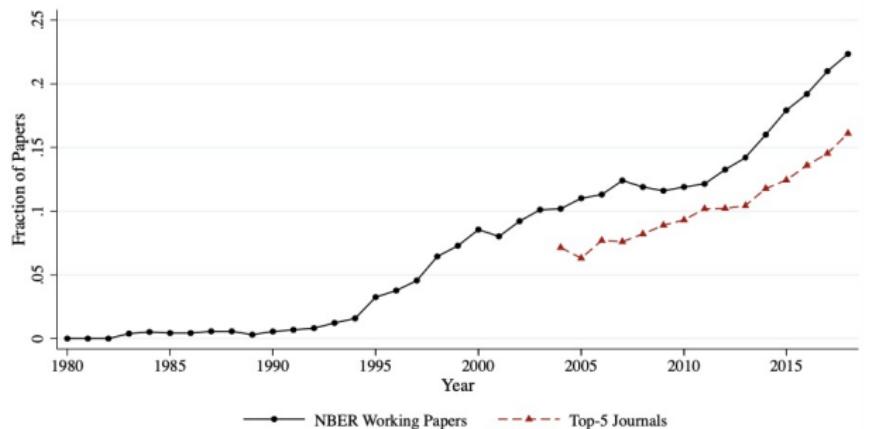
What is difference-in-differences (DiD)

- DiD is a very old, relatively straightforward, intuitive research design
- A group of units are assigned some treatment and then compared to a group of units that weren't
- One of the most widely used quasi-experimental methods in economics and increasingly in industry
- Mostly associated with “big shocks” happening in space over time

*“A good way to do econometrics is to look for good natural experiments and use statistical methods that can tidy up the confounding factors that nature has not controlled for us.” – Daniel McFadden
(Nobel Laureate recipient with Heckman 2002)*

Figure: Currie, et al. (2020)

A: Difference-in-Differences



Origins of diff-in-diff

- Difference-in-differences (DiD) was quietly and largely unnoticed introduced in the 19th century as a way to convince skeptics in health policy arguments
- Dominant disease theory in 19th century was *miasma* – disease caused by smelly vapor
- Keep in mind – microorganisms would not be identified until much later, partly caused by poor resolution in microscopes (Freedman 2007)

Miasma I: Ignaz Semmelweis and washing hands

- 1840s, Vienna maternity wards had high postpartum infections in one wing compared to other wings
- One division had doctors and trainee doctors, but another had midwives and trainee midwives

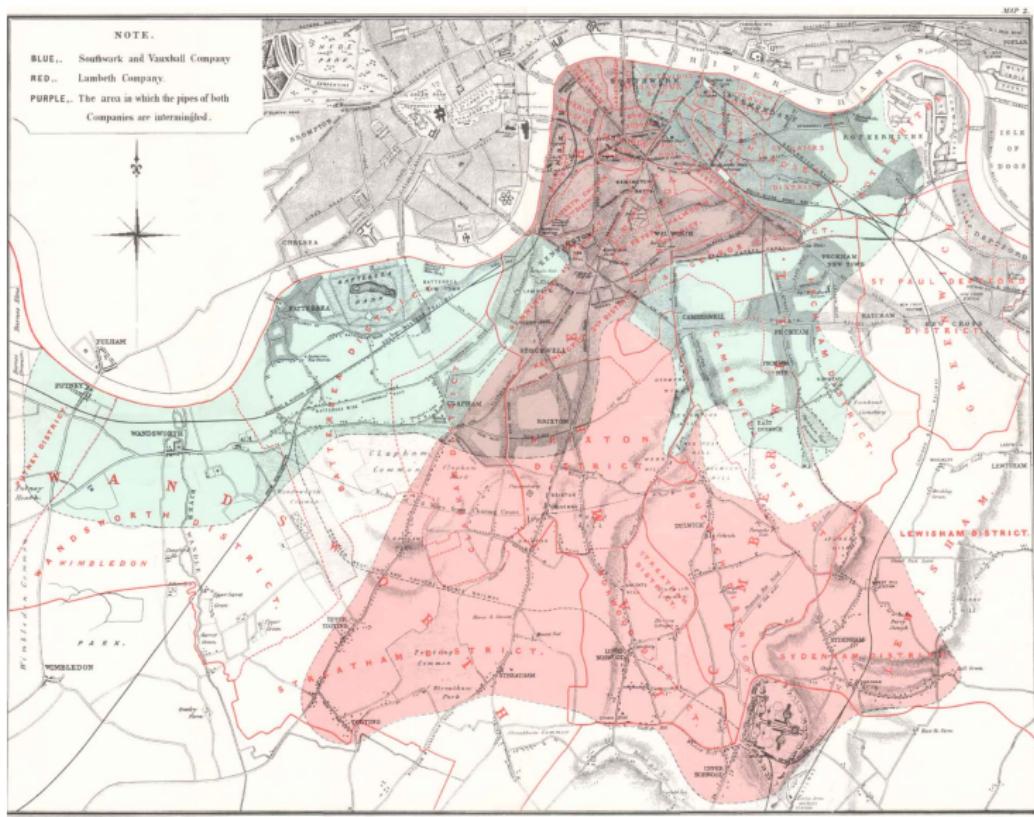
Miasma I: Ignaz Semmelweis and washing hands

- Ignaz Semmelweis notes the difference in 1841 when hospitals moved to “anatomical” training involving cadavers (Pamela Jakeila lecture notes on DiD)
- New training happens to one but not the other and Semmelweis thinks the mortality is caused by working with cadavers
- Proposes washing hands with chlorine in 1847 in the midwives’ wing and uses a DiD design of pre and post

Miasma II: John Snow and cholera

- Three major waves of cholera in the early to mid 1800s in London
- John Snow believed cholera was spread through the Thames water supply which contradicted dominant theory about “dirty air” transmission
- Grand experiment: Lambeth moves its pipe between 1849 and 1854; Southwark and Vauxhall delay
- He can evaluate the effect in three ways (one of which is DiD)

Figure: Two water utility companies in London 1854



1) Simple cross-sectional design

Table: Lambeth and Southwark and Vauxhall, 1854

Company	Cholera mortality
Lambeth	$Y = L + D$
Southwark and Vauxhall	$Y = SV$

$$\hat{\delta}_{cs} = D + (L - SV)$$

What is L and SV ?

1) Simple cross-sectional design

Table: Lambeth and Southwark and Vauxhall, 1854

Company	Cholera mortality
Lambeth	$Y = L + D$
Southwark and Vauxhall	$Y = SV$

$$\widehat{\delta}_{cs} = D + (L - SV)$$

This is biased if $L \neq SV$ (selection bias). Give an example when we're pretty sure they are equal.

2) Interrupted time series design

Table: Lambeth, 1849 and 1854

Company	Time	Cholera mortality
Lambeth	1849	$Y = L$
	1854	$Y = L + (L_t + D)$

$$\hat{\delta}_{its} = D + L_t$$

What is required for this estimator to be unbiased?

3) Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

Companies	Time	Outcome	D_1	D_2
Lambeth	Before	$Y = L$	$L_t + D$	$D + (L_t - SV_t)$
	After	$Y = L + L_t + D$		
Southwark and Vauxhall	Before	$Y = SV$	SV_t	
	After	$Y = SV + SV_t$		

$$\hat{\delta}_{did} = D + (L_t - SV_t)$$

This method yields an unbiased estimate of D if $L_t = SV_t$

Difference-in-differences and empirical crisis in labor economics

- Empirical crisis in empirical labor back in the 1970s (26:31 to 32:00)
https://youtu.be/1soLdywFb_Q?t=1579
- Orley Ashenfelter graduated from Princeton in the 1970s, takes a job in Washington DC and begins studying “job trainings programs” where he develops the difference-in-differences design
- Listen to Orley explain the connection he made between two way fixed effects and difference-in-differences; it was born out of a need to explain OLS to an American bureaucrat
<https://youtu.be/WnB3EJ8K7lg?t=126>

Steps of a project

1. Convert research question into causal parameter
2. Deduce beliefs needed to estimate that causal parameter with data
3. Create a calculator that will use data and estimate the causal parameter

Most of us skip (1) and maybe even (2) and instead simply “run regressions” and cross our fingers that that coefficient is causal, but is it? And why is it? And what is it? Let’s dig into Orley’s comment a little more.

Equivalence

$$Y_{ist} = \alpha_0 + \alpha_1 Treat_{is} + \alpha_2 Post_t + \delta(Treat_{is} \times Post_t) + \varepsilon_{ist}$$

$$\hat{\delta} = \left(\bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)} \right) - \left(\bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)} \right)$$

- Orley claims that the OLS estimator of δ and the “four averages and three subtractions” are the same thing numerically
- And they are – they are numerically *identical*
- And under a particular assumption, they are also unbiased estimates of an aggregate causal parameter
- But to see this we need new notation – potential outcomes

Potential outcomes notation

- Let the treatment be a binary variable:

$$D_{i,t} = \begin{cases} 1 & \text{if in job training program } t \\ 0 & \text{if not in job training program at time } t \end{cases}$$

where i indexes an individual observation, such as a person

Potential outcomes notation

- Potential outcomes:

$$Y_{i,t}^j = \begin{cases} 1: \text{wages at time } t \text{ if trained} \\ 0: \text{wages at time } t \text{ if not trained} \end{cases}$$

where j indexes a counterfactual state of the world

Treatment effect definitions

Individual treatment effect

The individual treatment effect, δ_i , equals $Y_i^1 - Y_i^0$

Missing data problem: I don't know my own counterfactual

Conditional Average Treatment Effects

Average Treatment Effect on the Treated (ATT)

The average treatment effect on the treatment group is equal to the average treatment effect conditional on being a treatment group member:

$$\begin{aligned} E[\delta | D = 1] &= E[Y^1 - Y^0 | D = 1] \\ &= E[Y^1 | D = 1] - \textcolor{red}{E[Y^0 | D = 1]} \end{aligned}$$

This is one of the most important policy parameters, if not the most important, and coincidentally it's also the parameter you get with diff-in-diff (even with heterogeneity)

Potential outcomes vs data

- ATT is expressed in terms of potential outcomes, but we do not use potential outcomes for estimation; we use data
- Potential outcomes are unknown and *hypothetical* possibilities describing states of the world but our data are realized outcomes, or "data", that actually occurred
- Potential outcomes become realized under treatment assignment

$$Y_{it} = D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0$$

- Depending on how the treatment is assigned really dictates whether correlations reveal causal effects or bias

DiD equation

Orley's "four averages and three subtractions", or what Bacon will call the 2x2

$$\hat{\delta} = \left(E[Y_k|Post] - E[Y_k|Pre] \right) - \left(E[Y_U|Post] - E[Y_U|Pre] \right)$$

k are the people in the job training program, U are the untreated people not in the program, $Post$ is after the trainees took the class, Pre is the period just before they took the class, and $E[y]$ is mean earnings.

Does $\hat{\delta}$ equal the ATT? If so when? If not why not?

Potential outcomes and the switching equation

$$\hat{\delta} = \underbrace{\left(E[Y_k^1|Post] - E[Y_k^0|Pre] \right) - \left(E[Y_U^0|Post] - E[Y_U^0|Pre] \right)}_{\text{Switching equation}} + \underbrace{E[Y_k^0|Post] - E[Y_k^0|Post]}_{\text{Adding zero}}$$

Parallel trends bias

$$\hat{\delta} = \underbrace{E[Y_k^1|Post] - E[Y_k^0|Post]}_{\text{ATT}} + \underbrace{\left[E[Y_k^0|Post] - E[Y_k^0|Pre] \right] - \left[E[Y_U^0|Post] - E[Y_U^0|Pre] \right]}_{\text{Non-parallel trends bias in 2x2 case}}$$

Identification through parallel trends

Parallel trends

Assume two groups, treated and comparison group, then we define parallel trends as:

$$E(\Delta Y_k^0) = E(\Delta Y_U^0)$$

In words: “The evolution of earnings for our trainees *had they not trained* is the same as the evolution of mean earnings for non-trainees”.

It's in red because parallel trends is untestable and critically important to estimation of the ATT using any method, OLS or “four averages and three subtractions”

Work together #1

Work together:

$$\hat{\delta} = \left(E[Y_k|Post] - E[Y_k|Pre] \right) - \left(E[Y_U|Post] - E[Y_U|Pre] \right)$$

Assume that the U group is treated in both periods. Replace expectations with potential outcomes and rewrite using the “add zero” trick we did. How is this similar to what we did before? Is parallel trends enough?

Work together #2

Work together:

$$\widehat{\delta} = \left(E[Y_k|Post] - E[Y_k|Pre] \right) - \left(E[Y_U|Post] - E[Y_U|Pre] \right)$$

In the pre-period, how important is it that the U group and k group have the same treatment status?

What is parallel trends

- Parallel trends assumes away the selection bias associated with comparisons
- The assumption is thought to be more plausible than simply assuming simple comparisons held equal
$$E[Y^0|D = 0] = E[Y^0|D = 1]$$
- But it is still a strong assumption, and differs from the assumptions have in the RCT which though also untestable, is nearly guaranteed by randomization
- Most of the hard part of the work involves the old fashioned detective work and the work of making good arguments with good exhibits (tables and figures)

Understanding parallel trends through worksheets

Before we move into regression, let's go through a simple exercise to really pin down these core ideas with simple calculations

[https://docs.google.com/spreadsheets/d/
1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=
sharing](https://docs.google.com/spreadsheets/d/1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=sharing)

No Anticipation

- Additional assumption is “no anticipation” – poorly named as it doesn’t require literally no anticipation
- No anticipation means that the treatment effect happens only at the time that the treatment occurs or after, but not before
 - **Example 1:** Tomorrow I win the lottery, but don’t get paid yet. I decide to buy a new house today. That violates NA
 - **Example 2:** Next year, a state lets you drive without a driver license and you know it. But you can’t drive without a driver license today. This satisfies NA.
- We need NA because we are comparing to a baseline period and it needs to not be treated

SUTVA

- Stable Unit Treatment Value Assumption (Imbens and Rubin 2015) focuses on what happens when in our analysis we are combining units (versus defining treatment effects)
 1. **No Interference:** a treated unit cannot impact a control unit such that their potential outcomes change (unstable treatment value)
 2. **No hidden variation in treatment:** When units are indexed to receive a treatment, their dose is the same as someone else with that same index
 3. **Scale:** If scaling causes interference or changes inputs in production process, then #1 or #2 are violated
- Shifts from defining treatment effects to estimating them, which means being careful about who is the control group, how you define treatments and what questions can and cannot be answered with this method

OLS Specification

- Simple DiD equation will identify ATT under parallel trends
- But so will a particular OLS specification (two groups and no covariates)
- OLS was historically preferred because
 - OLS estimates the ATT under parallel trends
 - Easy to calculate the standard errors
 - Easy to include multiple periods
- People liked it also because of differential timing, continuous treatments and covariates, but those are more complex so we address them later

Minimum wages

- Card and Krueger (1994) have a famous study estimating causal effect (ATT) of minimum wages on employment
- Let's listen to Card tell its origin
https://youtu.be/1soLdywFb_Q?t=2680
- Exploited a policy change in New Jersey between February and November in mid-1990s where minimum wage was increased, but neighbor PA did not
- Using DiD, they do not find a negative effect of the minimum wage on employment which is part of its legacy today, but I mainly present it to illustrate the history and the design principles



Binyamin Appelbaum

@BCAppelbaum



Replies to @BCAppelbaum

The Nobel laureate James Buchanan wrote in the Wall Street Journal that Card and Krueger were undermining the credibility of economics as a discipline. He called them and their allies "a bevy of camp-following whores."

