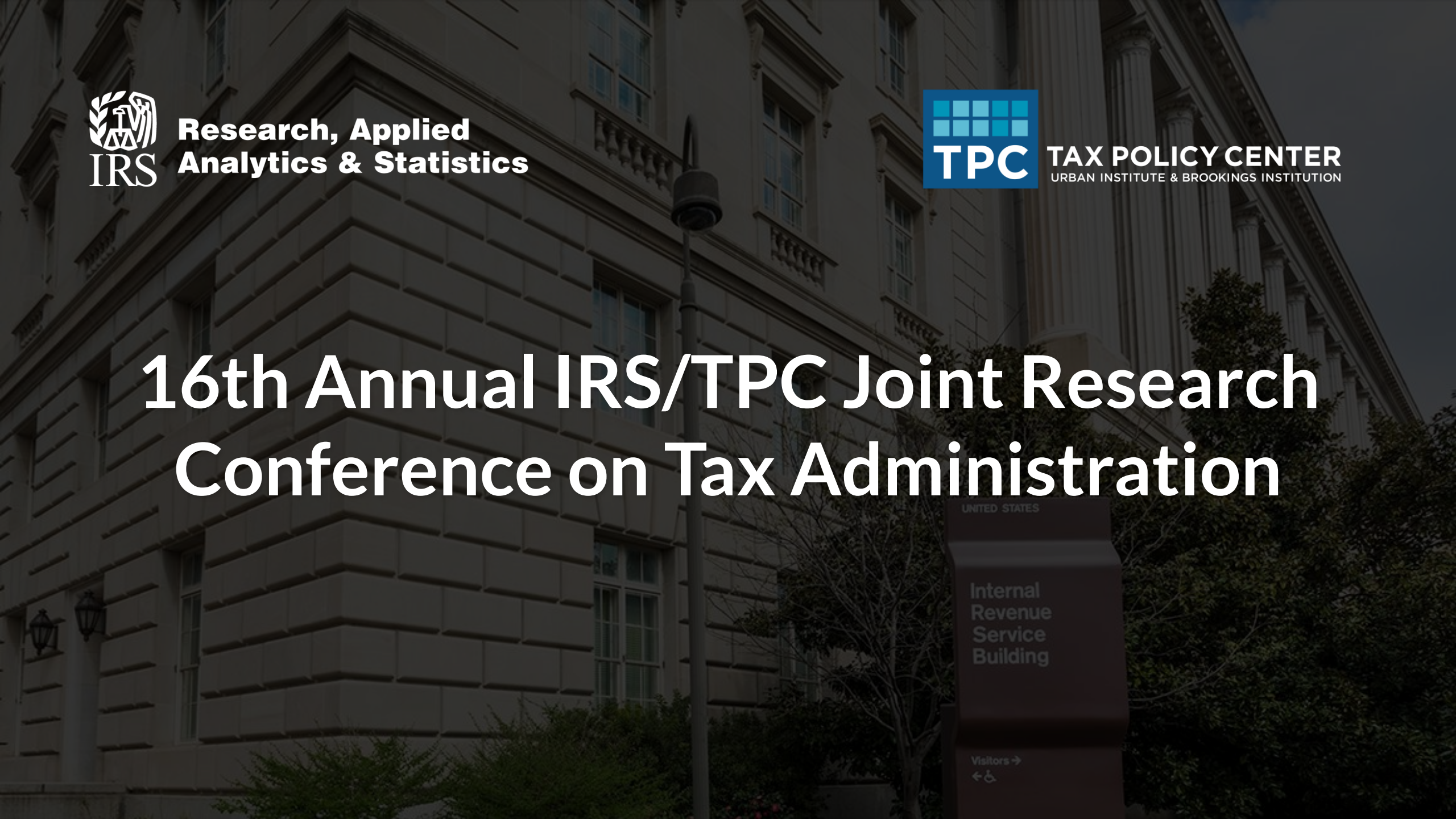


Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

16th Annual IRS/TPC Joint Research Conference on Tax Administration

IRS Research, Applied Analytics & Statistics

TPC TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

Session 3

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

Experimental Evidence on the Specific Indirect Effects of Audits

June 2026

IRS – RAAS: Tom Hertz

MITRE: Miguel Sarzosa, Damon Frezza, Justin Nave

NOTICE

This (software/technical data) was produced for the U. S. Government under Contract Number TIRNO-99-D-00005, and is subject to Federal Acquisition Regulation Clause 52.227-14, Rights in Data—General, Alt. I, II, III and IV (MAY 2014) [Reference 27.409(a)].

No other use other than that granted to the U. S. Government, or to those acting on behalf of the U. S. Government under that Clause is authorized without the express written permission of The MITRE Corporation.

For further information, please contact The MITRE Corporation, Contracts Management Office, 7515 Colshire Drive, McLean, VA 22102-7539, (703) 983-6000.

Introduction

- We assess the **specific indirect effect** of auditing **Self Employed taxpayers** that file a Schedule C with large deductions on certain expenses.
- We identify causal effects using two sources of variation:

Random allocation of taxpayers to a holdout sample that prevents them from being audited in subsequent years

IRS' auditing capacity constraints + case selection based on priority metrics → **discontinuities in the probability of treatment**

- Multiple populations and thresholds → **explore treatment effects heterogeneity**
  - ➢ Characterize populations for which indirect effect of audits is larger

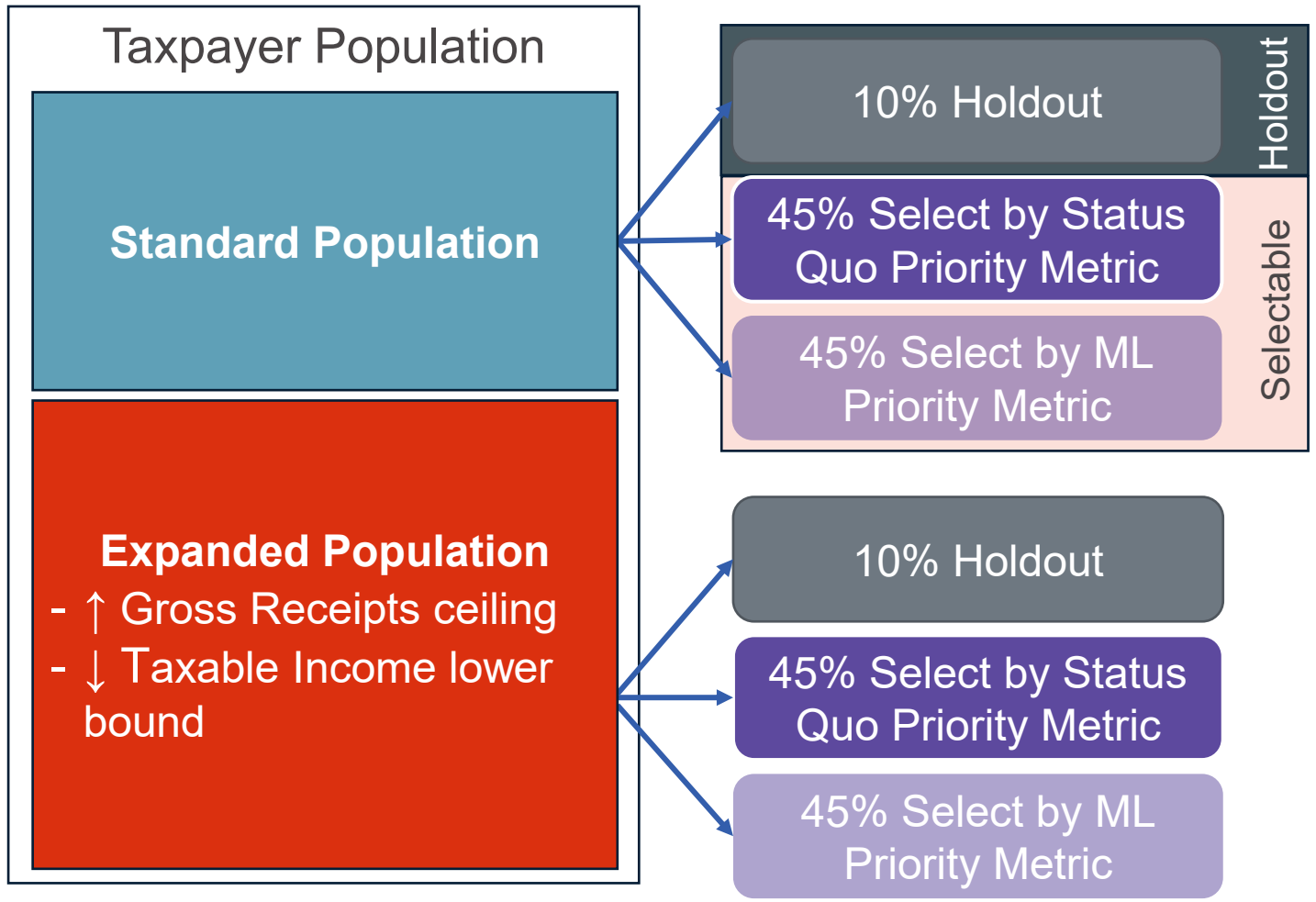
Data

- We use Form 1040, Schedule C, and Exam Enforcement microdata for a population selected based on the project code specific filter criteria and tax law exclusion criteria
  - Schedule C and Exam Enforcement Criteria are collected for three catchment years: 2017, 2018, and 2019
  - Form 1040 longitudinal data is collected for all the tax years between 2011 and 2024
- Much of the empirical strategies we use rely on the calculation of priority scores, we produce using feature engineering and XGboost modeling techniques to provide estimates of expected non-compliance
- The IRS then performs final filtering and prioritization based on the provided estimates

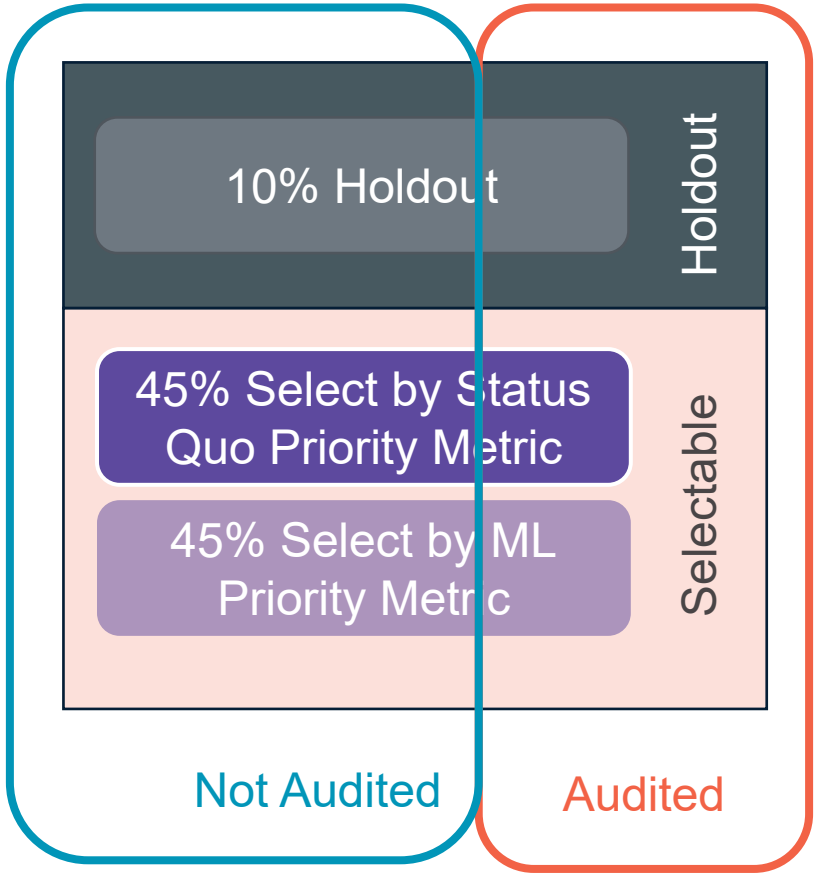
Experimental Design

The experimental design includes **three main** components:

1. Expansion to the auditable population (not the focus of presentation)
2. New **priority metric**
   - Sum of examined expenses: Used historically, Status Quo
   - ML: New metric tested under this experiment.
3. ***Random assignment to Holdout*** of would-be-audited population



Eligibility Vs. Treatment



(not to scale)

- Some holdout TINs are audited (e.g., referrals).
- Not all selectable TINs were audited (capacity constraints).

Identification

- Use of randomized eligibility to obtain causal estimates
- Priority-based selection + capacity constraints = Discontinuities in probability of treatment.

2017 Experiment (holdout vs. selectable) did not include Extended Population and did not stratify by priority metric.

Empirical Strategy

1. DiD with Eligibility as Instrument for Audit

- Being audited is *not random*, but $\Pr(Audit|\text{Selectable}) \gg \Pr(Audit|\text{Holdout})$
- **Instrument audit with eligibility** in a longitudinal context. Let $D_i = \mathbf{1}[Audit_i = 1]$

$$D_i = \pi_0 + \pi_1 Sel_i + \gamma X_i + u_i$$

$$Y_{it} = \beta \text{Post}_t + \theta \widehat{D}_i + \tau \text{Post}_t \widehat{D}_i + \lambda X_i + \mu_t + \varepsilon_{it}$$

2. Regression Discontinuity – DiD

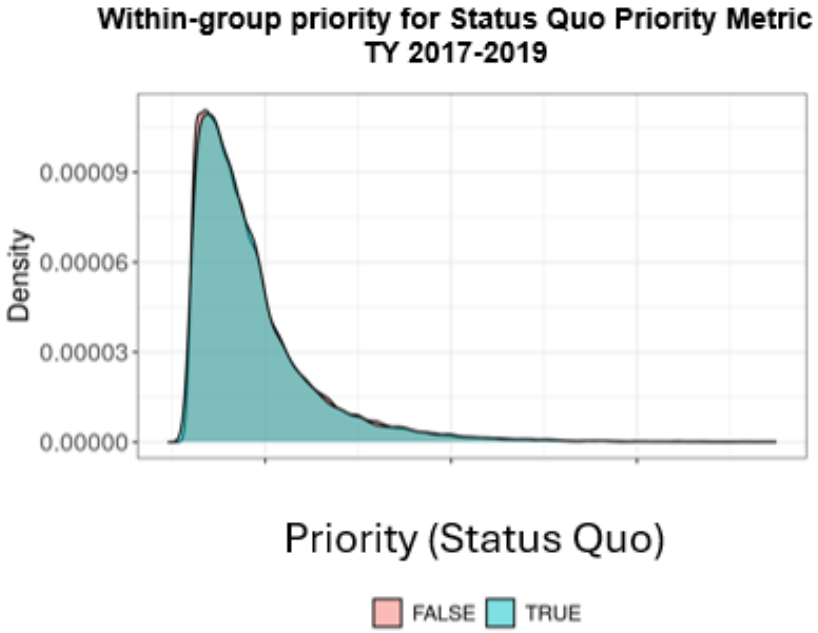
- Selection into treatment closely follows priority metrics *PM* (sum, ML)
- Capacity constraints create **jumps** in $\Pr(Audit|PM)$
- Use an RD approach in longitudinal context: i s.t. $PM_i \in (\text{cutoff} - \text{BW}, \text{cutoff} + \text{BW})$

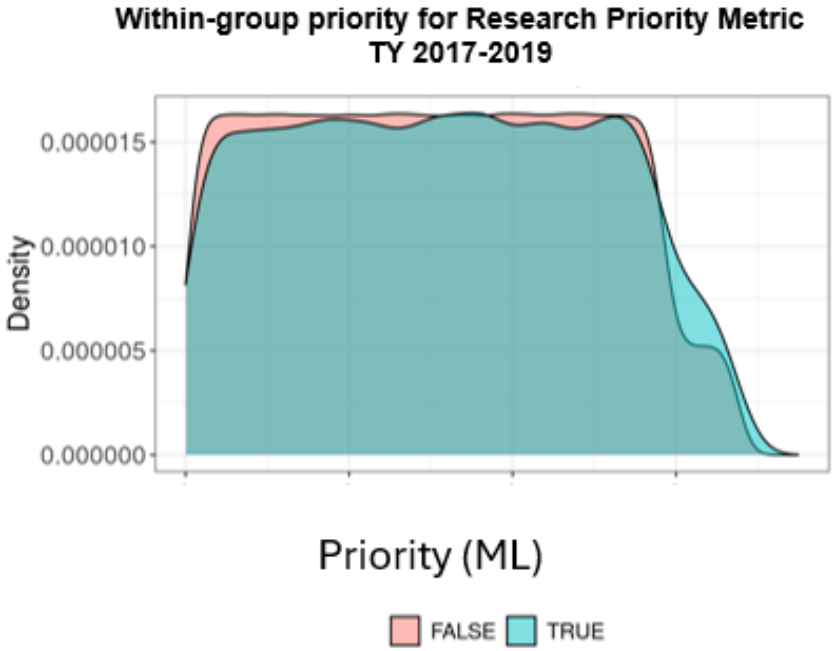
$$Y_{it} = \beta \text{Post}_t + \gamma \mathbf{1}\left[\text{PM}_i \geq \widetilde{\text{PM}}\right] + \tau \text{Post}_t \mathbf{1}\left[\text{PM}_i \geq \widetilde{\text{PM}}\right] + f_0(\text{PM}_i) + f_1(\text{PM}_i) + \mu_t + \varepsilon_{it}$$

MITRE

Balance

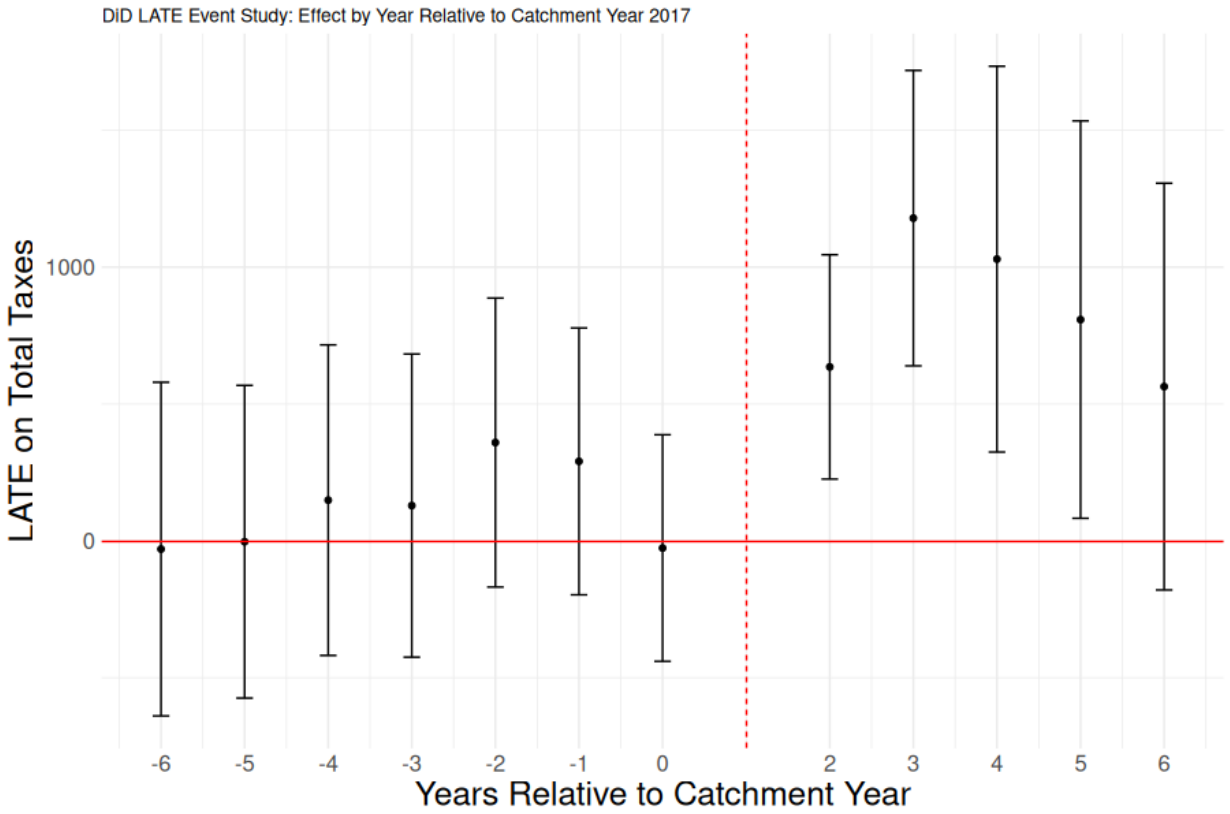
Variable	Estimate	Std. Error	t value	Pr(>|t|)
Priority Score (Research Sample)	-255.442	473.626	-0.539332	0.58966
Priority Score (Status Quo)	-20.6219	69.2648	-0.297725	0.76591
Total Tax (2017)	21.3634	99.1527	0.21546	0.8649
Total Tax (2018)	-5.58836	44.6008	-0.125297	0.92065
Total Tax (2019)	-14.3287	4.11926	-3.47846	0.17821



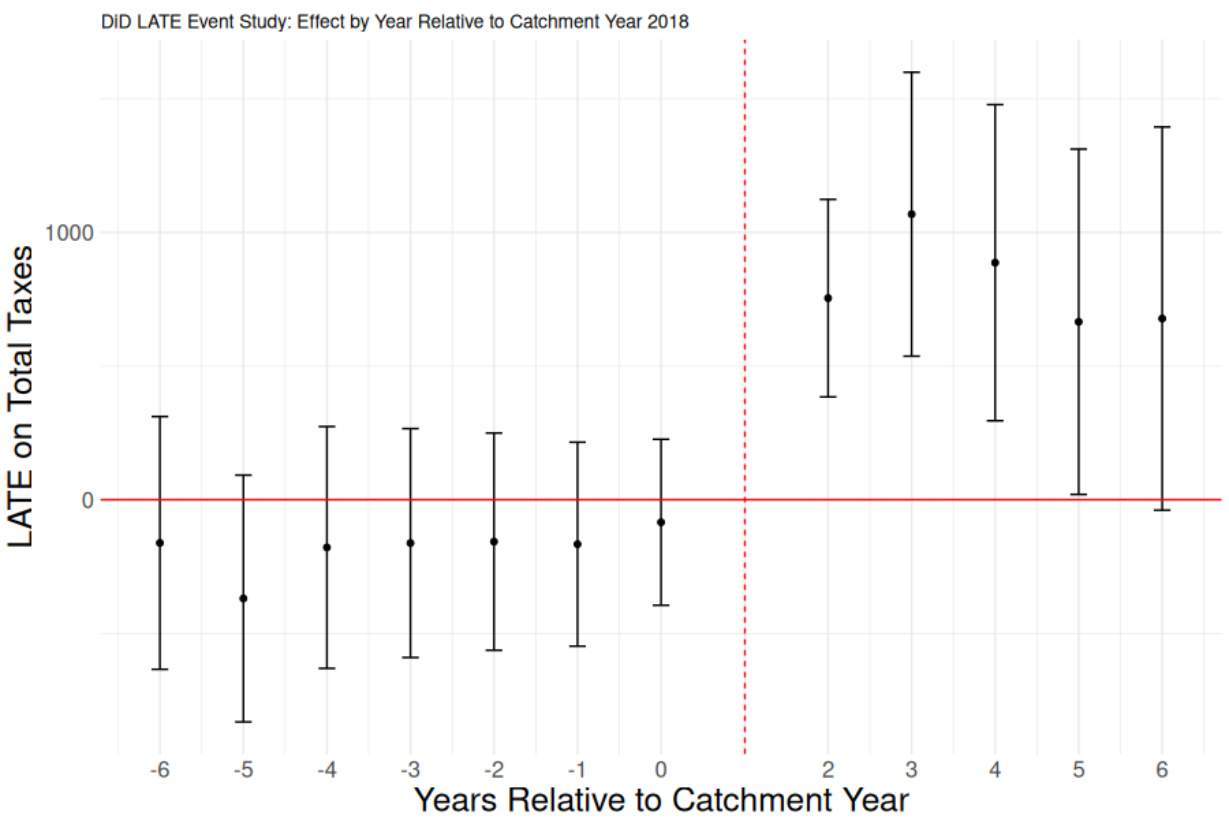


DiD LATE on Yearly Total Taxes by Catchment Year

2017



2018



Intent-to-treat (ITT) estimates show pre-treatment balance and significant post-treatment differences between selectable and holdout samples