3:49 PM · Mar 18, 2019



179



Reply



Share

[Read 18 replies](#)

Reaction to the paper

- Was it or was it not controversial? Yes and no
- Orley's opinion about the paper's controversy at the time.
<https://youtu.be/M0tbuRX4eyQ?t=1882>
- What Card and Krueger experienced
https://youtu.be/1soLdywFb_Q?t=3076

Card on that study

"I've subsequently stayed away from the minimum wage literature for a number of reasons. First, it cost me a lot of friends. People that I had known for many years, for instance, some of the ones I met at my first job at the University of Chicago, became very angry or disappointed. They thought that in publishing our work we were being traitors to the cause of economics as a whole."

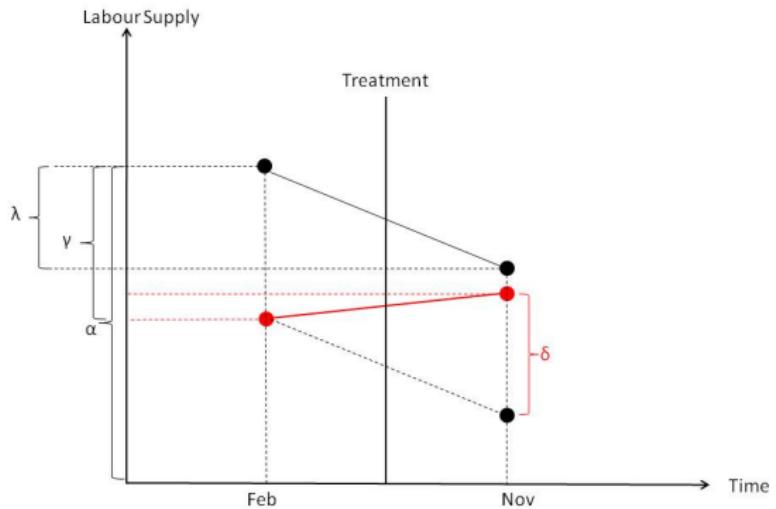
OLS specification of the DiD equation

- The correctly specified OLS regression is an interaction with time and group fixed effects:

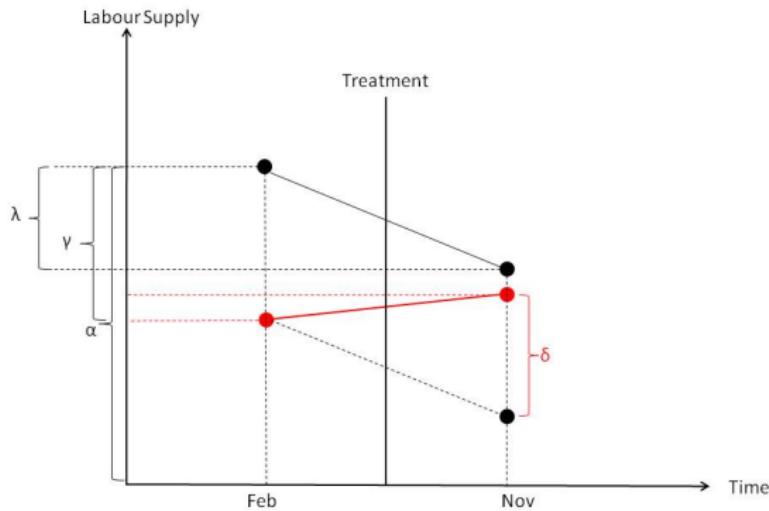
$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{its}$$

- NJ is a dummy equal to 1 if the observation is from NJ
- d is a dummy equal to 1 if the observation is from November (the post period)
- This equation takes the following values
 - PA Pre: α
 - PA Post: $\alpha + \lambda$
 - NJ Pre: $\alpha + \gamma$
 - NJ Post: $\alpha + \gamma + \lambda + \delta$
- DiD equation: $(NJ \text{ Post} - NJ \text{ Pre}) - (PA \text{ Post} - PA \text{ Pre}) = \delta$

$$Y_{ist} = \alpha + \gamma N J_s + \lambda d_t + \delta (N J \times d)_{st} + \varepsilon_{ist}$$



$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



Notice how OLS is “imputing” $E[Y^0|D = 1, Post]$ for the treatment group in the post period? It is only “correct”, though, if parallel trends is a good approximation

Introduction to Inference in DID

When dealing with clustered data, a crucial concept is the difference between correlated observations and correlated errors. While they may seem similar, they are distinct, and it's essential to focus on the errors when clustering standard errors.

Correlated Observations

- Correlated observations occur when the observed variables themselves are correlated within a cluster.
- For instance, incomes within a specific region might be positively correlated.
- Correlated observations do not necessarily violate OLS assumptions.

Correlated Errors

- Correlated errors occur when the unobserved errors are correlated within a cluster.
- This violates the classical OLS assumption of independent errors, leading to inefficient and possibly biased standard errors.
- Clustering standard errors accounts for this within-cluster correlation.

Why Focus on Correlated Errors?

- OLS assumptions are about errors, not the observed variables.
- Failing to account for correlated errors can lead to misleading inference.
- Clustering standard errors at the appropriate level corrects for this, giving more reliable hypothesis tests and confidence intervals.

Serial correlation creates problems

- So errors within a cluster (e.g., same treatment group) might be correlated.
- Even if between clusters, they are assumed to be independent, the within correlation in errors biases standard errors downward
- Bertrand, Duflo and Mullainathan (2004) show that conventional standard errors will often severely underestimate the standard deviation of the estimators
- They proposed three solutions, but most only use one of them (clustering)

Inference

- 1 Block bootstrapping standard errors (if you analyze states the block should be the states and you would sample whole states with replacement for bootstrapping)
- 2 Clustering standard errors at the group level (in Stata one would simply add `, cluster(state)` to the regression equation if one analyzes state level variation)

Most people will simply cluster, but there are issues if you have too few clusters. They mention a third way but it's only a curiosity.

Computing Cluster-Robust Standard Errors

1. Run your regression to get u and \hat{Y} .
2. Calculate cluster-level residuals \hat{u}_c .
3. Calculate the "meat" (as in bread-meat-bread sandwich) M as $\sum_c X'_c \hat{u}_c \hat{u}'_c X_c$.
4. Your standard error covariance matrix is $\hat{V} = (X'X)^{-1} M (X'X)^{-1}$.

Mixtape reference (chapter 2):

https://mixtape.scunning.com/02-probability_and_regression#cluster-robust-standard-errors

Compressing Into a Single Number

- Diagonal elements of \hat{V} contain variances for each coefficient.
- Standard errors are the square root of these diagonal elements.
- Non-diagonal elements are used for hypothesis tests and confidence intervals for combinations of coefficients.

Roadmap

Introduction

Managing expectations

Introducing difference-in-differences

Potential outcomes

Estimation

Inference

Parallel trends violations

How parallel trends can get violated

DiD in Court

Falsifications

Interesting Applications using staggered adoption

Facebook and Mental Health

Generative AI Randomized Experiment

ChatBot Rollout

Conclusion

Violating parallel trends exercise

- Parallel trends are needed so we can impute the missing $E[Y^0|D = 1]$ with $E[Y^0|D = 0]$ either explicitly or implicitly
- Which means if parallel trends isn't true, then the imputation isn't correct and therefore estimates are biased
- To illustrate this, let's go through the document again

[https://docs.google.com/spreadsheets/d/
1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=
sharing](https://docs.google.com/spreadsheets/d/1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=sharing)

Violating parallel trends

- Parallel trends are in expectation only – we don't rely everybody to follow the same trend, just that the group average for Y^0 be approximately the same for treated and control
- Violations are a form of selection bias and there are two straightforward ways that parallel trends will be violated
 1. Compositional differences in samples associated with repeated cross-sections
 2. Policy endogeneity

Repeated cross-sections and compositional change

- One of the risks of a repeated cross-section is that the composition of the sample may have changed between the pre and post period in ways that are correlated with treatment
- Hong (2013) uses repeated cross-sectional data from the Consumer Expenditure Survey (CEX) containing music expenditure and internet use for a random sample of households
- Study exploits the emergence of Napster (first file sharing software widely used by Internet users) in June 1999 as a natural experiment
- Study compares internet users and internet non-users before and after emergence of Napster

Introduction of Napster and spending on music

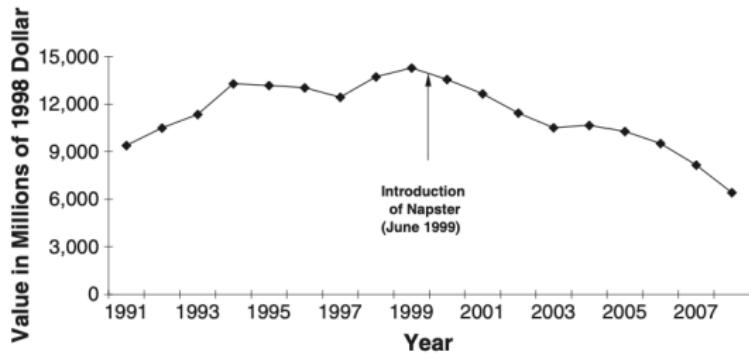


Figure 2. Total real value of record shipments in the USA. Refer to the RIAA's year-end statistics. Total sales include CDs, cassettes, LPs, and music videos. Starting from 2004, total sales also include digital formats such as legitimate download

Figure 1: Internet Diffusion and Average Quarterly Music Expenditure in the CEX

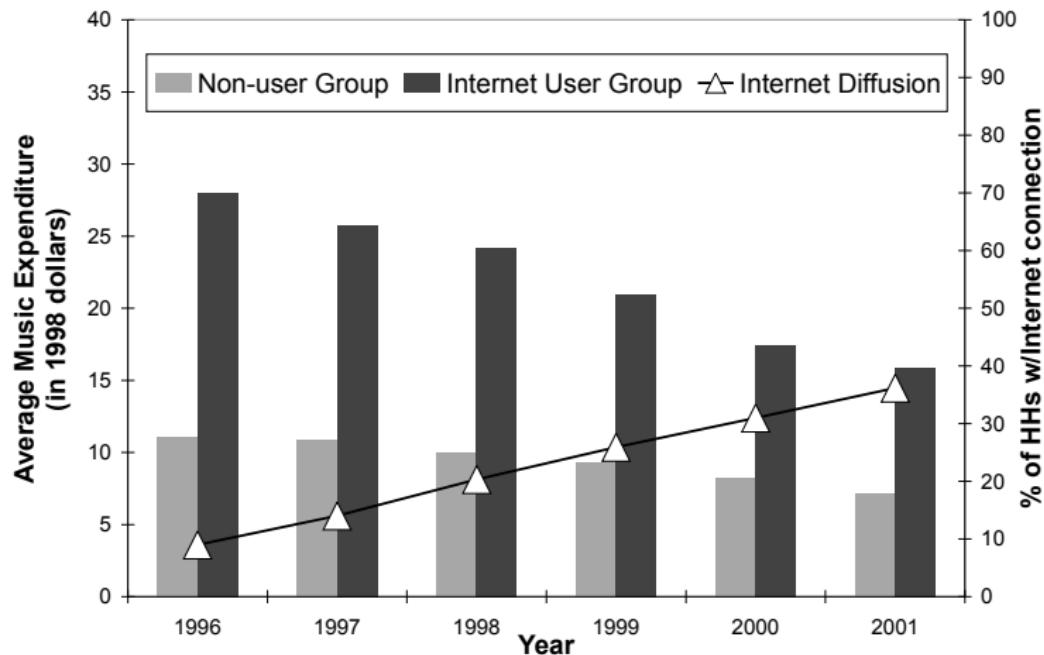


Table 1: Descriptive Statistics for Internet User and Non-user Groups^a

Year	1997		1998		1999	
	Internet User	Non-user	Internet User	Non-user	Internet User	Non-user
Average Expenditure						
Recorded Music	\$25.73	\$10.90	\$24.18	\$9.97	\$20.92	\$9.37
Entertainment	\$195.03	\$96.71	\$193.38	\$84.92	\$182.42	\$80.19
Zero Expenditure						
Recorded Music	.56	.79	.60	.80	.64	.81
Entertainment	.08	.32	.09	.35	.14	.39
Demographics						
Age	40.2	49.0	42.3	49.0	44.1	49.4
Income	\$52,887	\$30,459	\$51,995	\$28,169	\$49,970	\$26,649
High School Grad.	.18	.31	.17	.32	.21	.32
Some College	.37	.28	.35	.27	.34	.27
College Grad.	.43	.21	.45	.21	.42	.20
Manager	.16	.08	.16	.08	.14	.08

Diffusion of the Internet changes samples (e.g., younger music fans are early adopters)