Aggregate Indirect Effects Results, CY 2017

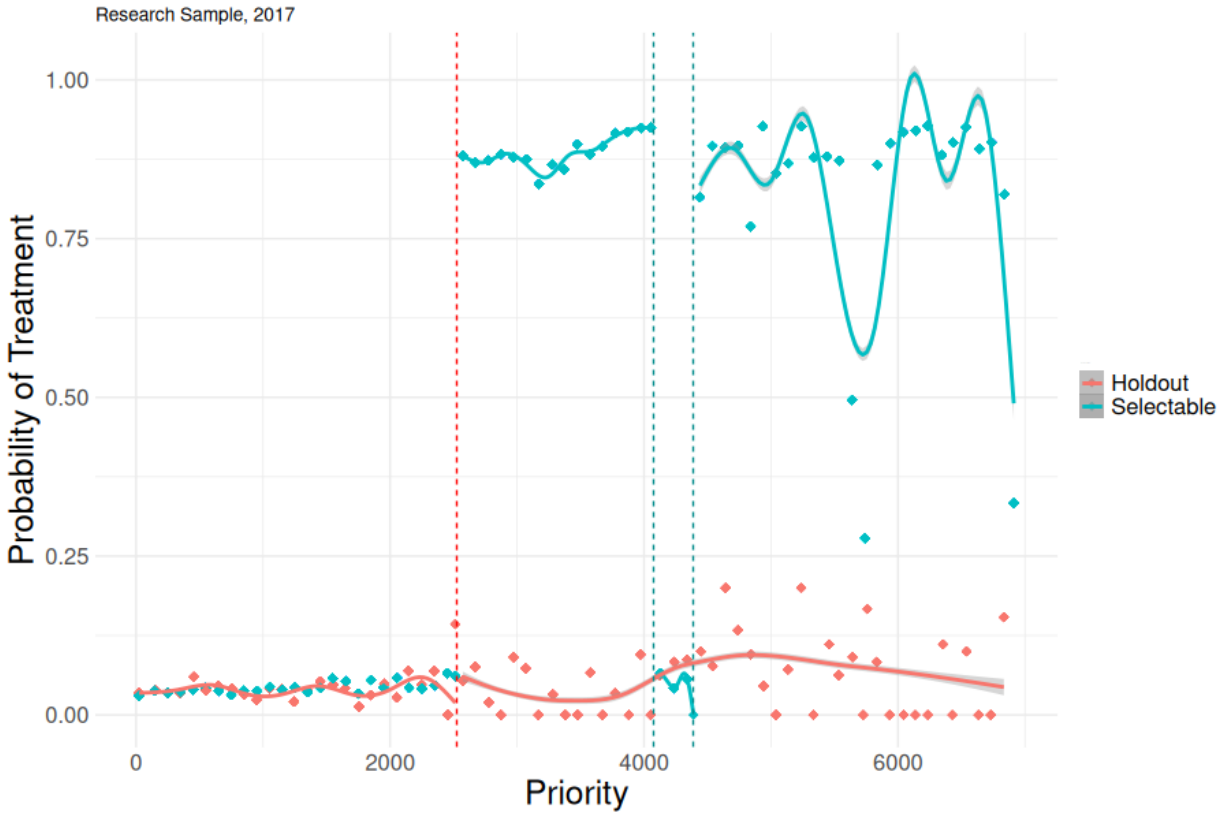
Dependent Var.:	Total Tax		
	Whole Sample		
	OLS	IV	Matching
Audited x post = 1	1,784.7***	726.9***	677.95***
	(45.1)	(263.8)	(169.45)
priority.metric FE	Yes	Yes	No
event_time FE	Yes	Yes	Yes
Observations	1,334,163	1,334,163	348,514
R2	0.04468	0.03069	0.06954

Note: Standard errors clustered at taxpayer level in parentheses. Signif. codes: 0 '***' 0.01 '**' 0.05 '*' 0.1

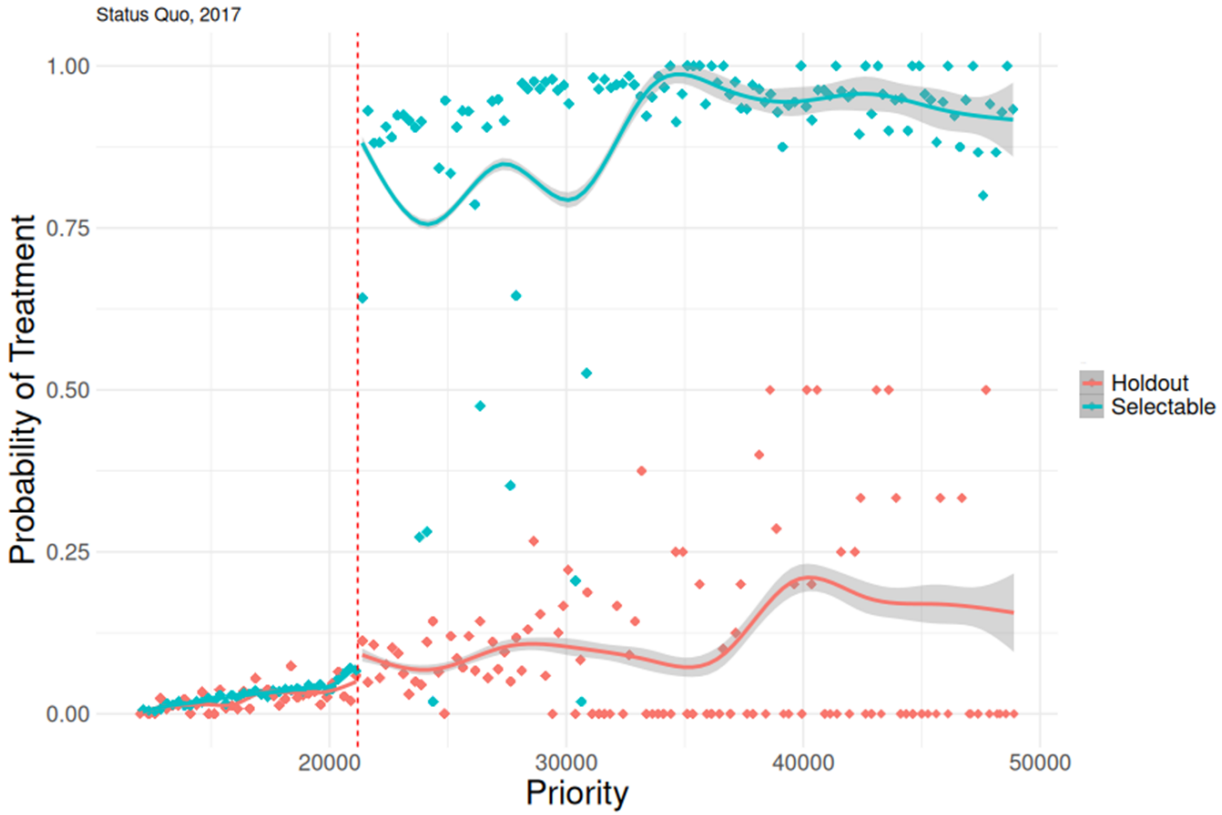
RD-DiD 2017

Pr(Audit | Priority), 2017

Research Sample



Status Quo Sample



!

Targeting based on priority scores and capacity constraints create significant jumps in Pr(Audit) among the selectable group.

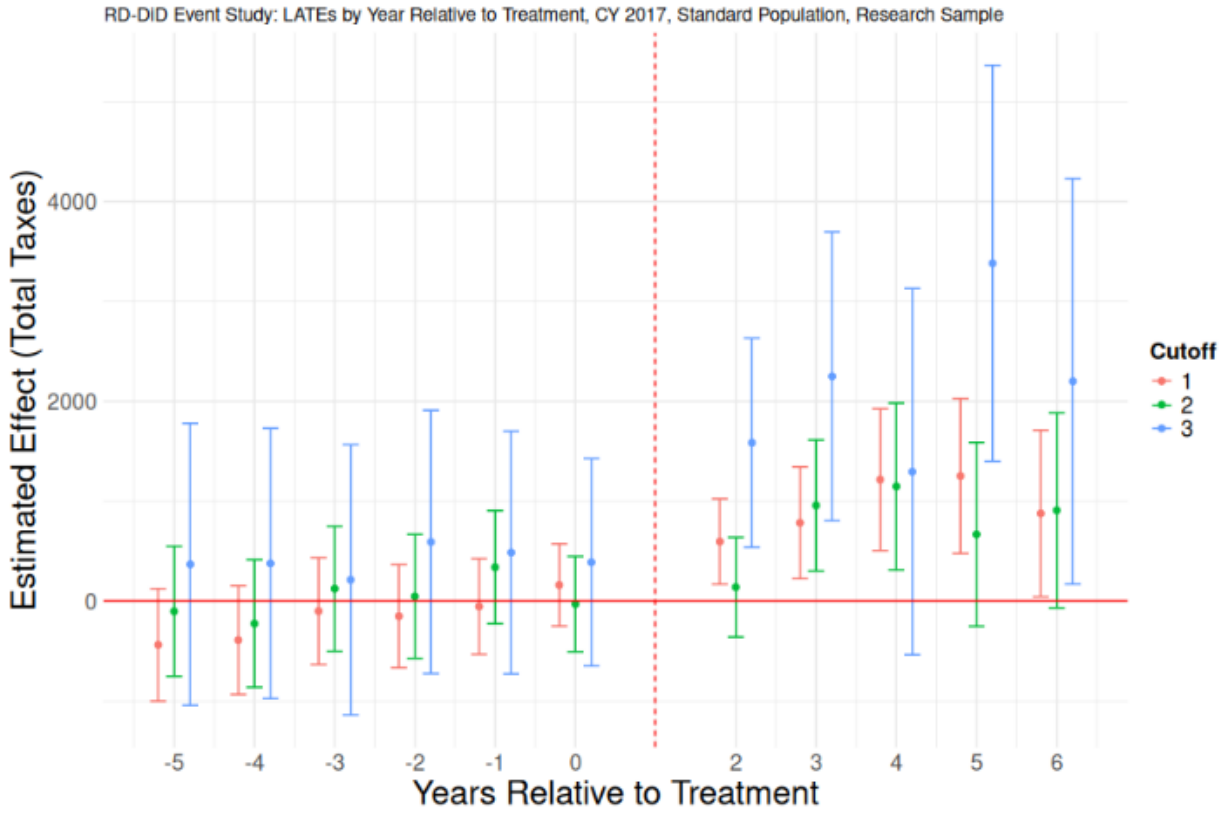
2017 LATEs by Sample and Cutoff

	Status Quo	Research		
Total Tax		Cutoff 1	Cutoff 2	Cutoff 3
LATE	1427.620***	1129.873***	686.174**	1749.050**
	(257.191)	(288.938)	(345.372)	(712.439)
Observations	50,602	41,611	21,213	6,804

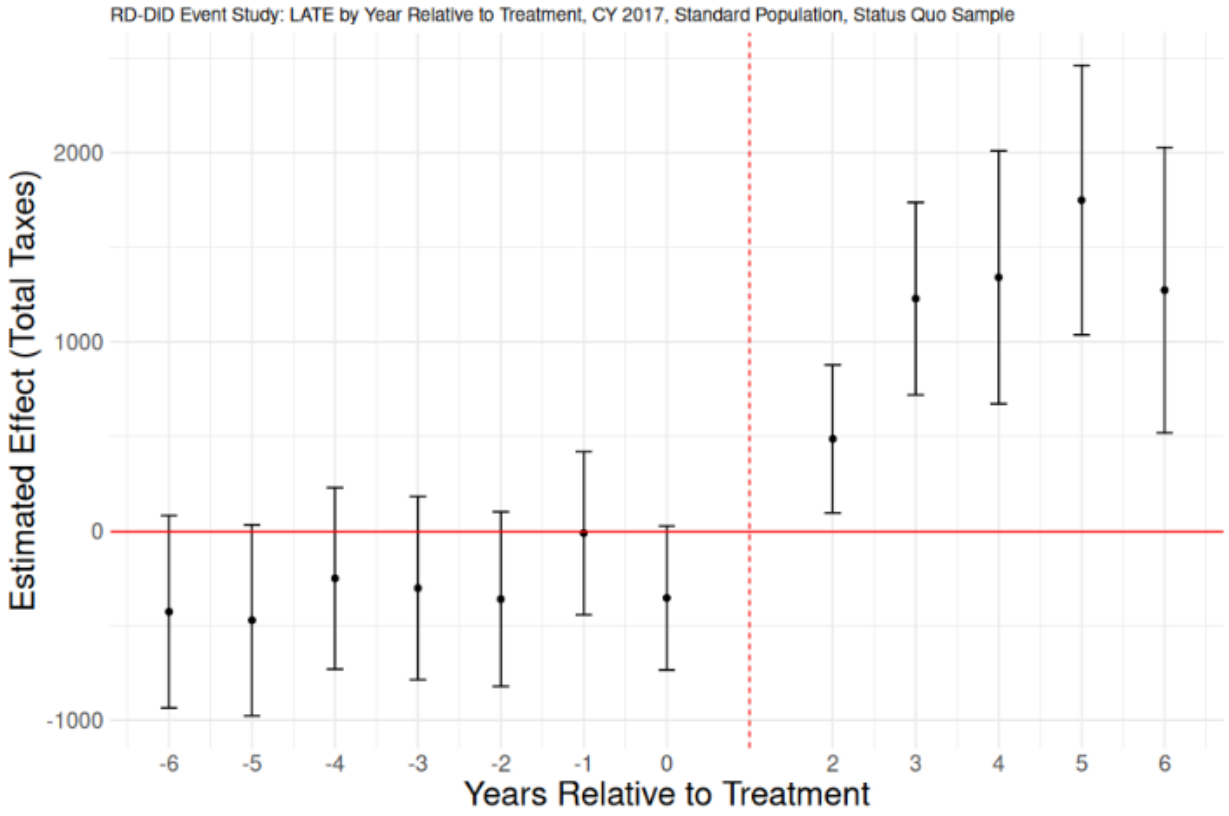
Note: All regressions include two quadratic functions of the distance to each threshold, one for the negative side and another one on the positive range. Sample limited to cases in the Standard Population that fall within the RD bandwidth. Standard errors clustered at taxpayer level in parentheses. $* p < 0.1$, $** p < 0.5$, $*** p < 0.01$

RD-DiD LATE on Yearly Total Taxes by Sample, 2017

Research Sample



Status Quo Sample



Conclusions

- We build an experiment that randomly assigns taxpayers to a holdout sample that prevents them from being audited.
- We use experimental variation to provide causal estimates of the indirect effect of audits
- We complement the experiment with effect identification via discontinuities in the probability of audit along priority metrics
- Audits cause taxpayers to report about $1000 extra in taxes each year, for the first four years after the audit.
- The effect is remarkably consistent across the Standard population

Appendix

Aggregate Indirect Effects Results, CY 2018

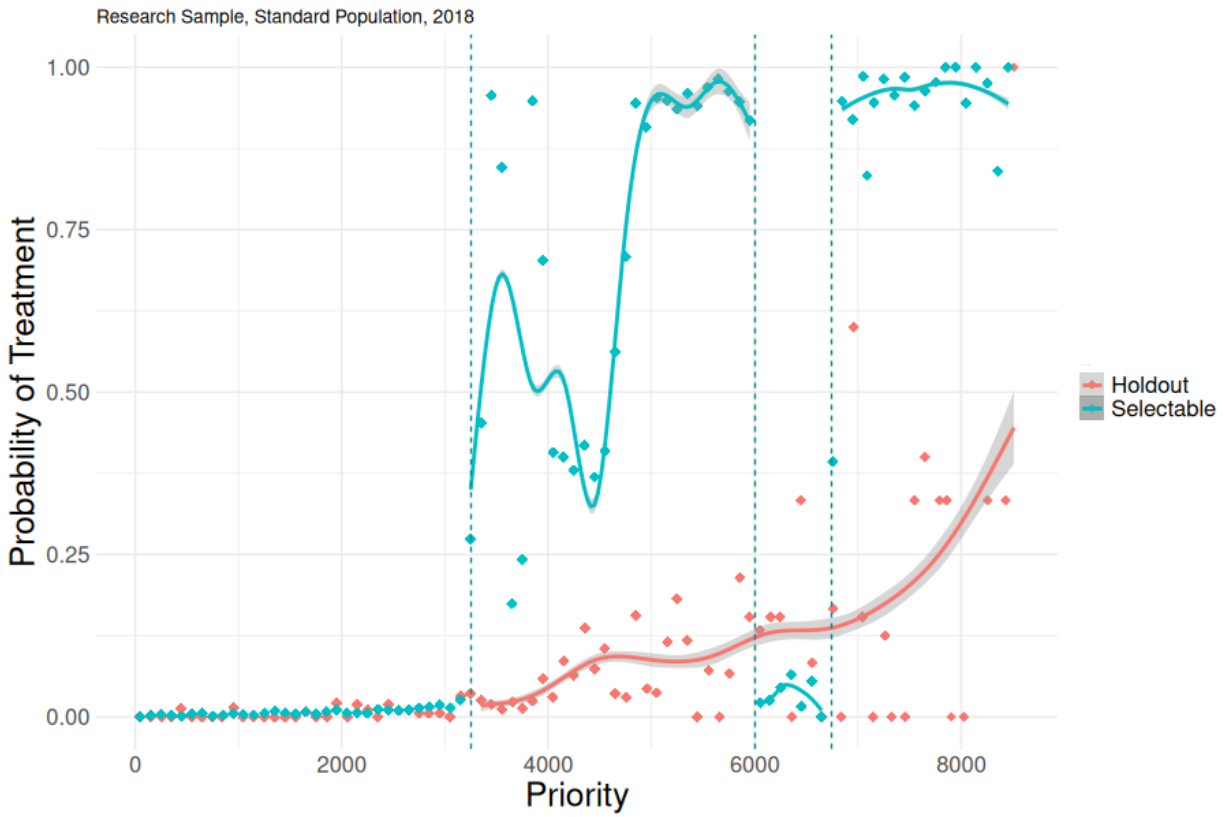
Dependent Var.:	Total Tax							
	Whole Sample		Research Sample			Status Quo Sample		
	OLS	IV	OLS	Matching	IV	OLS	Matching	IV
Audited x post = 1	1,823.6***	1,140.2***	2,171.7***	2,072.4***	1,173.8***	1,430.7***	1,277.1***	1,094.8***
	(46.6)	(366.2)	(65.1)	(153.5)	(457.2)	(66.3)	(179.5)	(590.5)
priority.metric FE	Yes	Yes	Yes	No	No	Yes	No	No
event_time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,581,493	1,581,493	773,817	158,063	773,817	807,676	123,816	807,676
R2	0.07546	0.06575	0.09034	0.11096	0.06959	0.06603	0.09746	0.06327

Note: Standard errors clustered at taxpayer level in parentheses. Signif. codes: 0 '***' 0.01 '**' 0.05 '*' 0.1

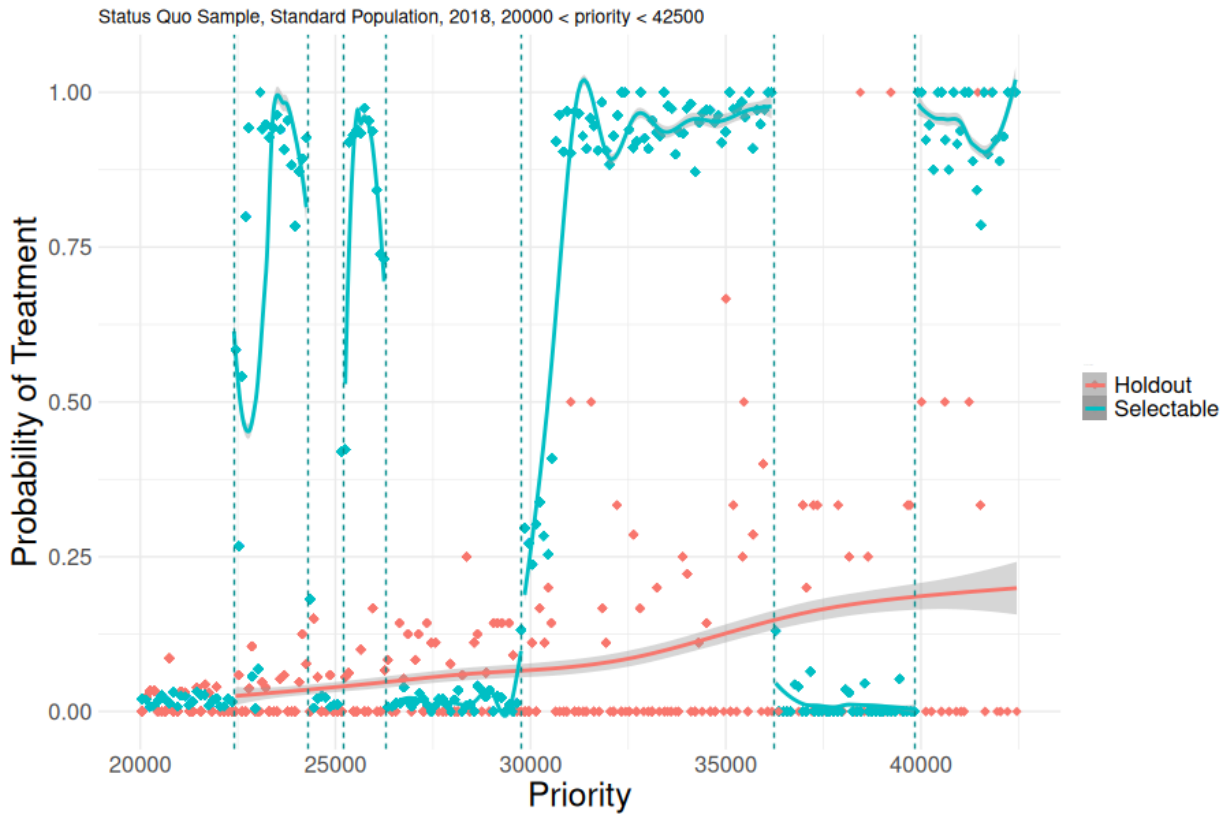
2018: RD-DiD

Pr(Audit | Priority), Selectable, 2018

Research Sample



Status Quo Sample



!

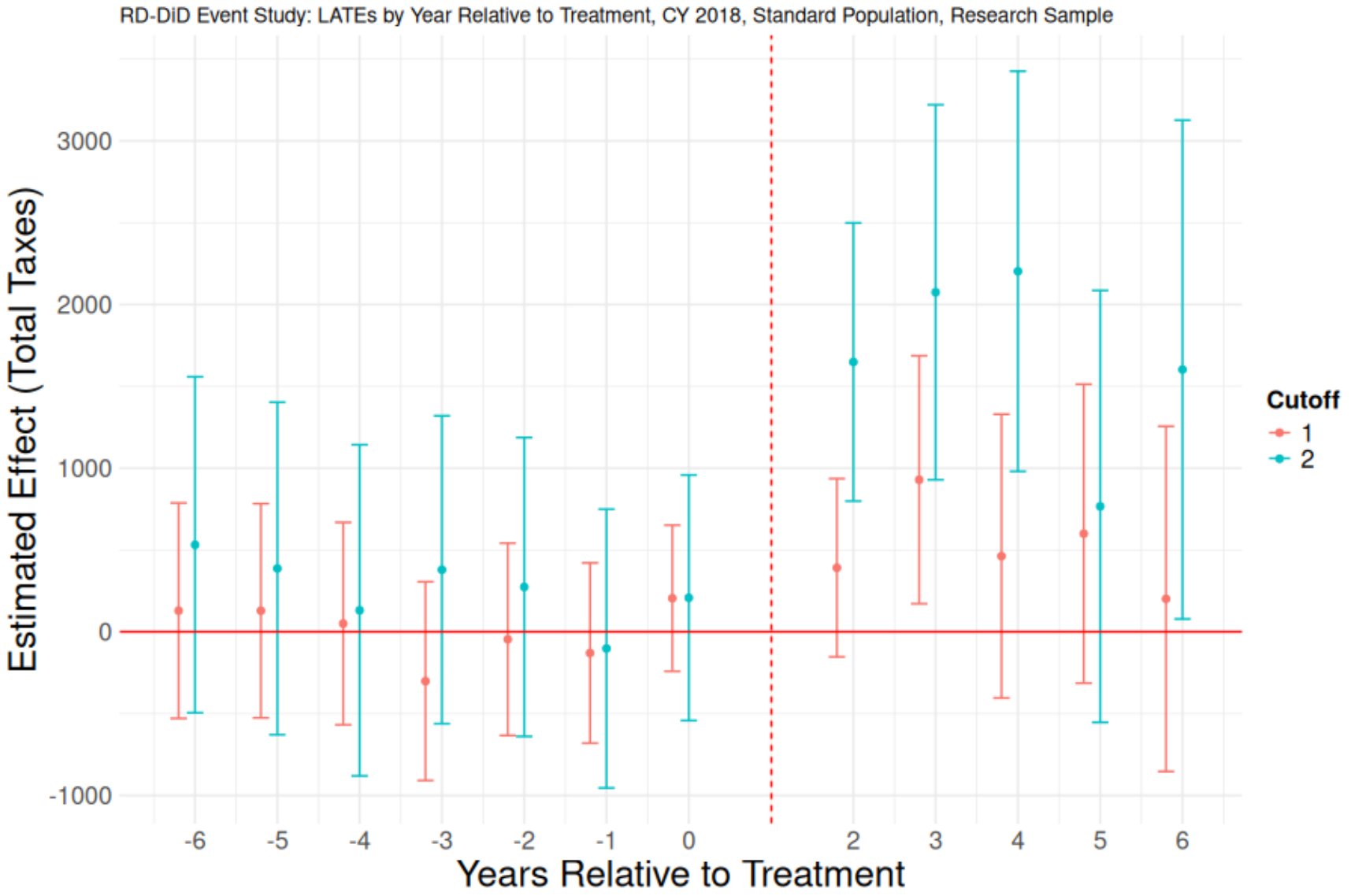
- Significant jumps in Pr(Audit) among the selectable group in Standard and Expanded populations.
- Odd selection into audit choices yield multiple discontinuities in Pr(Audit) in the Standard-Status Quo sample

2018 LATEs by Sample and Cutoff

	Status Quo					Research	
	Cutoff 1	Cutoff 2	Cutoff 3	Cutoff 5	Cutoff 6	Cutoff 1	Cutoff 2
LATE	904.923*	709.795*	989.965**	1054.76**	1508.37***	522.730	1391.47***
	(463.50)	(425.82)	(455.30)	(530.13)	(579.48)	(341.57)	(498.74)
Obs.	7092	9514	11692	5430	5055	35926	9238