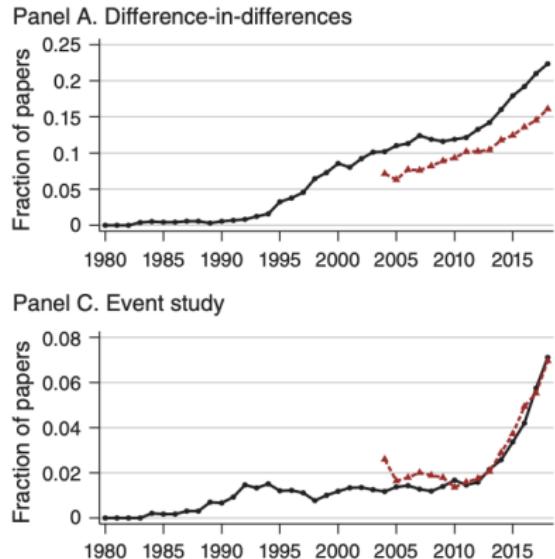
Repeated cross-sections

- Surprisingly underappreciated problem with almost no literature around it
- So what can you do? Check covariate balance by regressing the time-varying covariates instead of the outcome onto the treatment using your OLS specification
- They should be exogenous remember, so this covariate regression can be a helpful test of whether this is a problem
- “Difference-in-differences with Compositional Changes” by Pedro Sant’anna and Qi Xu (not yet released) is the only paper I’ve ever seen to look into it

Types of evidence

- You are building a case, the prosecutor before a judge and jury, always in battle with the defense attorney
- Evidence has particular broadly defined forms that can help you on the front end
- Your goal in my humble opinion should be mixing tight logic based falsifications with particular kinds of data visualization, starting with the event study

Event studies have become mandatory in DiD



Intuition behind event studies

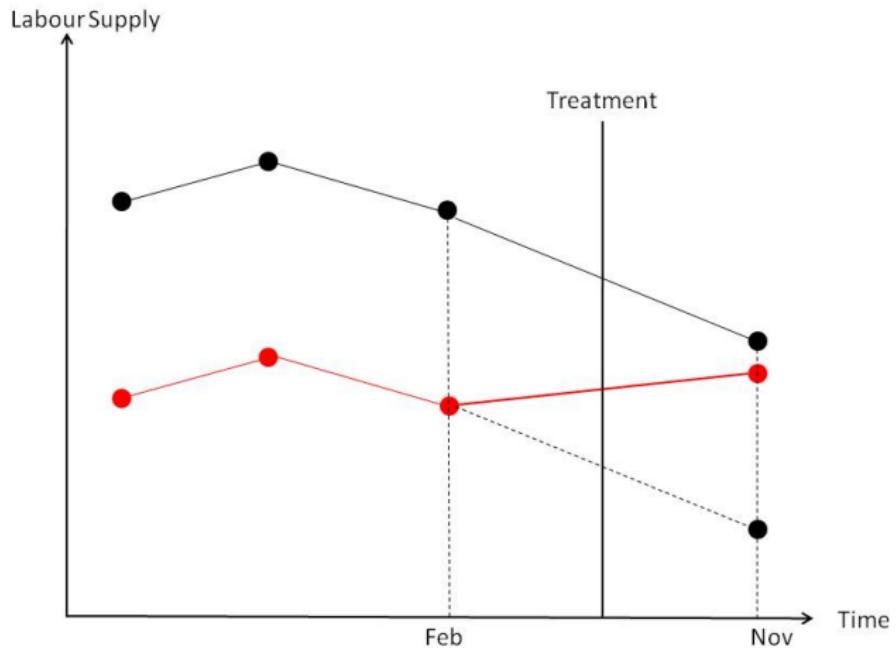
- Princeton Industrial Relations Section seems to be behind this – this intense focus on research design but also verifying assumptions
- The identifying assumption for all DD designs is parallel trends , but since we cannot verify parallel trends, we often look at pre-trends
- It's a type of check for selection bias, but you must understand what it is and what it isn't to see its value but not be naive about it (it is not a silver bullet)
- Even if pre-trends are the same one still has to worry about other policies changing at the same time (omitted variable bias is a parallel trends violation)

Ashenfelter's dip

Orley: <https://youtu.be/M0tbuRX4eyQ?t=960>

Card: https://youtu.be/1soLdywFb_Q?t=3333

Plot the raw data when there's only two groups



Evidence for parallel trends: pre-trends

Let's do the bonus questions on first and second tab now

[https://docs.google.com/spreadsheets/d/
1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=
sharing](https://docs.google.com/spreadsheets/d/1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=sharing)

Beware of naive reasoning

- Parallel pre-trends \neq parallel trends – these are often thought to be the same thing, and they aren't
- Equating them is a kind of *post hoc ergo propter hoc* fallacy
- Parallel pre-trends is more like a smoking gun based on things "looking the same" before
- Similar to checking for covariate balance in RCTs
- But **cannot** substitute for domain knowledge!

Event study regression

- Event studies have a simple OLS specification with only one treatment group and one never-treated group

$$Y_{its} = \alpha + \sum_{\tau=-2}^{-q} \mu_\tau D_{s\tau} + \sum_{\tau=0}^m \delta_\tau D_{s\tau} + \varepsilon_{ist}$$

- where D is an interaction of the treatment group s with the calendar year τ
- Treatment occurs in year 0, no anticipation, drop baseline $t - 1$
- Includes q leads or anticipatory effects and m lags or post treatment effects

Event study regression

$$Y_{its} = \alpha + \sum_{\tau=-2}^{-q} \mu_\tau D_{s\tau} + \sum_{\tau=0}^m \delta_\tau D_{s\tau} + \varepsilon_{ist}$$

Typically you'll plot the coefficients and 95% CI on all leads and lags
(binned or not, trimmed or not)

Under no anticipation, then you expect $\hat{\mu}$ coefficients to be zero, which gives you confidence that parallel trends holds (but is not a guarantee, and there are still specification issues – see Jon Roth's work)

Under parallel trends, $\hat{\delta}$ are estimates of the ATT at points in time

Medicaid and Affordable Care Act example



Volume 136, Issue 3
August 2021

< Previous Next >

Medicaid and Mortality: New Evidence From Linked Survey and Administrative Data [Get access >](#)

Sarah Miller, Norman Johnson, Laura R Wherry

The Quarterly Journal of Economics, Volume 136, Issue 3, August 2021, Pages 1783–1829,

<https://doi.org/10.1093/qje/qjab004>

Published: 30 January 2021

[Cite](#) [Permissions](#) [Share ▾](#)

Abstract

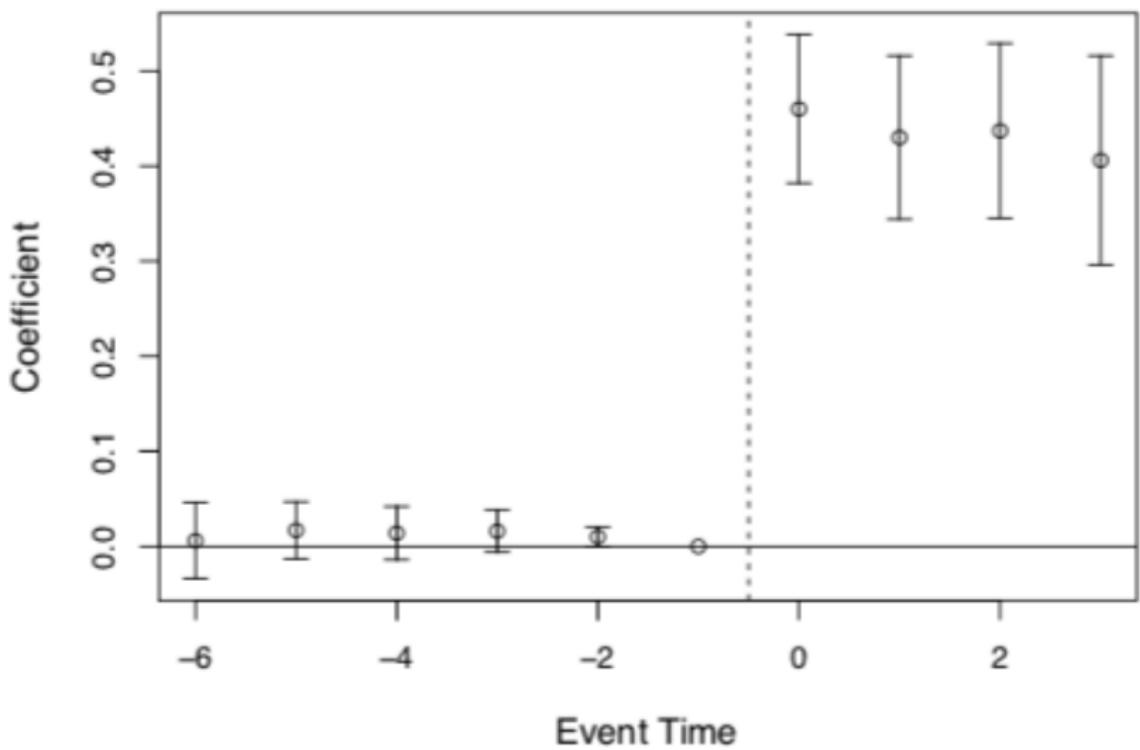
We use large-scale federal survey data linked to administrative death records to investigate the relationship between Medicaid enrollment and mortality. Our analysis compares changes in mortality for near-elderly adults in states with and without Affordable Care Act Medicaid expansions. We identify adults most likely to benefit using survey information on socioeconomic status, citizenship status, and public program participation. We find that prior to the ACA expansions, mortality rates across expansion and nonexpansion states trended similarly, but beginning in the first year of the policy, there were significant reductions in mortality in states that opted to expand relative to nonexpander states. Individuals in expansion states experienced a 0.132 percentage point decline in annual mortality, a 9.4% reduction over the sample mean, as a result of the Medicaid expansions. The effect is driven by a reduction in disease-related deaths and grows over time. A variety of alternative specifications, methods of inference, placebo tests, and sample definitions confirm our main result.

JEL: H75 - State and Local Government: Health; Education; Welfare; Public Pensions, I13 - Health Insurance, Public and Private, I18 - Government Policy; Regulation; Public Health

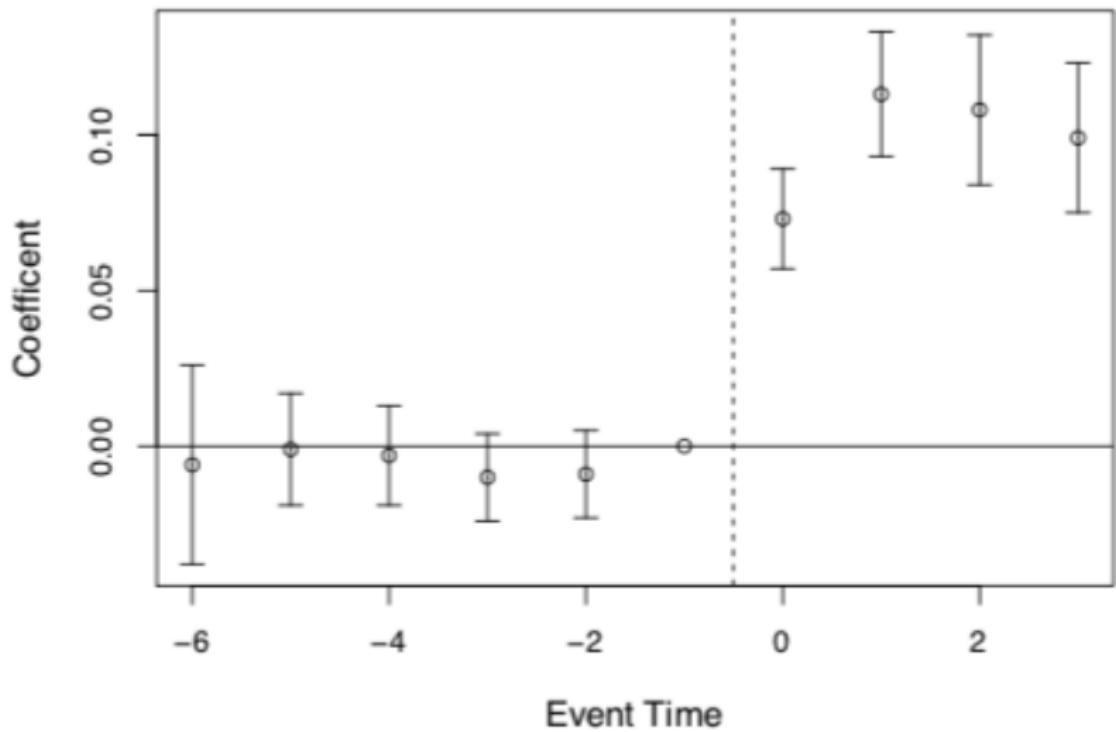
Issue Section: Article

Types of evidence

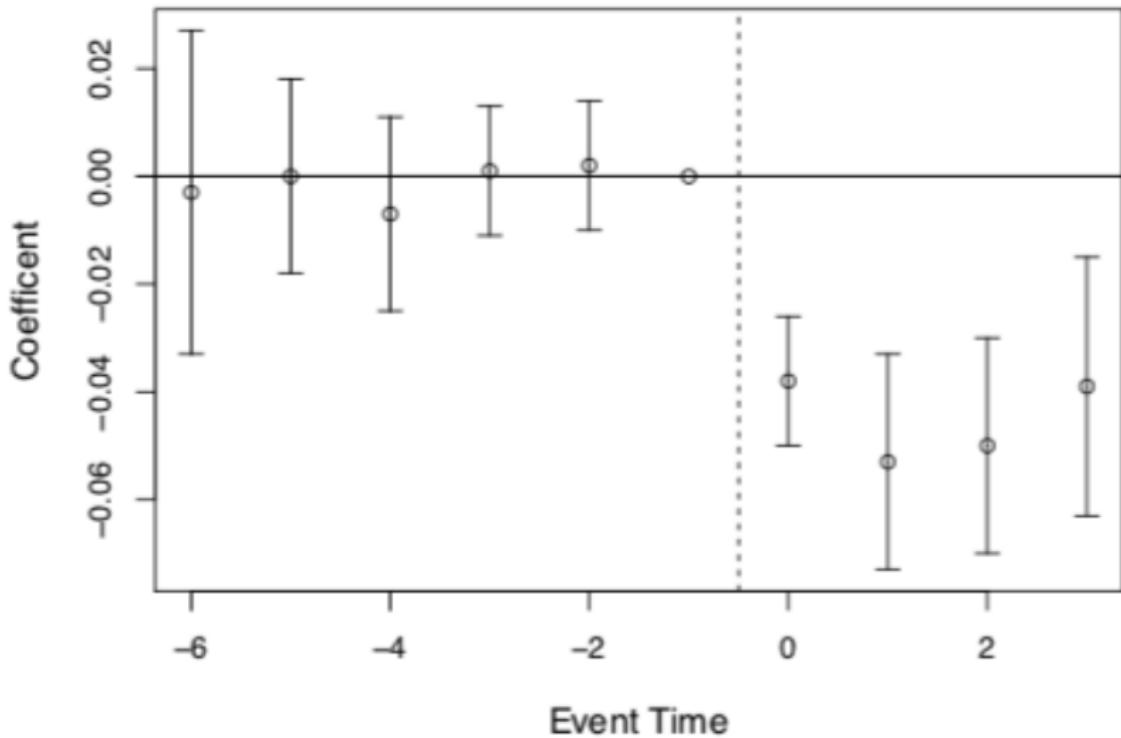
- **Bite** – show that the expansion shifted people into Medicaid and out of uninsured status
- **Main Results** – Show your main results (the point of the paper)
- **Placebos** – Show that there's no effect on mortality for groups it shouldn't be affecting (people 65+)
- **Mechanisms** – Find some reason explaining why the treatment affects the outcome via some “mechanism”
- **Event study** – Show leads and lags on mortality