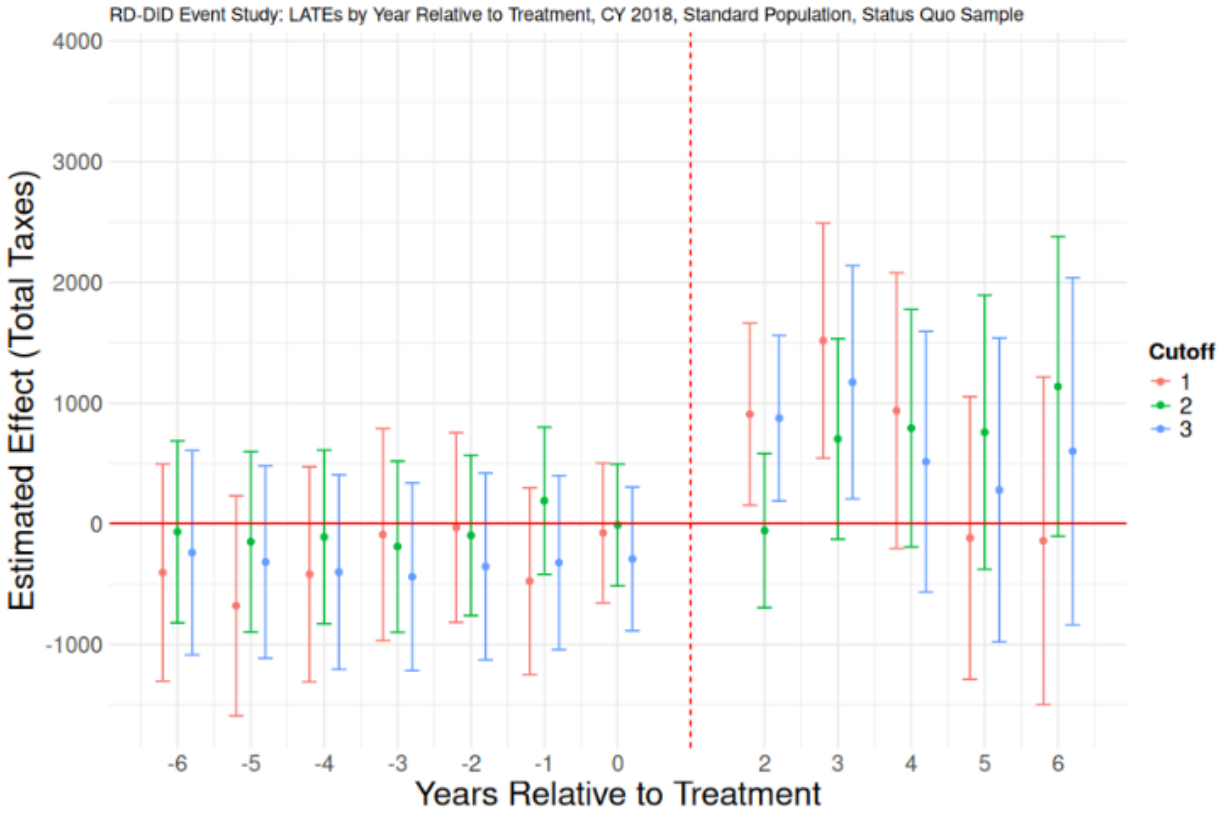
Note: All regressions include two quadratic functions of the distance to each threshold, one for the negative side and another one on the positive range. Sample limited to cases in the Standard Population that fall within the RD bandwidth. Standard errors clustered at taxpayer level in parentheses. $* p < 0.1$, $** p < 0.5$, $*** p < 0.01$

RD-DiD LATE, 2018 Research Sample

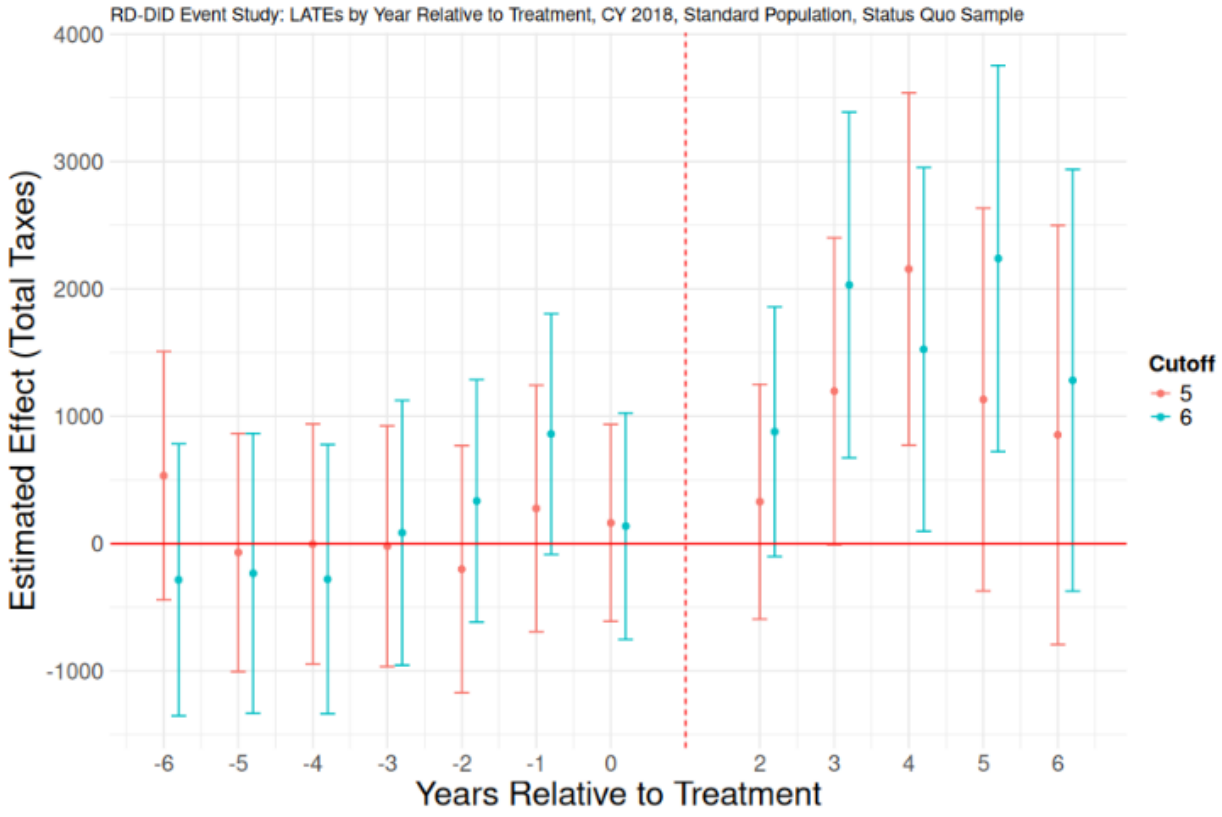


RD-DiD LATE, 2018 Status Quo Sample

Cutoffs 1 to 3



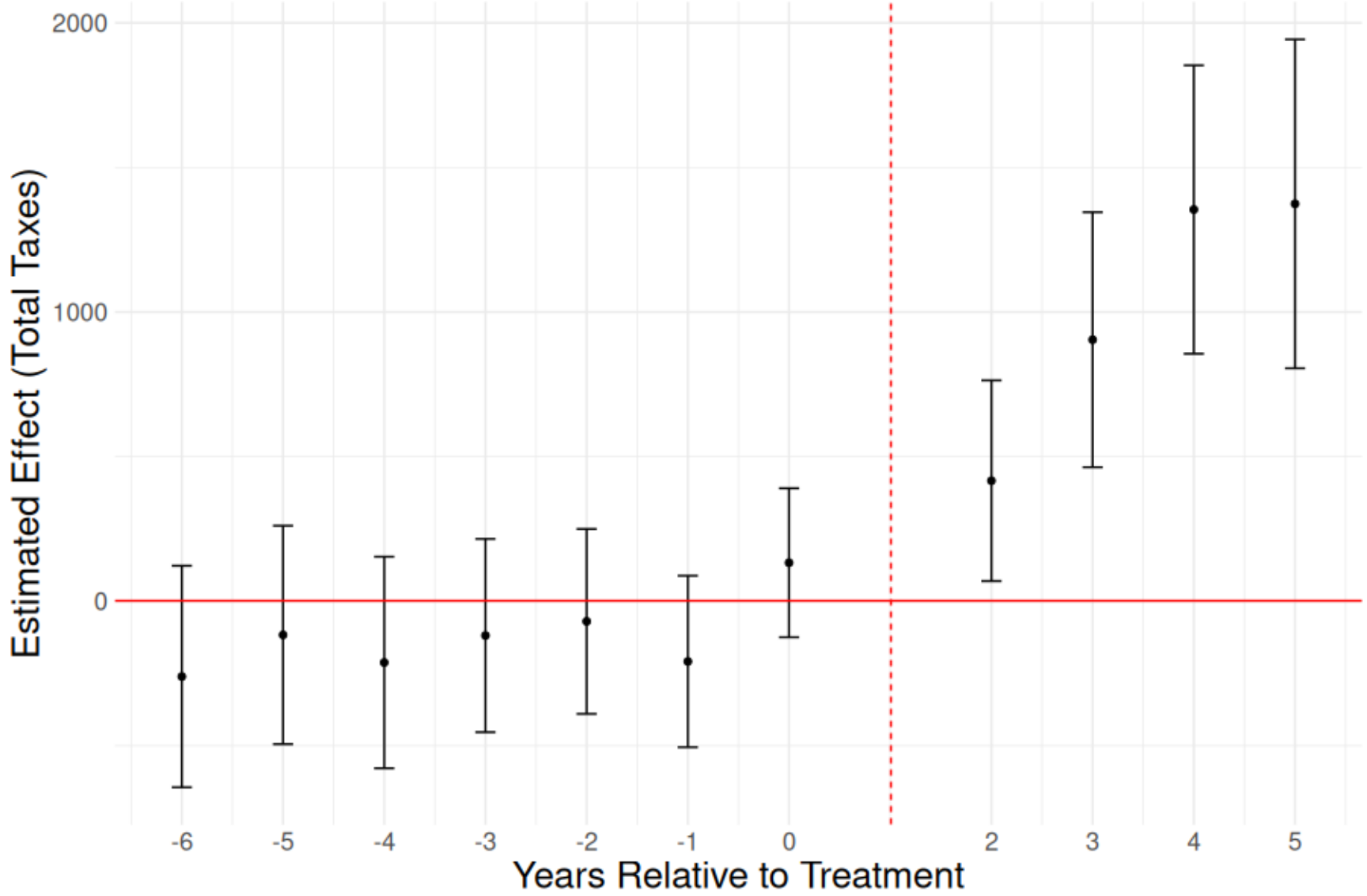
Cutoffs 5 to 6



2019: RD-DiD

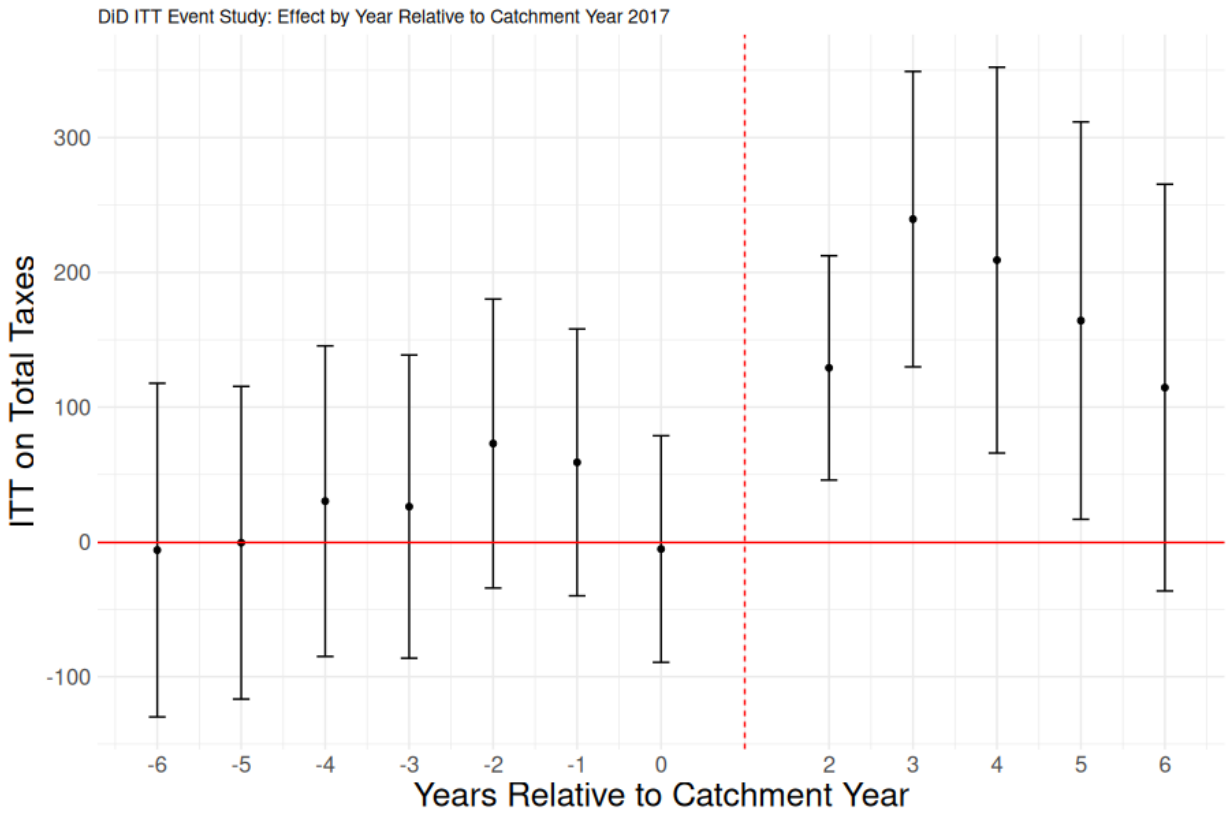
Pr(Audit | Priority), Research Sample, 2019