(a) Medicaid Eligibility



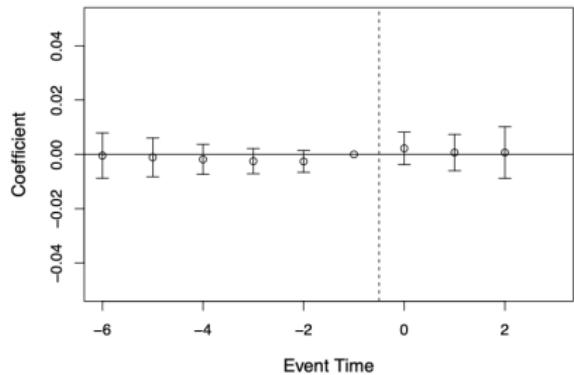
(b) Medicaid Coverage



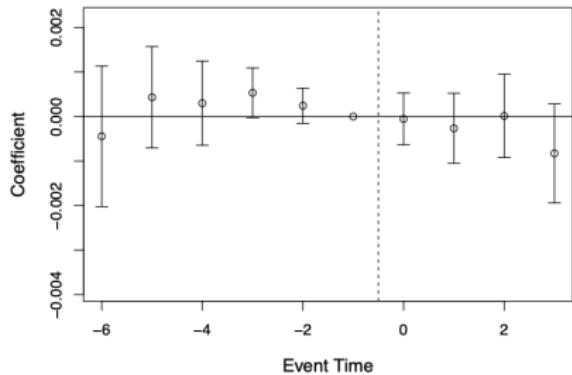
(c) Uninsured

Falsifications on elderly

Age 65+ in 2014



(c) Medicaid Coverage



(d) Annual Mortality

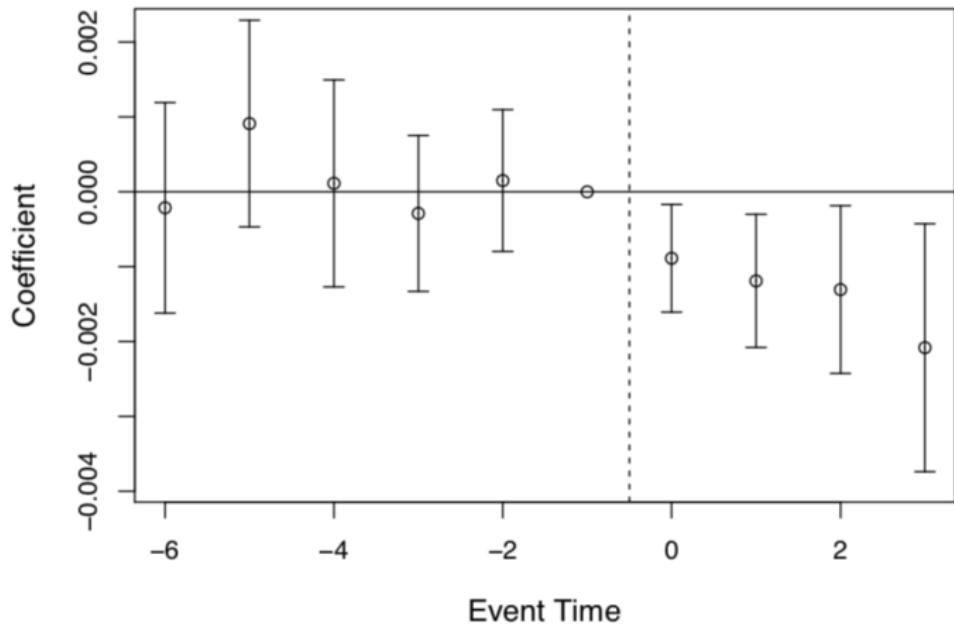


Figure: Miller, et al. (2019) estimates of Medicaid expansion's effects on annual mortality

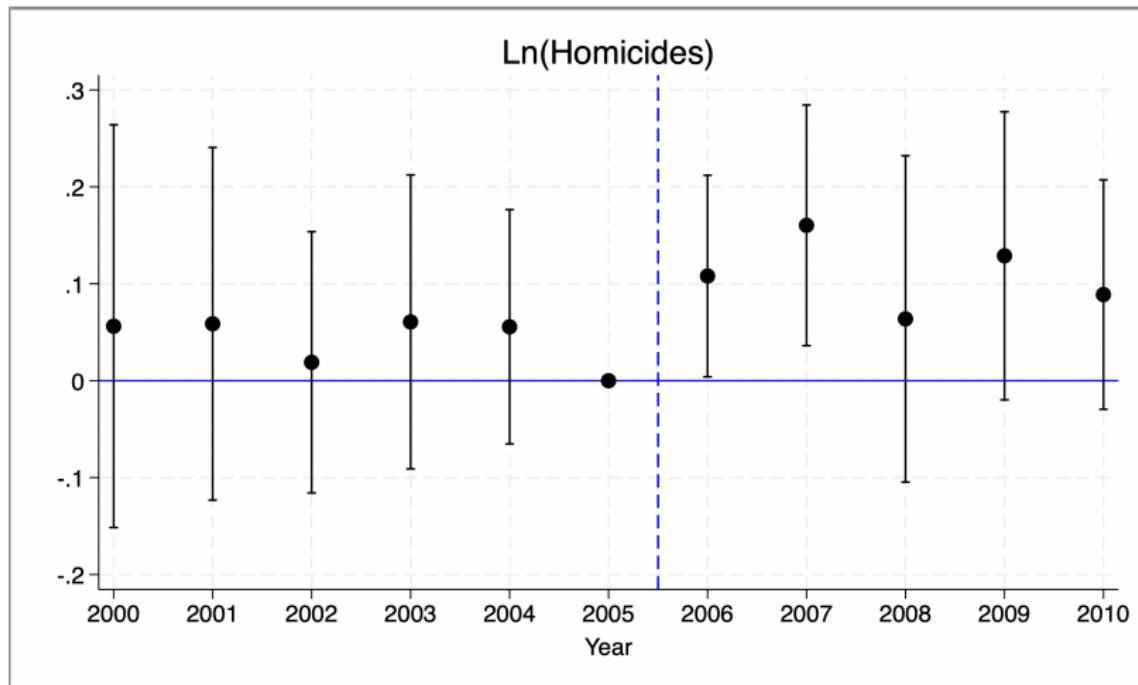
Mechanism

- Bite: Increases in enrollment and reductions in uninsured support that there is adoption of the treatment
- Main Results: 9.2% reduction in mortality among the near-elderly
- Falsifications: no effect on a similar group who isn't eligible
- Mechanism: "The effect is driven by a reduction in disease-related deaths and grows over time."
- Event studies: Compelling once the others are established

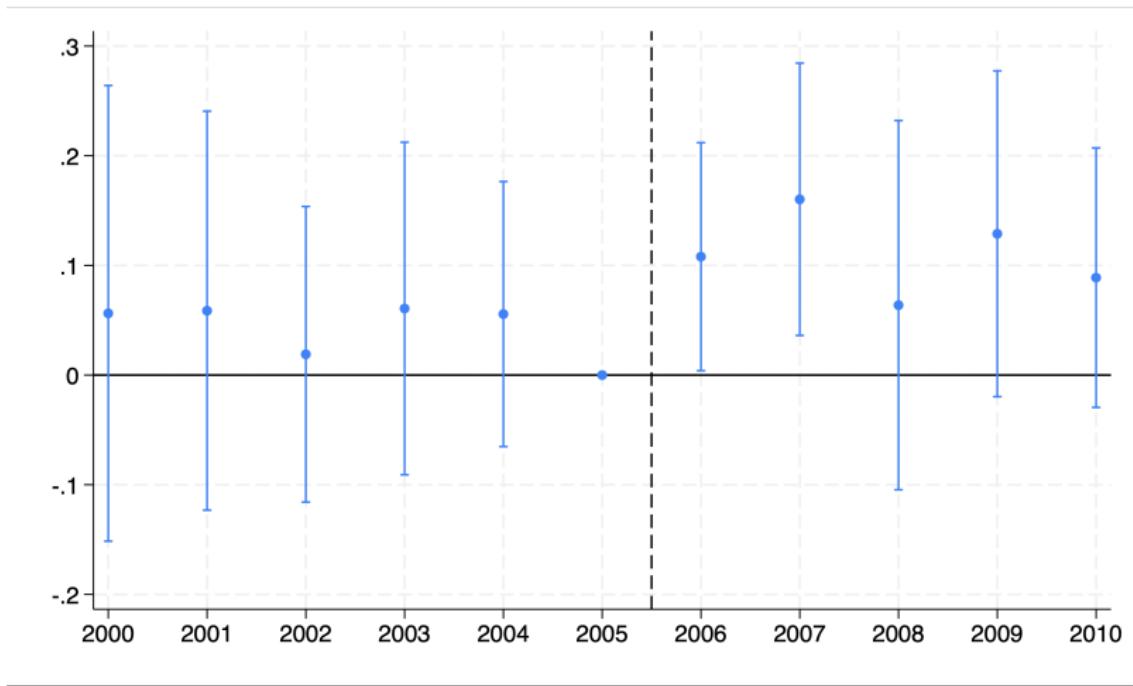
Making event study

- All the simple event study is an interaction between the treatment group dummy and the calendar year dummies
- You must drop a $t - \tau$ as the baseline (e.g., $t - 1$) must be Y^0 untreated comparisons recall
- I have included in a do file that will do it for you either manually or using coefplot in `simple_eventstudy.do` at github labs

Manually creating the event study



Creating the event study with Ben Jann's coefplot



Falsifications on more comparison groups

- Very common for readers and others to request a variety of “robustness checks” from a DD design
- We saw some of these just now (e.g., falsification test using data for alternative control group, the Medicare population)
- Triple differences is another way to do that, but it will have a somewhat different parallel trends assumption

Biased diff-in-diff #1

Table: Biased diff-in-diff #1: comparing states

States	Period	Outcomes	D_1	D_2
Experimental states	Before	$Y = NJ$		
	After	$Y = NJ + NJ_t + D$	$NJ_t + D$	$D + (NJ_t - PA_t)$
Non-experimental states	Before	$Y = PA$		
	After	$Y = PA + PA_t$	PA_t	

$$\widehat{\delta}_{did} = D + (NJ_t - PA_t)$$

The ATT is D. Diff-in-diff is an unbiased estimate of D if $NJ_t = PA_t$. Call parallel trends.

Biased diff-in-diff #2

Table: Biased diff-in-diff #2: comparing groups

Groups	Period	Outcomes	D_1	D_2
Married women	Before	$Y = MW$	$MW_t + D$	$D + (MW_t - SO_t)$
	After	$Y = MW + MW_t + D$		
Single men and older women	Before	$Y = SO$	SO_t	$D + (MW_t - SO_t)$
	After	$Y = SO + SO_t$		

$$\hat{\delta}_{did} = D + (MW_t - SO_t)$$

Also unbiased estimate of D if $MW_t = SO_t$. A different parallel trends assumption.

Two biased diff-in-diffs

- If parallel trends of one doesn't hold, use the other, because both identify the ATT under their respective parallel trends
- But what if both of them are biased – then you can't use diff-in-diff because diff-in-diff requires at least one parallel trends
- Okay, but what if both biases are the same?

$$(NJ_t - PA_t) = (MW_t - SO_t)$$

- Then you can use triple differences, first developed by Gruber (1994)

Triple differences by Gruber (1995)

TABLE 3—DDD ESTIMATES OF THE IMPACT OF STATE MANDATES
ON HOURLY WAGES

Location/year	Before law change	After law change	Time difference for location
A. Treatment Individuals: Married Women, 20–40 Years Old:			
Experimental states	1.547 (0.012) [1,400]	1.513 (0.012) [1,496]	−0.034 (0.017)
Nonexperimental states	1.369 (0.010) [1,480]	1.397 (0.010) [1,640]	0.028 (0.014)
Location difference at a point in time:	0.178 (0.016)	0.116 (0.015)	
Difference-in-difference:		−0.062 (0.022)	
B. Control Group: Over 40 and Single Males 20–40:			
Experimental states	1.759 (0.007) [5,624]	1.748 (0.007) [5,407]	−0.011 (0.010)
Nonexperimental states	1.630 (0.007) [4,959]	1.627 (0.007) [4,928]	−0.003 (0.010)
Location difference at a point in time:	0.129 (0.010)	0.121 (0.010)	
Difference-in-difference:		−0.008 (0.014)	
DDD:		−0.054 (0.026)	

Triple differences commentary

- In Gruber's 1994 article, it isn't clear why he needed triple differences in the first place – his triple differences yielded -0.054 which is almost the same as what he found with his first diff-in-diff (-0.062)
- The main value of triple differences is that you use it when you believe the parallel trends assumption doesn't hold

Table: Difference-in-Difference-in-Differences (Gruber version)

Groups	States	Period	Outcomes	D_1	D_2	D_3
Married women 20-40	Experimental states	After	$NJ + MW + \textcolor{blue}{NJ}_t + \textcolor{red}{MW}_t + D$	$\textcolor{blue}{NJ}_t + MW_t + D$	$D + \textcolor{blue}{NJ}_t - PA_t$	D
		Before	$NJ + MW$			
	Non-experimental states	After	$PA + MW + PA_t + MW_t$	$PA_t + MW_t$	$NJ_t - PA_t$	D
		Before	$PA + MW$			
Single men Older women	Experimental states	After	$NJ + SO + NJ_t + SO_t$	$NJ_t + SO_t$	$NJ_t - PA_t$	D
		Before	$NJ + SO$			
	Non-experimental states	After	$PA + SO + PA_t + SO_t$	$PA_t + SO_t$	$NJ_t - PA_t$	D
		Before	$PA + SO$			

Triple diff assumption

$$\hat{\delta}_{DDD} = D + [(\textcolor{blue}{NJ}_t^{MW} - PA_t^{MW}) - (NJ_t^{SO} - PA_t^{SO})]$$

Equally biased DiD #1 and #2

Triple differences requires two diff-in-diff, from different groups, with the same bias.
 Parallel bias

DDD in Regression

$$\begin{aligned} Y_{ijt} = & \alpha + \beta_2 \tau_t + \beta_3 \delta_j + \beta_4 D_i + \beta_5 (\delta \times \tau)_{jt} \\ & + \beta_6 (\tau \times D)_{ti} + \beta_7 (\delta \times D)_{ij} + \beta_8 (\delta \times \tau \times D)_{ijt} + \varepsilon_{ijt} \end{aligned}$$

- Your panel is now a group j state i (e.g., AR high wage worker 1991, AR high wage worker 1992, etc.)
- Assume we drop τ_t but I just want to show it to you for now.
- If the placebo DD is non-zero, it might be difficult to convince the reviewer that the DDD removed all the bias

Great new paper to learn more



Econometrics Journal (2022), volume 00, pp. 1–23.
<https://doi.org/10.1093/econj/utac010>

The triple difference estimator

ANDREAS OLDEN AND JARLE MØEN

*Dept. of Business and Management Science, NHH Norwegian School of Economics, Hellevn.
30, N-5045 Bergen, Norway.*
Email: andreasolden@gmail.com, jarle.moen@nhh.no

First version received: 14 May 2020; final version accepted: 10 May 2021.