RD-DiD, ITT Research Sample, Selectable, Standard Population 2019

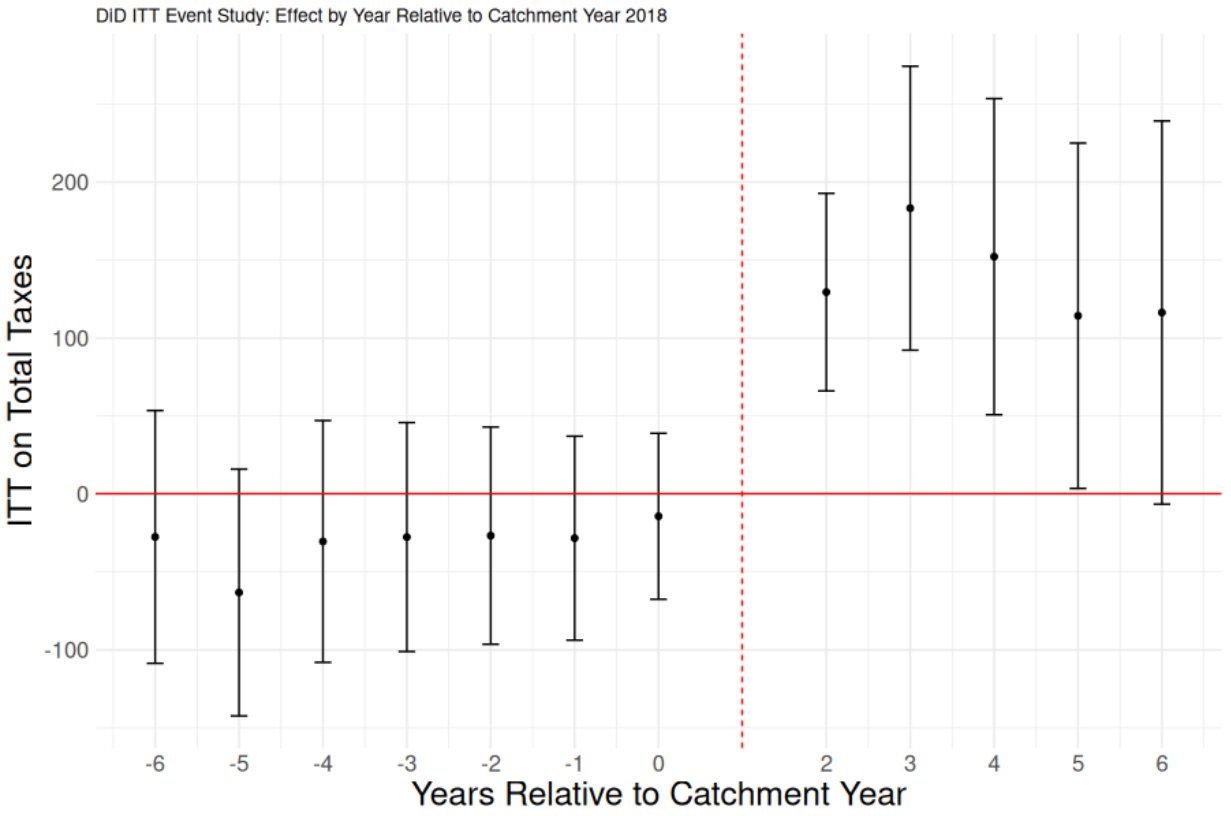


ITT on Yearly Total Taxes by Catchment Year

2017



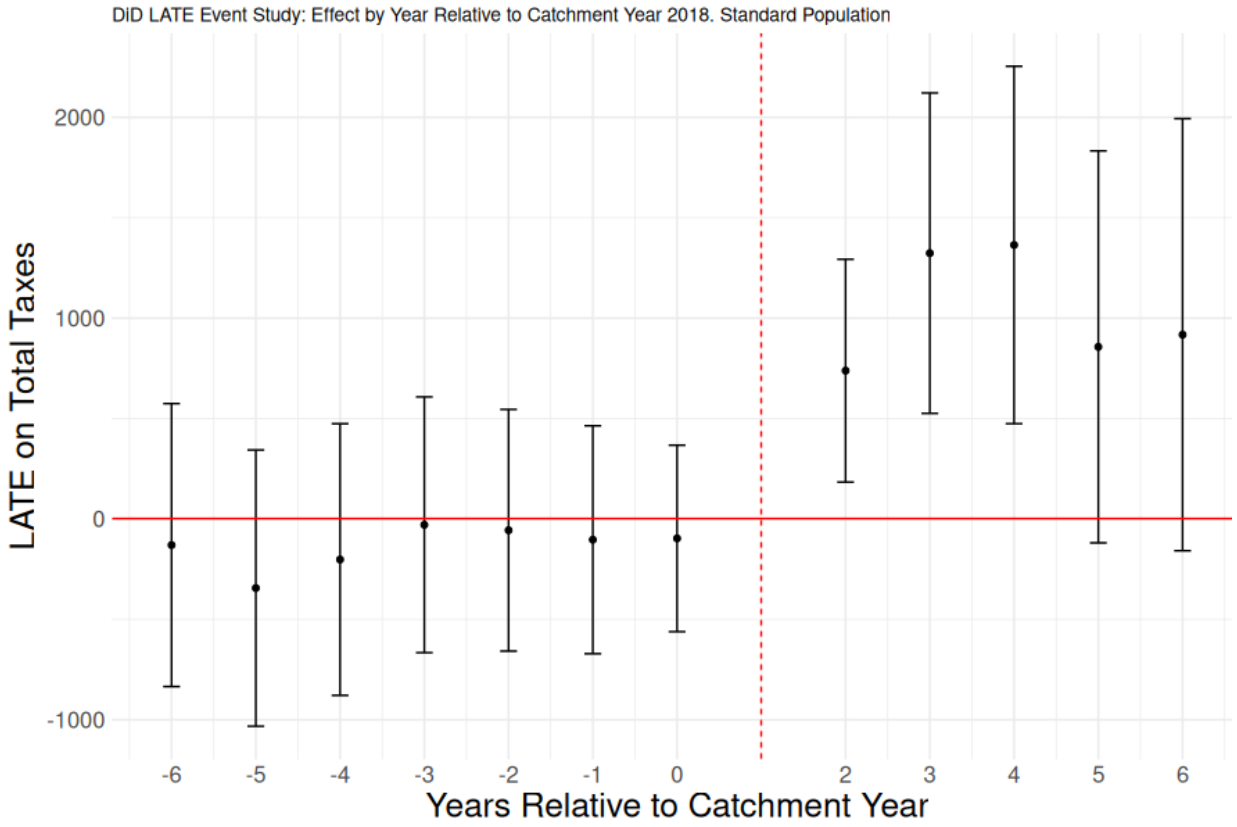
2018



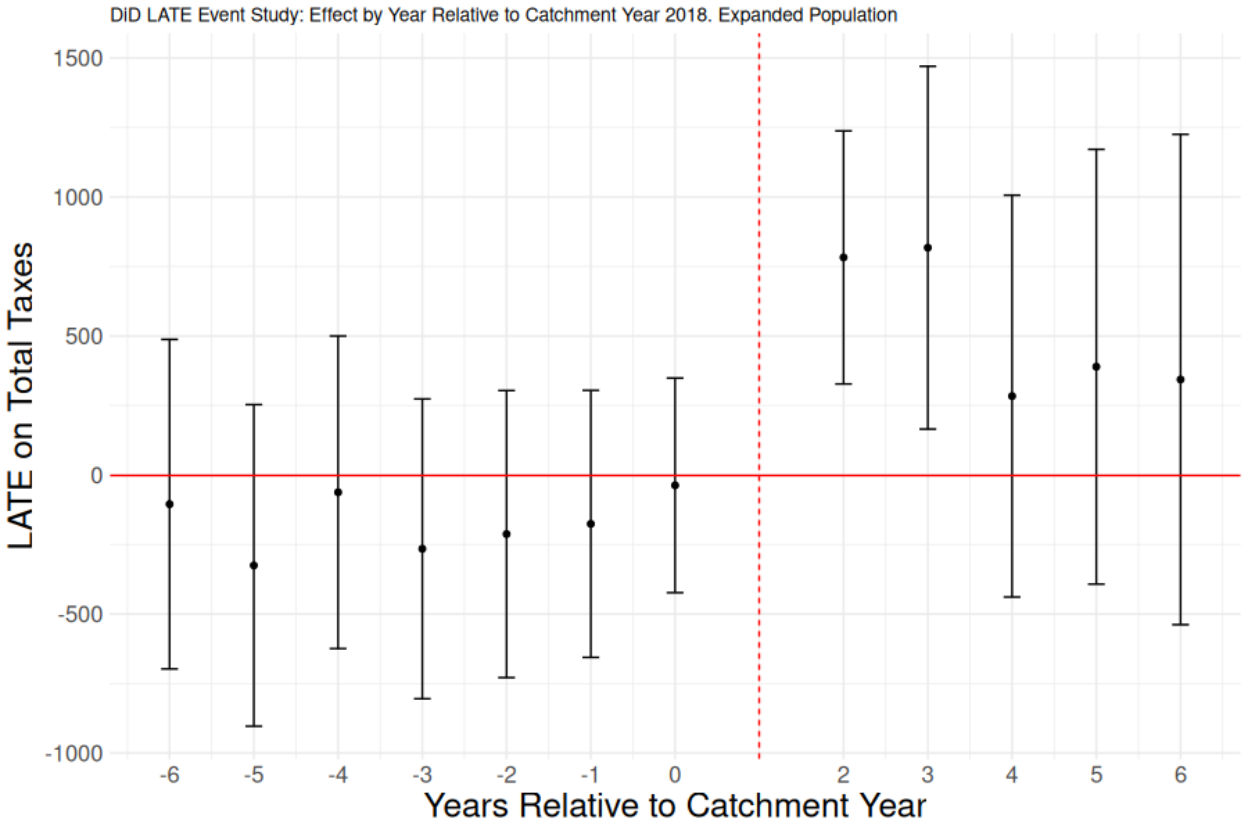
Intent-to-treat (ITT) estimates show pre-treatment balance and significant post-treatment differences between selectable and holdout samples