Summary: Triple difference has become a widely used estimator in empirical work. A close reading of articles in top economics journals reveals that the use of the estimator to a large extent rests on intuition. The identifying assumptions are neither formally derived nor generally agreed on. We give a complete presentation of the triple difference estimator, and show that even though the estimator can be computed as the difference between two difference-in-differences estimators, it does not require two parallel trend assumptions to have a causal interpretation. The reason is that the difference between two biased difference-in-differences estimators will be unbiased as long as the bias is the same in both estimators. This requires only one parallel trend assumption to hold.

Keywords: DD, DDD, DID, DiDID, difference-in-difference-in-differences, difference-in-differences, parallel trend assumption, triple difference.

JEL Codes: C10, C18, C21.

1. INTRODUCTION

The triple difference estimator is widely used, either under the name ‘triple difference’ (TD) or the name ‘difference-in-difference-in-differences’ (DDD), or with minor variations of these spellings. Triple difference is an extension of double differences and was introduced by Gruber (1994). Even though Gruber’s paper is well cited, very few modern users of triple difference credit him for his methodological contribution. One reason may be that the properties of the triple difference estimator are considered obvious. Another reason may be that triple difference was little more than a curiosity in the first ten years after Gruber’s paper. On Google Scholar, the annual number of references to triple difference did not pass one hundred until year 2007. Since then, the use of the estimator has grown rapidly and reached 928 unique works referencing it in the year 2017.¹

Looking only at the core economics journals *American Economic Review* (AER), *Journal of Political Economy* (JPE), and *Quarterly Journal of Economics* (QJE), we have found 32 articles using triple difference between 2010 and 2017, see Table A1 in Appendix A. A close reading of these articles reveals that the use of the triple difference estimator to a large extent rests on

¹ More details on the historical development of the use of the triple difference estimator can be found in the working paper version of Olden and Møen (2020, fig. 1). In the working paper, we also analyse naming conventions and suggest that there is a need to unify terminology. We recommend the terms ‘triple difference’ and ‘difference-in-difference-in-differences’.

Falsification on outcomes

- The within-group control group (DDD) is a form of placebo analysis using the same *outcome*
- But there are also placebos using a *different outcome* – but you need a hypothesis of mechanisms to figure out what is in fact a *different outcome*
- Figure out what those are, and test them – finding no effect on placebo outcomes tends to help people your other results interestingly enough
- Cheng and Hoekstra (2013) examine the effect of castle doctrine gun laws on non-gun related offenses like grand theft auto and find no evidence of an effect

Rational addiction as a placebo critique

Sometimes, an empirical literature may be criticized using nothing more than placebo analysis

"A majority of [our] respondents believe the literature is a success story that demonstrates the power of economic reasoning. At the same time, they also believe the empirical evidence is weak, and they disagree both on the type of evidence that would validate the theory and the policy implications. Taken together, this points to an interesting gap. On the one hand, most of the respondents claim that the theory has valuable real world implications. On the other hand, they do not believe the theory has received empirical support."

Placebo as critique of empirical rational addiction

- Auld and Grootendorst (2004) estimated standard “rational addiction” models (Becker and Murphy 1988) on data with milk, eggs, oranges and apples.
- They find these plausibly non-addictive goods are addictive, which casts doubt on the empirical rational addiction models.

Placebo as critique of peer effects

- Several studies found evidence for “peer effects” involving inter-peer transmission of smoking, alcohol use and happiness tendencies
- Christakis and Fowler (2007) found significant network effects on outcomes like obesity
- Cohen-Cole and Fletcher (2008) use similar models and data and find similar network “effects” for things that aren’t contagious like acne, height and headaches
- Ockham’s razor - given social interaction endogeneity (Manski 1993), homophily more likely explanation

Roadmap

Introduction

Managing expectations

Introducing difference-in-differences

Potential outcomes

Estimation

Inference

Parallel trends violations

How parallel trends can get violated

DiD in Court

Falsifications

Interesting Applications using staggered adoption

Facebook and Mental Health

Generative AI Randomized Experiment

ChatBot Rollout

Conclusion

Bringing them together

- An advanced area is when areas adopt policies at different points in time – called differential timing
- Very popular, but methodologically somewhat more complex than the one we've reviewed
- I'm going to just walk you through their findings, and if you're interested in learning more about this, then you can attend my workshop for \$1 in a few weeks!

POSTED JAN 31, 2024 AT 10:32 AM EST

0 Comments (0 New)

A



ADI ROBERTSON

Mark Zuckerberg is trying to reset the conversation on social media's mental health effects. His opening testimony emphasizes that "the existing body of scientific work has not shown a causal link between using social media and young people having worse mental health outcomes." (Which isn't necessarily wrong, partly because causal links are very hard to prove — the overall body of research is complicated and seems to suggest social media has a variety of possible effects.)



All the news from Congress' Big Tech child safety hearing

ADI ROBERTSON JAN 31

Mental health and Social Media

- Unclear what he means; he may mean there is no experimental evidence
- Very difficult to imagine a randomized experiment – especially once the claim out there is that it is harmful, Institutional Review Boards likely wouldn't approve it
- Quasi-experimental evidence can step in to answer important questions like this
- Braghieri, Levy and Makarin (2022), "Social Media and Mental Health", *American Economic Review*, 112(11): 3660-3693

Overview of design and data

- Authors take advantage of a clever quirk in Facebook (then “theFacebook”) targeted different universities from 2004 to 2006
- They found an online data source that allowed them to pin point precisely when a university was “treated” with theFacebook
- They then linked that data with a longrunning health survey of college students (both before and after) in a very clever way
- Estimated the effect of a new social media platform’s presence at a university on student revealed mental health problems

DID in Court

Five elements of a strong DiD

1. **Bite:** **Nothing.** They cannot really show much here. No data on Facebook usage. They had to rely on anecdote and Facebook as a "first mover", but there had been Friendster and MySpace so this does weaken the paper maybe
2. **Main Results:** Very strong evidence, mostly expressed using rich survey data and questions transformed into z-scores (standard deviations)
3. **Falsifications:** **None.** Authors do not perform falsifications. Remember Miller, Johnson and Wherry looking at Medicaid's effect on Medicare eligible population? There isn't anything like that here.
4. **Event studies:** Extremely compelling evidence and robustness across a half dozen different models
5. **Mechanism:** **Very weak in my opinion**

DiD in Court

- So in many ways the strength of the project lies in a few areas:
 1. Important question – social media and youth mental health problems is a major policy question (see Zuckerberg testifying before Congress about it)
 2. Excellent research design – difference-in-differences
 3. Meticulous data collection
 4. Data visualization is compelling
- And it publishes in the premiere journal in economics, which I think shows that the research question and high quality data combined with research design can lift a paper

$$Y_{icgt} = \alpha_g + \delta_t + \beta \times Facebook_{gt} + X_i \times \gamma + X_c \times \psi + \varepsilon_{icgt} \quad (1)$$

This is a version of the regression model we looked at called "twoway fixed effects". Somewhat complicated to dig into, so I will just say that they use it plus some other methods that are appropriate when you have several difference-in-differences events. But the focus is on β

Data on Facebook

- When does Facebook appear at a school?
 - Facebook only publishes a fraction of that information
 - They came up with a workaround
- The Wayback Machine has been taking almost daily photographs of every website since the Internet's beginning – including the frontpage of "TheFacebook"
- Guess what was on the front page of TheFacebook ...

[DONATE](#)

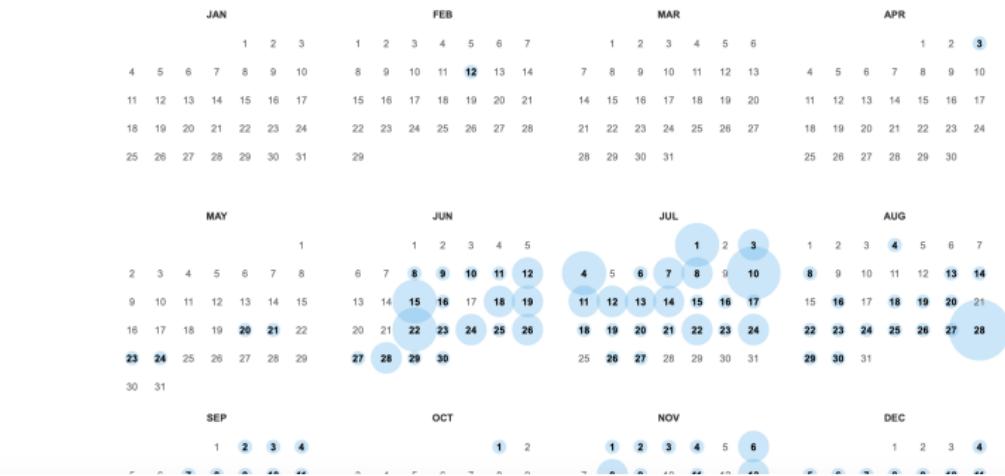
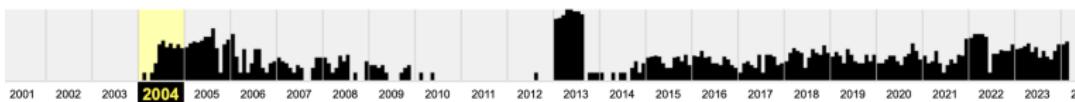
Explore more than 294 billion web pages saved over time

thefacebook.com

X

[Calendar](#)[Collections](#)[Changes](#)[Summary](#)[Site Map](#)[URLs](#)

Saved 7,057 times between February 12, 2004 and February 16, 2024.





[thefacebook]

[login](#) [register](#) [about](#)

Email:

Password:

[login](#) [register](#)

Welcome to Thefacebook!

[Welcome to Thefacebook]

Thefacebook is an online directory that connects people through social networks at colleges.

We have opened up Thefacebook for popular consumption at **Harvard, Columbia, Stanford, Yale, Cornell, Dartmouth, UPenn, MIT**, and now **BU** and **NYU**.

Your facebook is limited to your own college or university.

You can use Thefacebook to:

- Search for people at your school
- Find out who is in your classes
- Look up your friends' friends
- See a visualization of your social network

To get started, click below to register. If you have already registered, you can log in.

[Register](#)

[Login](#)

[about](#) [contact](#) [faq](#) [terms](#) [privacy](#)

a Mark Zuckerberg production

Thefacebook © 2004

Go JUN SEP OCT
◀ 02 ▶ 2004 2005 ▼ Abd

Welcome to Thefacebook!

[Welcome to Thefacebook]

Thefacebook is an online directory that connects people through social networks at colleges.

We have opened up Thefacebook for popular consumption at:

Amherst • BC • Berkeley • Bowdoin • Brown • BU • Bucknell
Caltech • Carnegie Mellon • Chicago • Columbia • Cornell • Dartmouth
Duke • Emory • Florida • Florida State • Georgetown • Georgia • GWU • Hamilton
Harvard • Illinois • Johns Hopkins • Maryland • Michigan
Michigan State • Middlebury • MIT • Northeastern • Northwestern • Notre Dame
NYU • Oberlin • Penn • Princeton • Rice • Rochester • South Florida
Stanford • Swarthmore • Syracuse • Tufts • Tulane • UCDavis
UCF • UCLA • UCSD • UNC • USC • UVA • Vanderbilt • WashU
Wellesley • Wesleyan • Williams • Yale

Your facebook is limited to your own college or university.

You can use Thefacebook to:

- Search for people at your school
- Find out who is in your classes
- Look up your friends' friends
- See a visualization of your social network

To get started, click below to register. If you have already registered, you can log in.

[Register](#) [Login](#)

about contact jobs faq advertise terms privacy
a Mark Zuckerberg production
Thefacebook © 2004

Go MAR APR MAY
◀ 28 ▶
2004 2005 2006



[thefacebook]

login register about faq

Email:

Password:

[login](#) [register](#)

Welcome to Thefacebook!

[Welcome to Thefacebook]

Thefacebook is an online directory that connects people through social networks at colleges.

We have recently opened up Thefacebook at the following schools:

Abilene Christian • Agnes Scott • Albright • Allegheny
Anderson • Angelo State • Arcadia • Arts • Azusa Pacific
Beloit • Boise State • C. Arkansas • C. Missouri • Cal. Lutheran
Campbell • Canisius • Capital • Carthage • Christian Brothers
Clark Atlanta • Cleveland State • Columbus State • CSCC
CUNY City • Delta State • DeSales • Edgerton • Endicott • ETSU
Evansville • Frostburg • Guilford • Gustavus • Hampden-Sydney
Hartwick • Hendrix • Illinois Tech • IPFW • Jacksonville
John Jay • Kettering • Lake Forest • Lamar • Liberty • Lock Haven
McDaniel • Messiah • Milligan • MSOE • Murray State • N. Georgia
N. Kentucky • NJIT • Nova • NYIT • Otterbein • Philadelphia
RIC • S. Alabama • S.E. Louisiana • Saginaw Valley • Salem State
Shepherd • St. Cloud • St. Rose • Sweet Briar • Tarleton
TN Chattanooga • TN Tech • UMass Lowell • Valencia • W. Florida
W. Oregon • Widener • WNEC • Wofford • Yeshiva

For a complete list of supported schools, click [here](#).

Your facebook is limited to your own college or university.

You can use Thefacebook to:

- Search for people at your school
- Find out who is in your classes
- Look up your friends' friends
- See a visualization of your social network

To get started, click below to register. If you have already registered, you can log in.

[Register](#) [Login](#)

[about](#) [contact](#) [jobs](#) [announce](#) [advertise](#) [terms](#) [privacy](#)
a Mark Zuckerberg production
Thefacebook © 2005

Timing Dates

- They went through three years of daily screenshots on Wayback machine to find when a school appeared on the front page
- The first time Agnes Scott, or Covenant, appears on the front page, the authors mark that as the date when the school got Facebook
- But now they need information on mental health outcomes
- They find it with an old long running repeated cross section survey of college students

NCHA survey by ACHA

*Our second main data source consists of more than 430,000 responses to the NCHA survey, a survey administered to college students on a semi-annual basis by the American College Health Association (ACHA). The NCHA survey was developed in 1998 by a team of college health professionals with the purpose of obtaining information from college students about their mental and physical health. Specifically, the NCHA survey inquires about demographics, physical health, **mental health**, alcohol and drug use, sexual behaviors, and perceptions of these behaviors among one's peers.*

No evidence of bite

The NCHA survey does not include any questions on social media use; therefore, it is not possible for us to determine whether a particular survey respondent had a Facebook account.

This is probably the problem in any study in which your treatment is more or less the first of its kind – most likely the standard surveys have not yet incorporated the questions into their surveys

Linking Facebook data with NCHA data

In order to protect the privacy of the institutions that participate in the NCHA survey while still allowing us to carry out the analysis, the ACHA kindly agreed to provide us with a customized dataset that includes a variable indicating the semester in which Facebook was rolled out at each college. Specifically, the ACHA adopted the following procedure: (i) merge our dataset containing the Facebook introduction dates to the NCHA dataset; (ii) add a variable listing the semester in which Facebook was rolled out at each college;¹⁵ (iii) strip away any information that could allow us to identify colleges (including the specific date in which Facebook was introduced at each college).

Basic facts about early and late adopters

- Colleges in earlier Facebook expansion groups are more selective in terms of test scores, larger, more likely to be on the East Coast, and have more residential undergraduate programs than colleges in later Facebook expansion groups.
- Colleges in earlier Facebook expansion groups enroll students from relatively more advantaged economic backgrounds.
- Students in colleges that received Facebook relatively earlier have worse baseline mental health outcomes than students attending colleges in later Facebook expansion groups.

Measurement

- The survey data is very rich with a lot of questions about mental health with different scales
- They create their own combinations of these questions into aggregate indices – “index of poor mental health” where higher numbers mean worse mental health
- Each outcome survey question is normalized into what is called a “z-score” which is interpreted as a fraction of a standard deviation
- Estimates are ATT parameters – average effect of Facebook on students at schools that got Facebook

TABLE 1—BASELINE RESULTS: INDEX OF POOR MENTAL HEALTH

	Index of poor mental health			
	(1)	(2)	(3)	(4)
Post-Facebook introduction	0.137 (0.040)	0.124 (0.022)	0.085 (0.033)	0.077 (0.032)
Observations	374,805	359,827	359,827	359,827
Survey-wave fixed effects	✓	✓	✓	✓
Facebook-expansion-group fixed effects	✓	✓		
Controls		✓	✓	✓
College fixed effects			✓	✓
FB-expansion-group linear time trends				✓

Notes: This table explores the effect of the introduction of Facebook at a college on student mental health. Specifically, it presents estimates of coefficient β from equation (1) with our index of poor mental health as the outcome variable. The index is standardized so that, in the preperiod, it has a mean of zero and a standard deviation of one. Column 1 estimates equation (1) without including controls; column 2 estimates equation (1) including controls; column 3, our preferred specification, replaces Facebook-expansion-group fixed effects with college fixed effects; column 4 includes linear time trends estimated at the Facebook-expansion-group level. Our controls consist of age, age squared, gender, indicators for year in school (freshman, sophomore, junior, senior), indicators for race (White, Black, Hispanic, Asian, Indian, and other), and an indicator for international student. Column 2 also includes indicators for geographic region of college (Northeast, Midwest, West, South); such indicators are omitted in columns 3 and 4 because they are collinear with the college fixed effects. For a detailed description of the outcome, treatment, and control variables, see online Appendix Table A.31. Standard errors in parentheses are clustered at the college level.

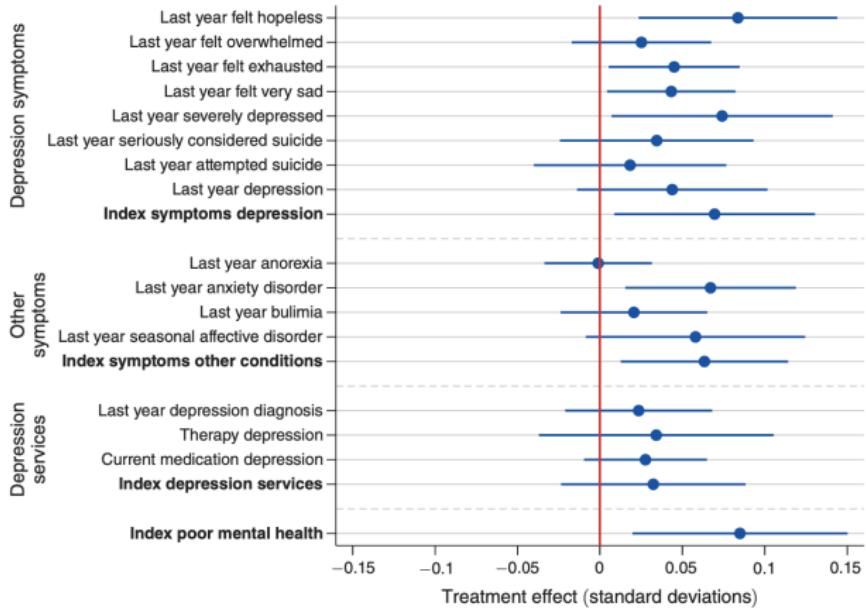


FIGURE 1. EFFECTS OF THE INTRODUCTION OF FACEBOOK ON STUDENT MENTAL HEALTH

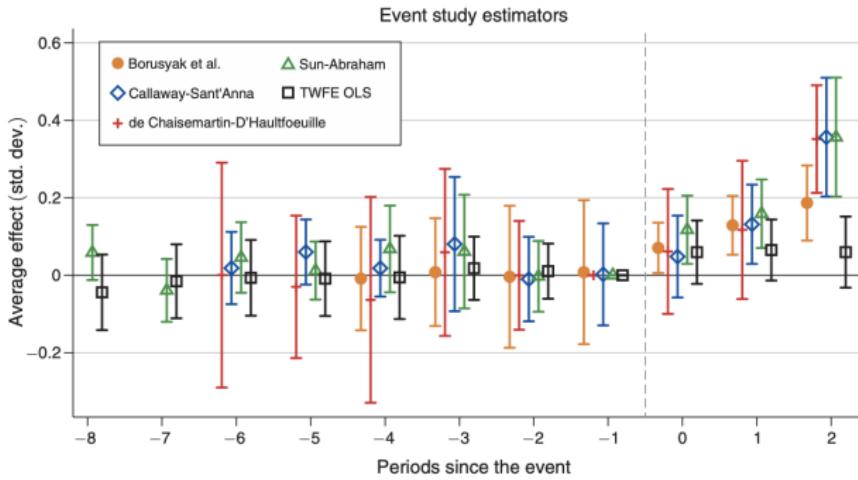


FIGURE 2. EFFECTS OF FACEBOOK ON THE INDEX OF POOR MENTAL HEALTH BASED ON DISTANCE TO/FROM FACEBOOK INTRODUCTION

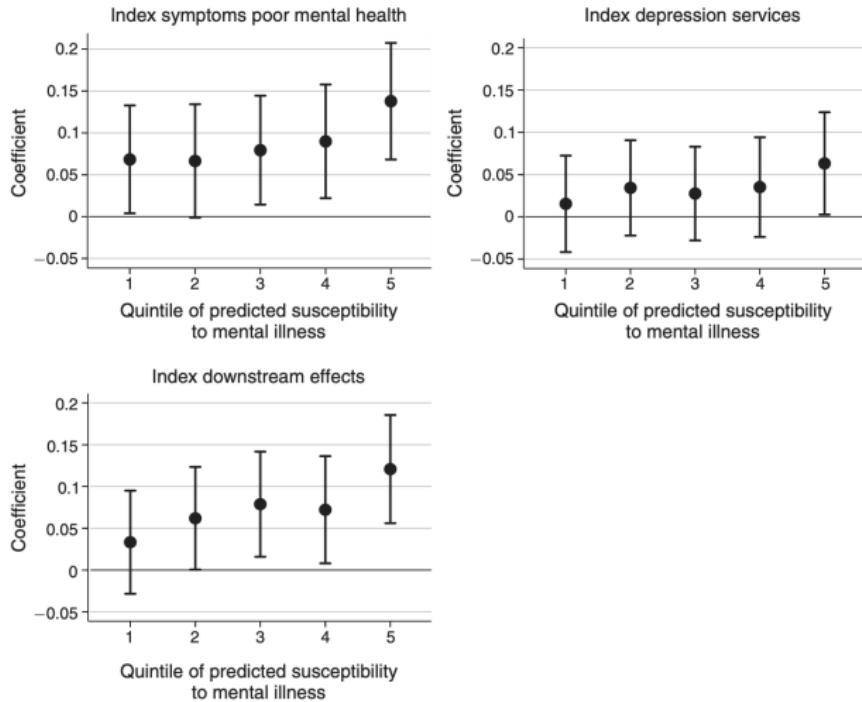


FIGURE 3. HETEROGENEOUS EFFECTS BY PREDICTED SUSCEPTIBILITY TO MENTAL ILLNESS

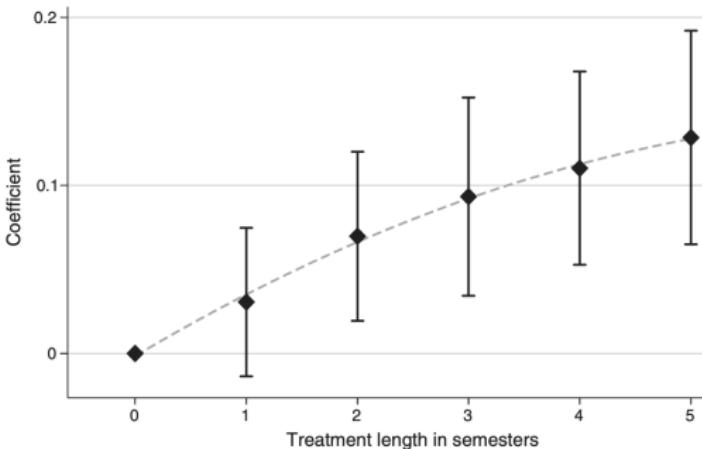


FIGURE 4. EFFECT ON POOR MENTAL HEALTH BY LENGTH OF EXPOSURE TO FACEBOOK

Notes: This figure explores the effects of length of exposure to Facebook on our index of poor mental health by presenting estimates of equation (4). The index is standardized so that, in the preperiod, it has a mean of zero and a standard deviation of one. The dashed curve is the quadratic curve of best fit. Our controls consist of age, age squared, gender, indicators for year in school (freshman, sophomore, junior, senior), indicators for race (White, Black, Hispanic, Asian, Indian, and other), and an indicator for international student. Students who entered college in 2006 might have been exposed to Facebook already in high school, because, starting in September 2005, college students with Facebook access could invite high school students to join the platform. Such students are excluded from the regression. For a detailed description of the outcome, treatment, and control variables, see online Appendix Table A.31. The bars represent 95 percent confidence intervals. Standard errors are clustered at the college level.

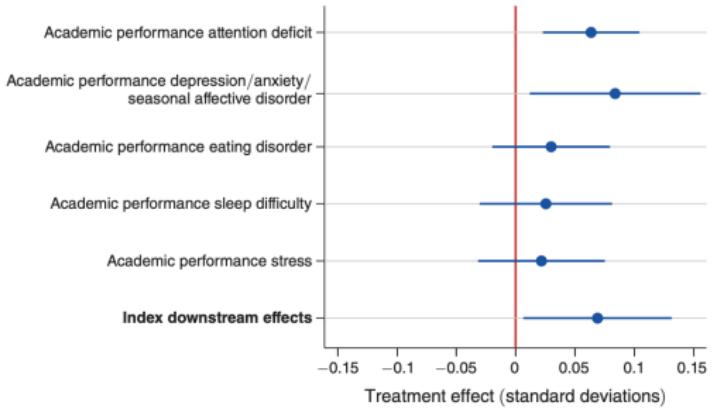


FIGURE 5. DOWNSTREAM EFFECTS ON ACADEMIC PERFORMANCE

Notes: This figure explores downstream effects of the introduction of Facebook on the students' academic performance. It presents estimates of coefficient β from equation (1) using our preferred specification, including survey-wave fixed effects, college fixed effects, and controls. The outcome variables are answers to questions inquiring as to whether various mental health conditions affected the students' academic performance and our index of downstream effects. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. For a detailed description of the outcome, treatment, and control variables, see online Appendix Table A.31. The bars represent 95 percent confidence intervals. Standard errors are clustered at the college level.