LATE Catchment Year 2018 by Population Type

Standard

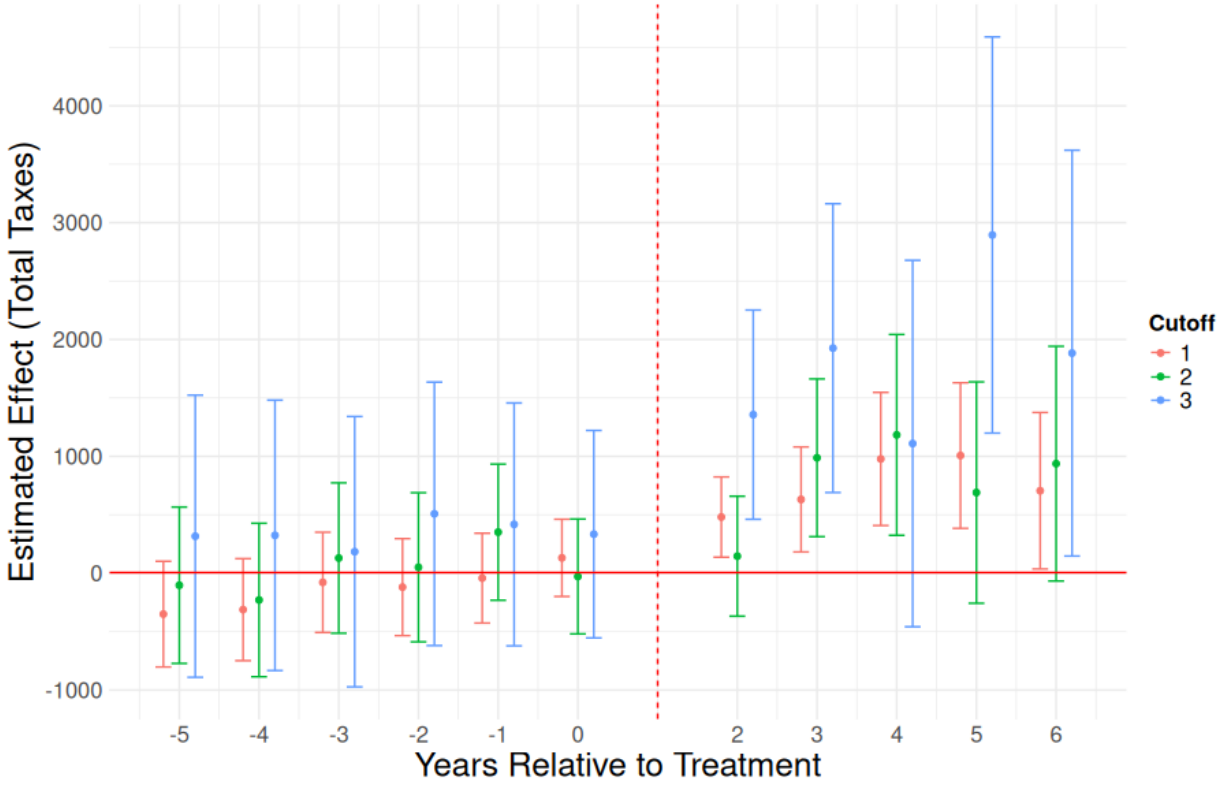


Expanded

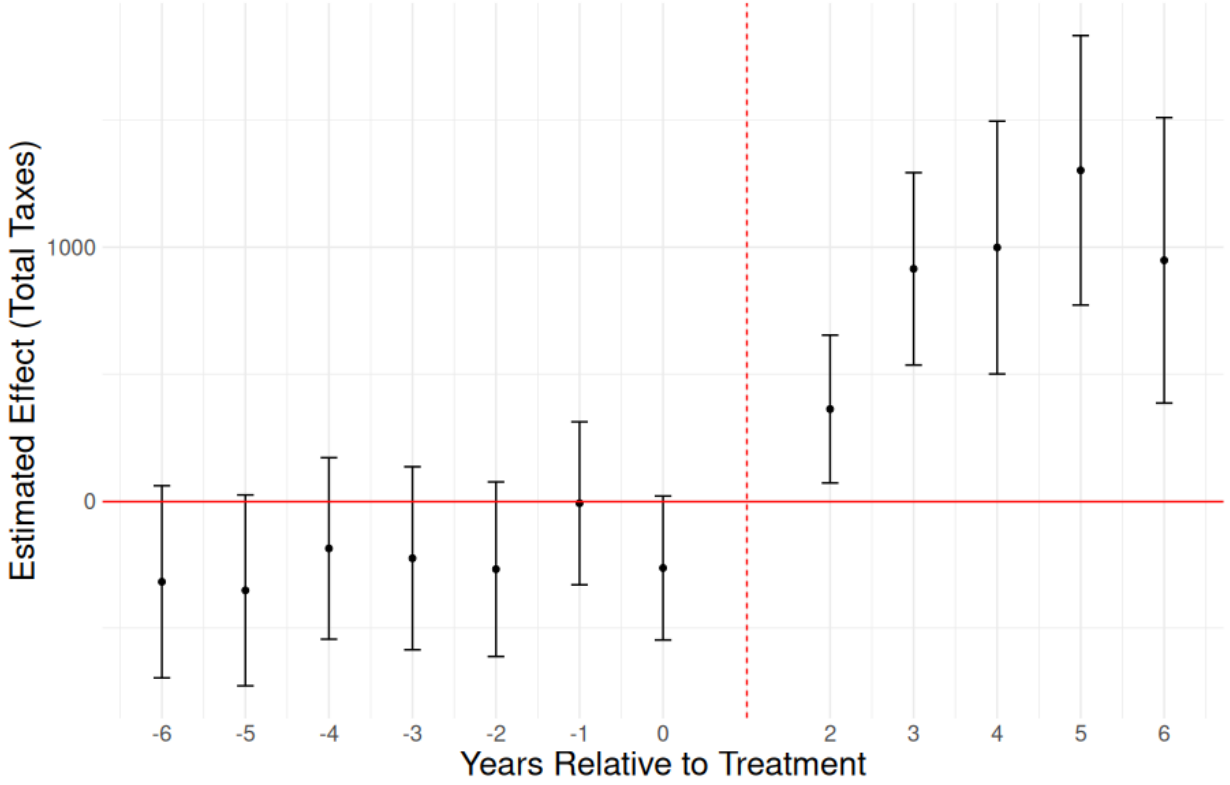


RD-DiD ITT on Yearly Total Taxes by Sample, 2017

Research Sample

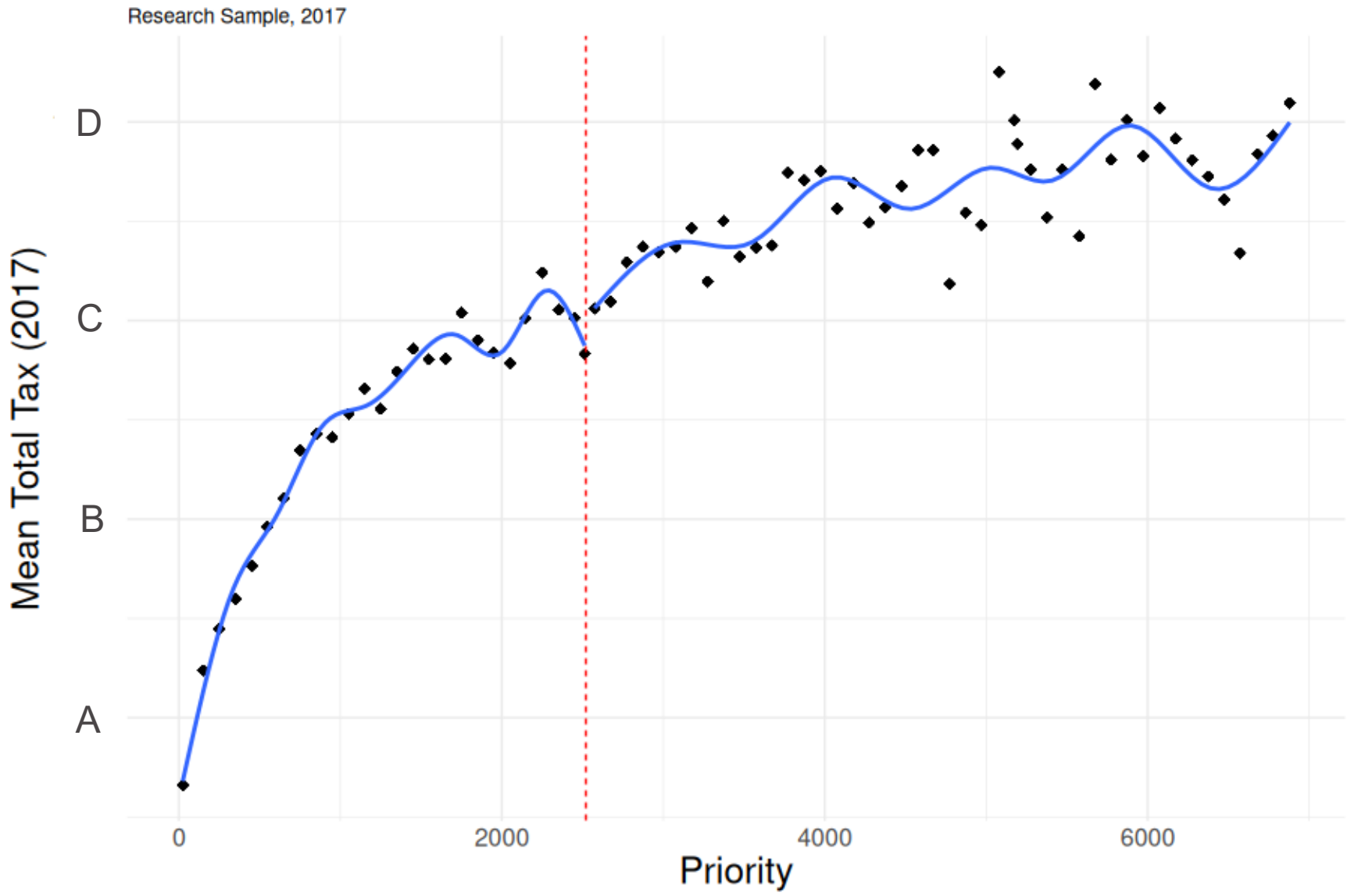


Status Quo Sample

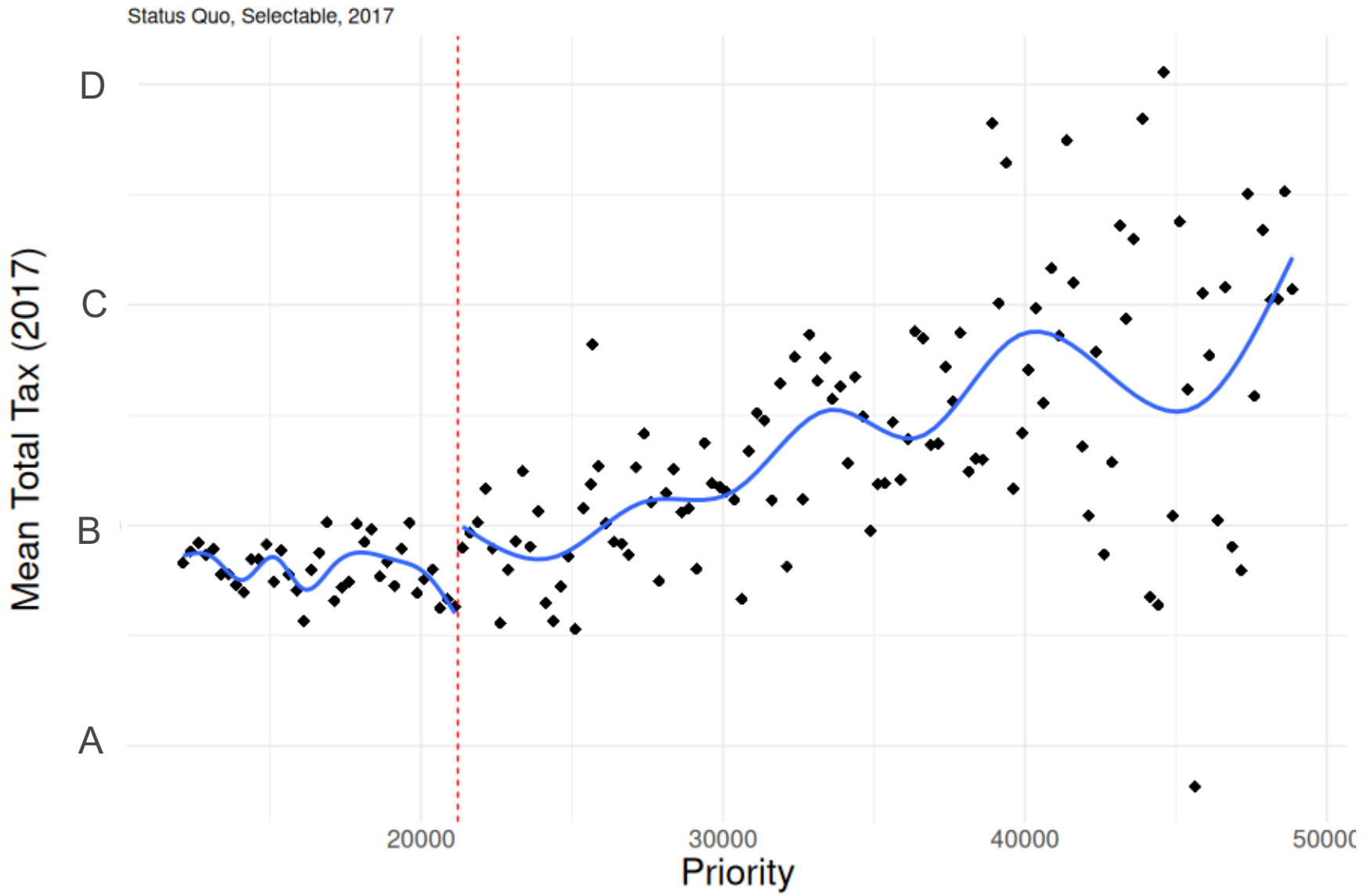


Note: To aid the comparison across cutoffs, estimates for cutoff 2 are multiplied by -1 so that they represent the change in taxes reported due to an increase in the probability of audit.