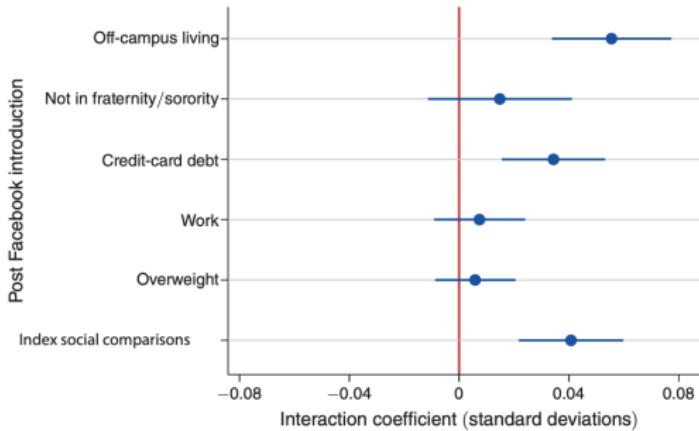


FIGURE 6. HETEROGENEOUS EFFECTS AS EVIDENCE OF UNFAVORABLE SOCIAL COMPARISONS

Notes: This figure explores the mechanisms behind the effects of Facebook on mental health. It presents estimates from a version of equation (1) in which our treatment indicator is interacted with a set of indicators for belonging to a certain subpopulation of students. The outcome variable is our overall index of poor mental health. The estimates are obtained using our preferred specification, namely the one including survey-wave fixed effects, college fixed effects, and controls. For a detailed description of the outcome, treatment, interaction, and control variables, see online Appendix Table A.31. The bars represent 95 percent confidence intervals. Standard errors are clustered at the college level.

Questions and Comments

- Mark Zuckerberg quote: “the existing body of scientific work has not shown a causal link between using social media and young people having worse mental health outcomes” – What is your reaction to his claim?
- What is your reaction to this study’s evidence?
- Which parts of this study do you think is more memorable and more convincing and why?
- How might you replicate this study yourself?

Generative AI and Workers

- Generative AI systems create content that mimics human creativity across various media.
- May have large impacts on some kinds of jobs, but unclear to what degree and how soon
- Unclear whether we should be concerned, and we can't devise a lot of policies if we don't understand what effects might be

What is a Large Language Model?

- Large Language Models (LLMs) like ChatGPT3, 3.5 and 4 are at the forefront of this change.
- LLMs learn from (“train on”) large datasets of writings to produce human-like language.
- Think of it like autocorrect on your phone, only it’s generating entire conversations, not just the next letter or word
- These models power diverse applications, from chatbots to data analysis tools.

The Evolution of OpenAI

- OpenAI was founded in 2015 by Sam Altman, Elon Musk, Greg Brockman, and several others as a non-profit aimed to ensure artificial intelligence benefits all of humanity, marking a commitment to openness and safety in AI development.
- Initially, OpenAI focused on reinforcement learning and unsupervised learning but quickly adapted to the transformative potential of transformers after 2017 when Google published a paper entitled "Attention is All You Need" introducing transformer models
- The organization's pivot towards transformer-based models like GPT was driven by the architecture's ability to process and generate natural language more effectively than previous models.

Transformers and Generative AI

- The introduction of the transformer model in the "Attention Is All You Need" paper represented a paradigm shift in AI, leading to the development of more advanced and efficient language models.
- OpenAI capitalized on this breakthrough, leading to the creation of the Generative Pretrained Transformer series, which significantly advanced the field of natural language processing.
- Quickly they realized that the more data they fed these models, the better they got – nonlinear, not linear, so they invested heavily in computing powers (billions of dollars raised, Microsoft partnership)

OpenAI's Response and Innovations Post-2017

- OpenAI's strategic shift to transformers culminated in the release of GPT, GPT-2, and the groundbreaking GPT-3, each setting new benchmarks for AI's capability in understanding and generating human-like text.
- The development and release of these models not only showcased OpenAI's leadership in AI research but also underscored the potential of large language models in various applications.

The Growth of OpenAI

- OpenAI's journey from a non-profit organization to the creation of a capped-profit arm in 2019 allowed it to attract significant investment and scale its operations, while still adhering to its mission.
- The subsequent development and launch of GPT-3.5 and GPT-4 illustrated not only technological advancements but also a significant increase in user adoption and commercial success.

Impact and Future Directions

- Exciting time in many ways, but also worrisome
- Technologies have always reallocated work – think of the old horse and buggy industry being replaced by vehicles
- But these technologies may be impacting high skill work which would make it different from the Industrial Revolution
- Two causal studies (one experiment and one diff-in-diff) are going to examine this

Two studies

- First is a randomized controlled trial (RCT) in which ChatGPT-4 was randomly assigned to consultants at Boston Consulting Group to examine effects on various performance measures
- Second is a diff-in-diff when a firm rolled out its own generative AI chatbot to assist customer service representatives
- In both we are trying to understand the impact this had on worker productivity, but we will ask ourselves in the end what we might expect to happen to demand too

Study 1: the RCT

- Professors at Penn, Harvard and MIT worked with Boston Consulting Group, a global management consulting firm, to examine the performance implications of ChatGPT-4 on realistic, complex and knowledge-intensive tasks
- Experiment was registered ahead of time and involved 758 consultants comprising 7% of the individual consultants at the company
- They then created a realistic, yet fictional, sneaker client and observed consultants' performance on tasks afterwards

Design

Each consultants' baseline skill on tasks was measured and then each was randomized to one of the three following "treatment" conditions

1. No AI access
2. ChatGPT-4 AI access
3. ChatGPT-4 access with prompt engineering overview

Two types of tasks

- We tested two distinct tasks: one situated outside the frontier of AI capabilities and the other within its bounds.
- The experiment aimed to understand how AI integration might reshape the traditional workflows of these high human capital professionals.

Findings

Our results show that this generation of LLMs are highly capable of causing significant increases in quality and productivity, or even completely automating some tasks, but the actual tasks that AI can do are surprising and not immediately obvious to individuals or even to producers of LLMs themselves. Because this frontier is expanding and changing, the overall results suggest that AI will have a large impact on work, one which will increase with LLM capabilities, but where the impacts occur will be uneven.

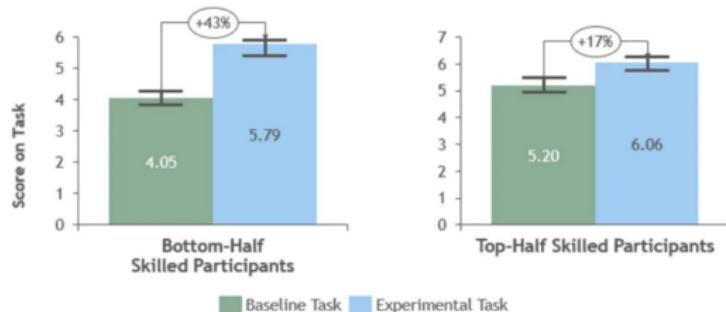
Inside the Frontier Tasks



Distribution of output quality across all the tasks. The blue group did not use AI, the green and red groups used AI, the red group got some additional training on how to use AI.

Inside the Frontier Tasks

Figure 5: Bottom-Half Skills and Top-Half Skills - Inside the Frontier



Outside the Frontier Tasks

- But a number of things in the study workers treated with ChatGPT-4 performed worse than the control group
- For those “outside the frontier” tasks, consultants using ChatGPT-4 were 19 percentage points less likely to produce correct solutions compared to control group

Summary

- So we have some workers – particularly the least skilled – becoming just as productive with ChatGPT-4 as the most skilled worker without it
- But not all types of work does AI help with – basic math, interestingly, it cannot help with; but for the creation of new ideas, and creative work more generally, for some reason it seems have some improvements
- That was an RCT with skilled workers – let's now move to a DiD with a generative AI chatbot rolled out to all of the workers over time

Working Paper on ChatBot

NBER WORKING PAPER SERIES

GENERATIVE AI AT WORK

Erik Brynjolfsson

Danielle Li

Lindsey R. Raymond

Working Paper 31161

<http://www.nber.org/papers/w31161>

NATIONAL BUREAU OF ECONOMIC RESEARCH

1050 Massachusetts Avenue

Cambridge, MA 02138

April 2023

Chatbots and Workers

- An unnamed firm released gradually a generative AI-based conversational assistant chatbot to its 5,179 customer support agents
- These chatbots provided assistance in handling complaints
- Very stressful job as the only time customers reached out was when they were very upset
- It isn't a randomized experiment so they're going to estimate the effect of the adoption of the chatbot using difference-in-differences

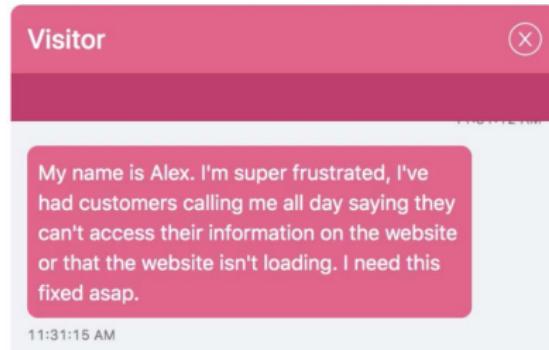
Outcomes and Pictures

- Main focus is on various measures of customer support agents handling of calls, which is the proxy for their productivity
- But they also focus on high and low skill workers (heterogeneity like before)
- Authors are going to present evidence almost entirely using event study graphs
- They also present regression tables, but the event study graphs are very powerful

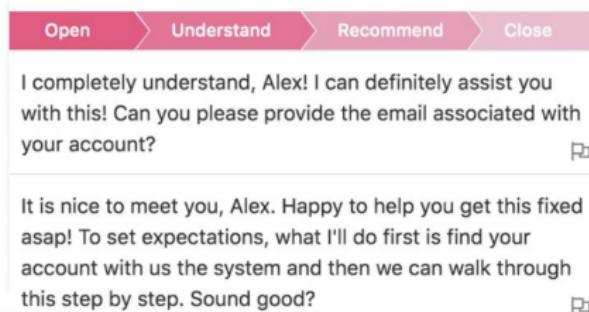
Example of ChatBot

FIGURE 1: SAMPLE AI OUTPUT

A. SAMPLE CUSTOMER ISSUE

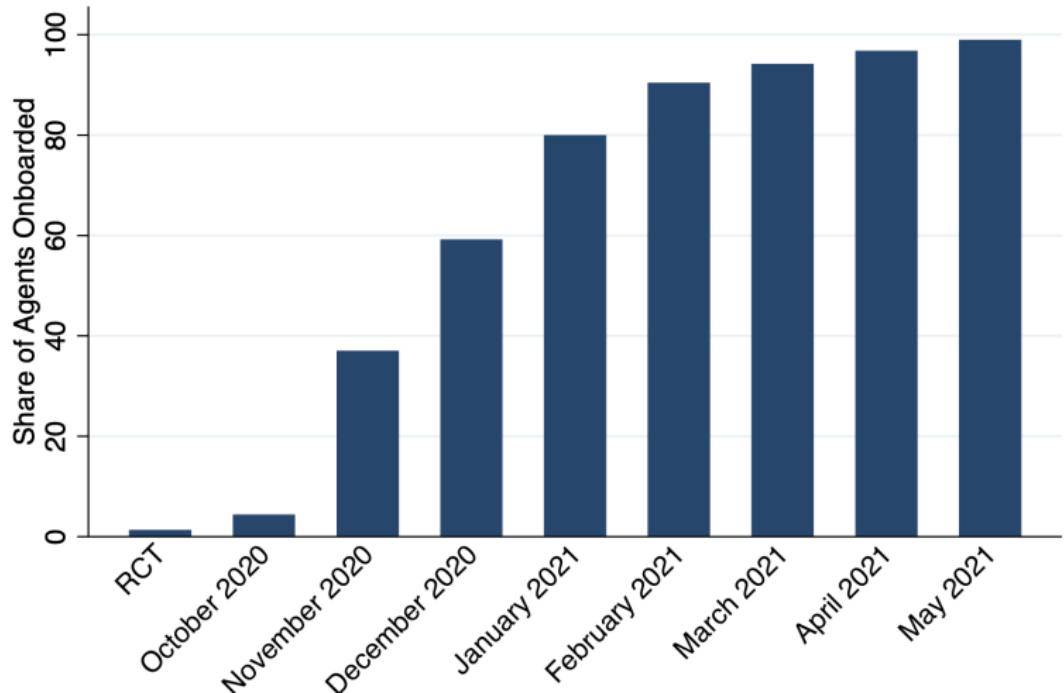


B. SAMPLE AI-GENERATED SUGGESTED RESPONSE



Rollout

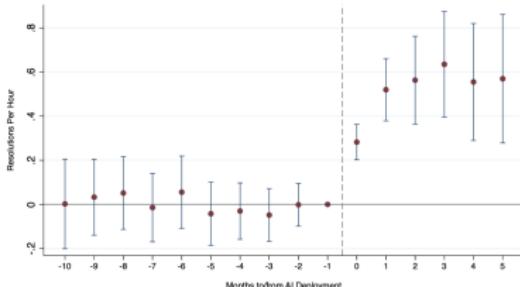
FIGURE 2: DEPLOYMENT TIMELINE



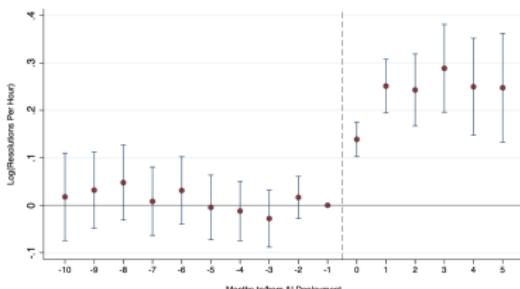
Resolutions of Customer Problems

FIGURE 4: EVENT STUDIES, RESOLUTIONS PER HOUR

A. RESOLUTIONS PER HOUR



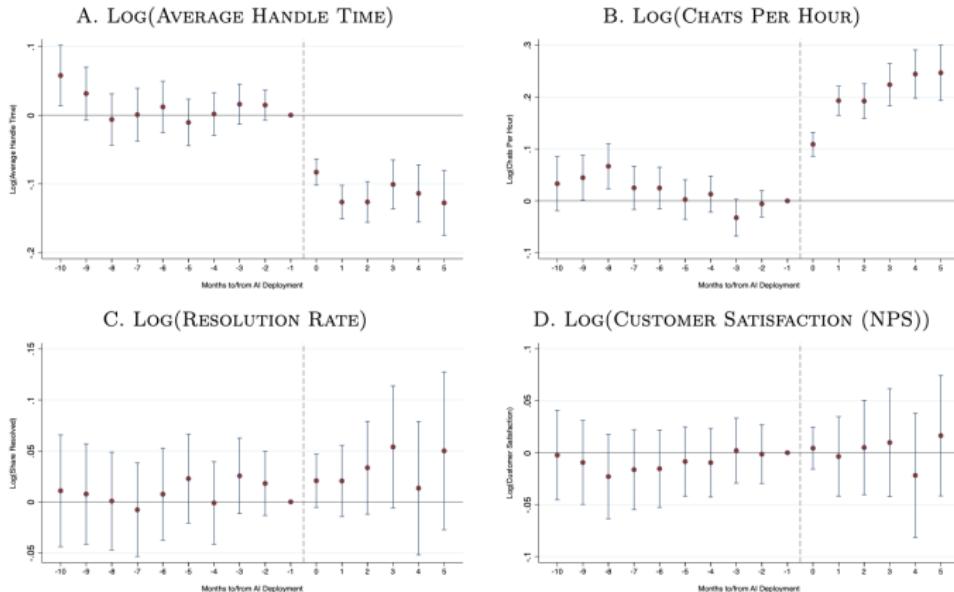
B. LOG(RESOLUTIONS PER HOUR)



NOTES: These figures plot the coefficients and 95 percent confidence interval from event study regressions of AI model deployment using the Sun and Abraham (2021) interaction weighted estimator. See text for additional details. Panel A plots the resolutions per hour and Panel B plots the natural log of the measure. All specifications include agent and chat year-month, location, agent tenure and company fixed effects. Robust standard errors are clustered at the agent level.

Additional Outcomes

FIGURE 5: EVENT STUDIES, ADDITIONAL OUTCOMES

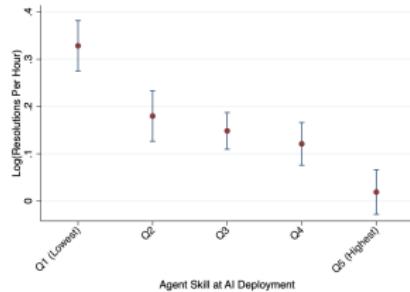


NOTES: These figures plot the coefficients and 95 percent confidence interval from event study regressions of AI model deployment using the Sun and Abraham (2021) interaction weighted estimator. See text for additional details. Panel A plots the average handle time or the average duration of each technical support chat. Panel B plots the number of chats an agent completes per hour, incorporating multitasking. Panel C plots the resolution rate, the share of chats successfully resolved, and Panel D plots net promoter score, which is an average of surveyed customer satisfaction. All specifications include agent and chat year-month, location, agent tenure and company fixed effects. Robust standard errors are clustered at the agent level.

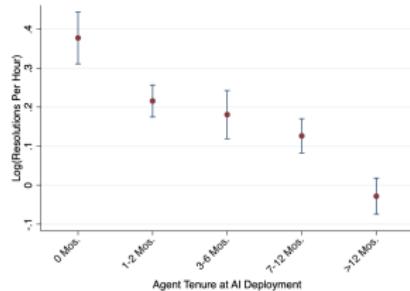
Heterogeneity by Skill

FIGURE 6: HETEROGENEITY OF AI IMPACT, BY SKILL AND TENURE

A. IMPACT OF AI ON RESOLUTIONS PER HOUR, BY SKILL AT DEPLOYMENT

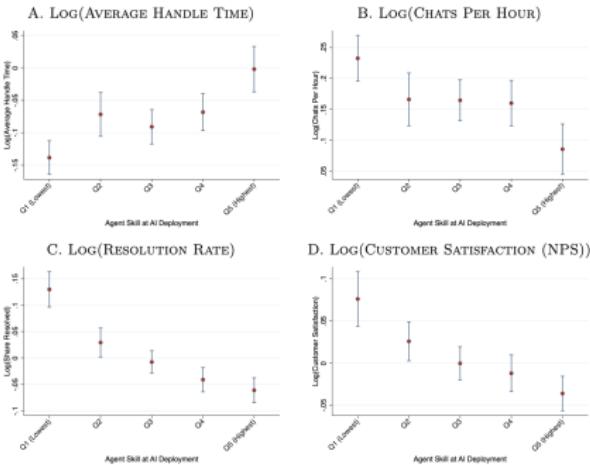


B. IMPACT OF AI ON RESOLUTIONS PER HOUR, BY TENURE AT DEPLOYMENT



Heterogeneity by Skill

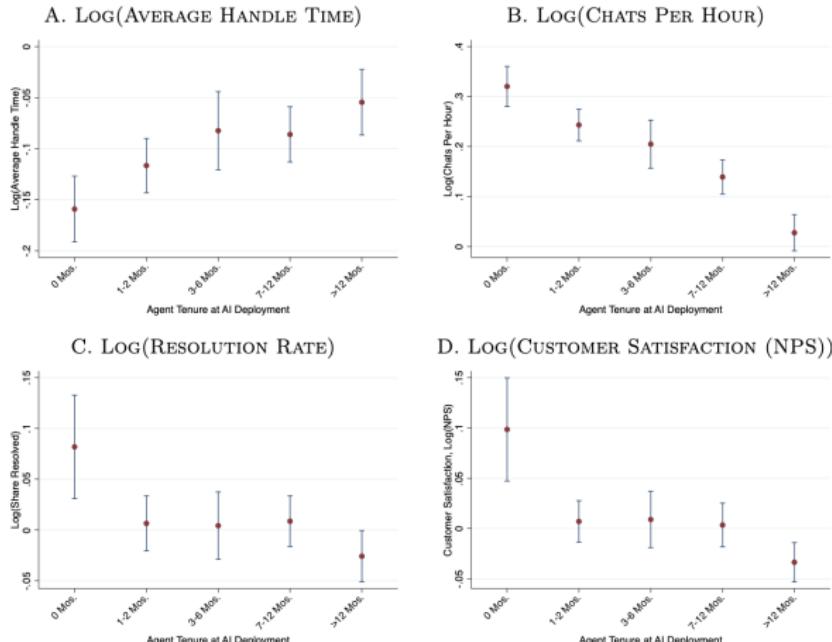
FIGURE 7: HETEROGENEITY OF AI IMPACT BY PRE-AI WORKER SKILL, ADDITIONAL OUTCOMES



NOTE: These figures plot the impacts of AI model deployment on four measures of productivity and performance, by pre-deployment worker skill. Agent skill is calculated as the agent's trailing three month average of performance on average handle time, call resolution, and customer satisfaction, the three metrics our firm uses for agent performance. Within each month and company, agents are grouped into quintiles, with the most productive agents within each firm in quintile 5 and the least productive in quintile 1. Panel A plots the average handle time or the average duration of each technical support chat. Panel B graphs chats per hour, or the number of calls an agent can handle per hour. Panel C plots the resolution rate, and Panel D plots net promoter score, an average of surveyed customer satisfaction. All specifications include agent and chat year-month, location, and company fixed effects and standard errors are clustered at the agent level.

Heterogeneity by Worker Tenure

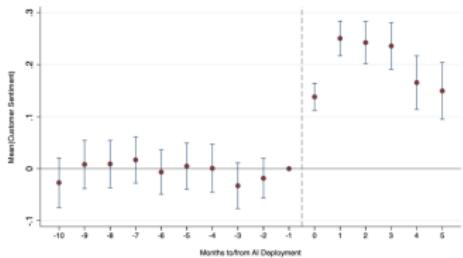
FIGURE 8: HETEROGENEITY OF AI IMPACT BY PRE-AI WORKER TENURE, ADDITIONAL OUTCOMES



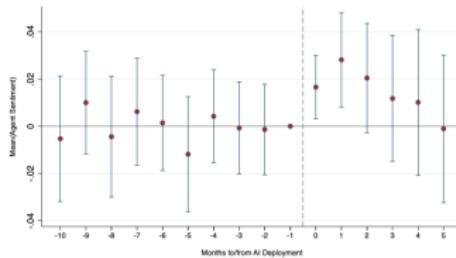
NOTES: These figures plot the impacts of AI model deployment on measures of productivity and performance by pre-AI worker tenure, defined as the number of months an agent has been employed when they receive access to the AI model. Panel A plots the average handle time or the average duration of each technical support chat. Panel B plots the number of chats per hour. Panel C plots the resolution rate. Panel D plots the customer satisfaction level, measured as the Net Promoter Score (NPS). Error bars represent standard errors.

Sentiment

C. CUSTOMER SENTIMENT, EVENT STUDY



D. AGENT SENTIMENT, EVENT STUDY

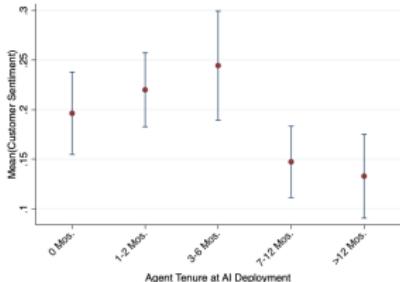


NOTES: Each panel of this figure plots the impact of AI model deployment on conversational sentiment. Panel A shows average customer sentiments. Panel B shows average agent sentiments. Panel C plots the event study of AI model deployment on customer sentiment and Panel D plots the corresponding estimate for agent sentiment. Sentiment is measured using SIEBERT, a fine-tuned checkpoint of a RoBERTA, an English language transformer model. All data come from the firm's internal software systems.

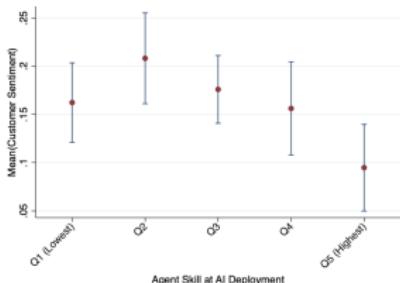
Sentiment

FIGURE A.7: HETEROGENEITY IN CUSTOMER SENTIMENT

A. BY TENURE AT AI MODEL DEPLOYMENT



B. BY PRODUCTIVITY AT AI MODEL DEPLOYMENT

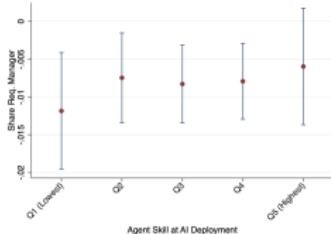


NOTES: Each panel of this figure plots the impact of AI model deployment on the mean sentiment per conversation. Sentiment refers to the emotion or attitude expressed in the text of the customer chat and ranges from -1 to 1 where -1 indicates very negative sentiment and 1 indicates very positive sentiment. Panel A plots the effects of AI model deployment on customer sentiment by agent tenure when AI deployed and Panel B plots the impacts by agent ex-ante productivity. All data come from the firm's internal software systems. Average sentiment is measured using SIEBERT, a fine-tuned checkpoint of a RoBERTa, an English language transformer model.

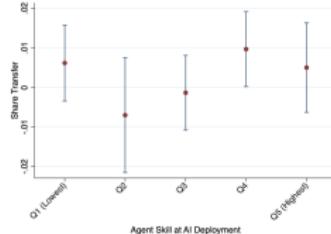
Manager Assistance

FIGURE A.8: ESCALATION AND TRANSFERS, HETEROGENEITY BY WORKER TENURE AND SKILL

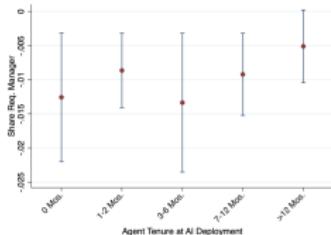
A. MANAGER ASSISTANCE, BY PRE-AI SKILL



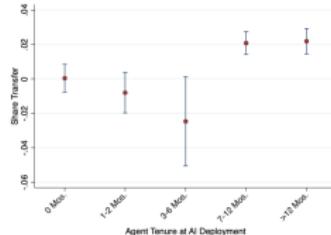
B. TRANSFERS, BY PRE-AI SKILL



C. MANAGER ASSISTANCE, BY PRE-AI TENURE



D. TRANSFERS, BY PRE-AI TENURE



NOTES: Panels A and C show the effects of AI on customer requests for manager assistance, by pre-AI agent skill and in by pre-AI agent tenure. Panels B and D show the impacts on transfers by pre-AI agent skill and pre-AI agent tenure. All robust standard errors are clustered at the agent location level. All data come from the firm's internal software systems.

Outcomes

- Across many dimensions, worker productivity rose
- And the productivity increases were higher for the least skilled workers – just like we had seen in the experiment
- They suggest that generative AI “reallocates experience” to the least experienced workers making them essentially appear as though they had been there awhile
- Findings suggest that it improves customer sentiment, reduces requests for managerial intervention, and improves employee retention
- Still unclear how generalizable this is, and what impact we should see on overall aggregate employment as this was AI assisted, not AI alone

Roadmap

Introduction

Managing expectations

Introducing difference-in-differences

Potential outcomes

Estimation

Inference

Parallel trends violations

How parallel trends can get violated

DiD in Court

Falsifications

Interesting Applications using staggered adoption

Facebook and Mental Health

Generative AI Randomized Experiment

ChatBot Rollout

Conclusion

DiD vs ATT

- We learned that difference-in-differences was just four averages and three subtractions
- But it was also a specific regression specification
- We saw that difference-in-differences could be used to estimate average treatment effects
- But the DiD equation is distinct from the ATT parameter we care about

Parallel Trends

- DiD only was equal to the ATT if the parallel trends assumption was true
- But it's not verifiable so it's a difficult assumption
- Parallel trends is not something a statistical model fixes – it's something a control group fixes
- Some comparison groups will satisfy parallel trends, but some won't

Evidence for parallel trends

1. Event study graphics – plot coefficients and confidence intervals to check if pre-trends are zero so that post-treatment coefficients are compelling evidence for causal effects
2. Falsifications – if possible, rule out competing hypotheses using falsifications (e.g., 65+ year olds can't be on Medicaid as they're already on Medicare)

Roadmap

1. Show bite – first order effects
2. Main results – What's your study about?
3. Event study graphs – This will be your main results and your evidence of parallel trends keeping in mind pre-trends and parallel trends are technically distinct
4. Falsifications – If you can find falsifications, use them
5. Mechanisms – can you find any explanation?

Synthetic control

- But what if parallel trends really isn't realistic – what then?
- Then you may need to create your own control group that follows the same approximately trajectory as your treatment group pre-treatment
- A method by Abadie and Gardazebal (2003) and follow up papers worked out a method for this called synthetic control
- If we have time, we will go through this together