2017: Total Tax Around Cutoff, Research Sample



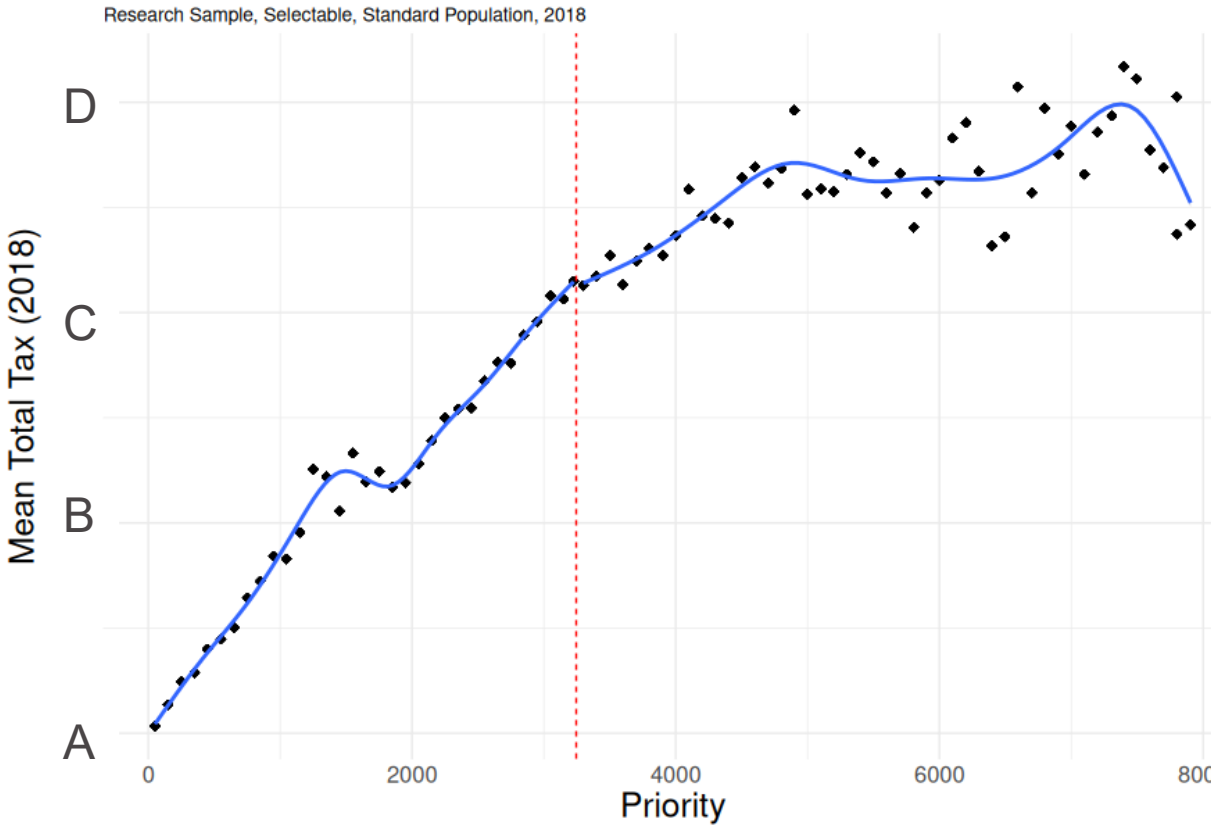
2017: Total Tax Around Cutoff, Status Quo



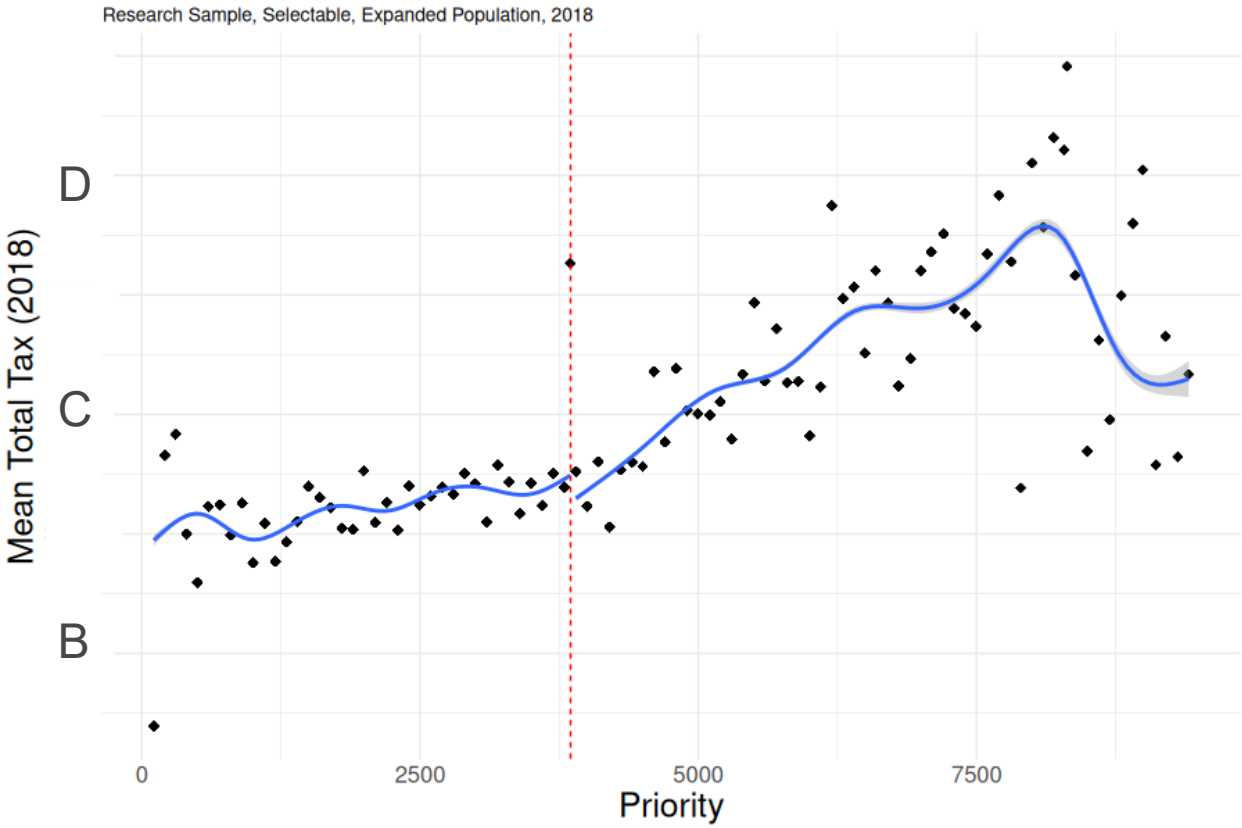
2018: RD-DiD

2018: Total Tax Around Cutoff, Research Sample, Selectable

Standard

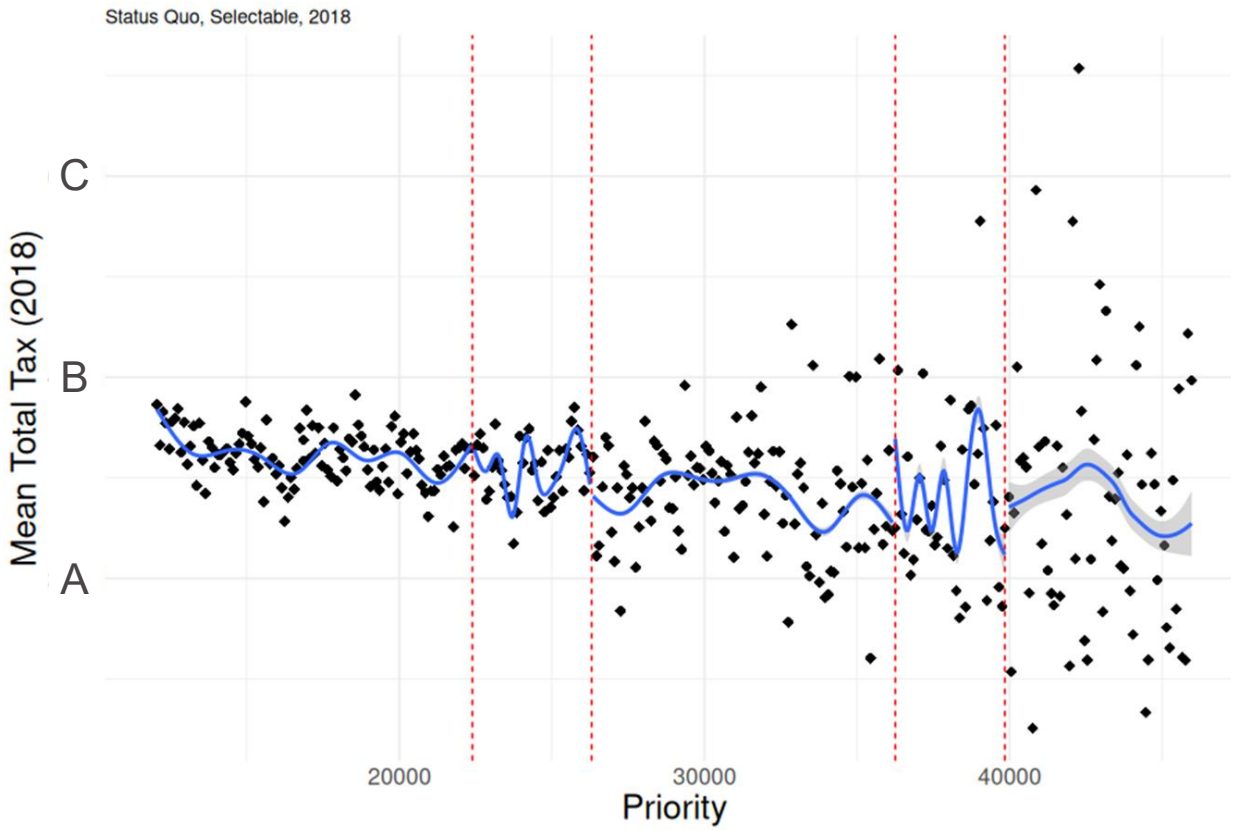


Expanded

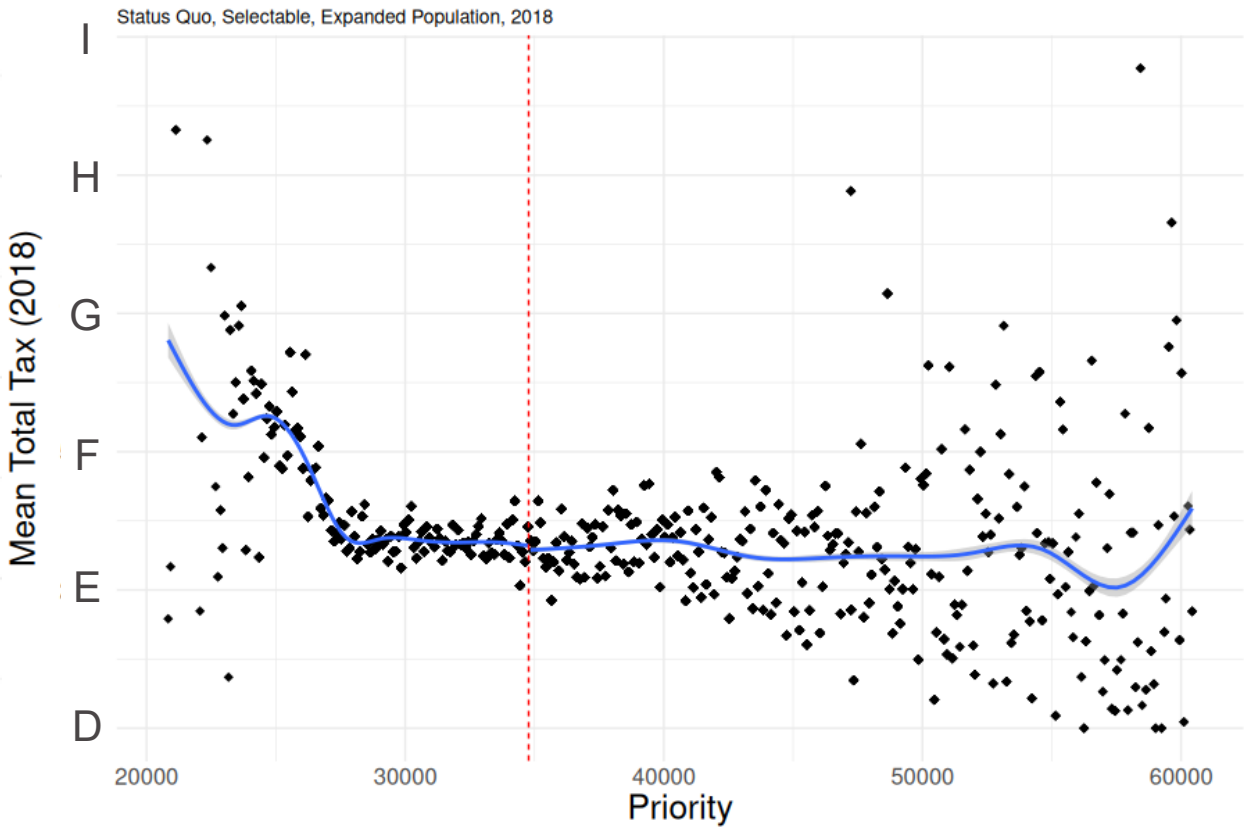


2018: Total Tax Around Cutoff, Status Quo Sample, Selectable

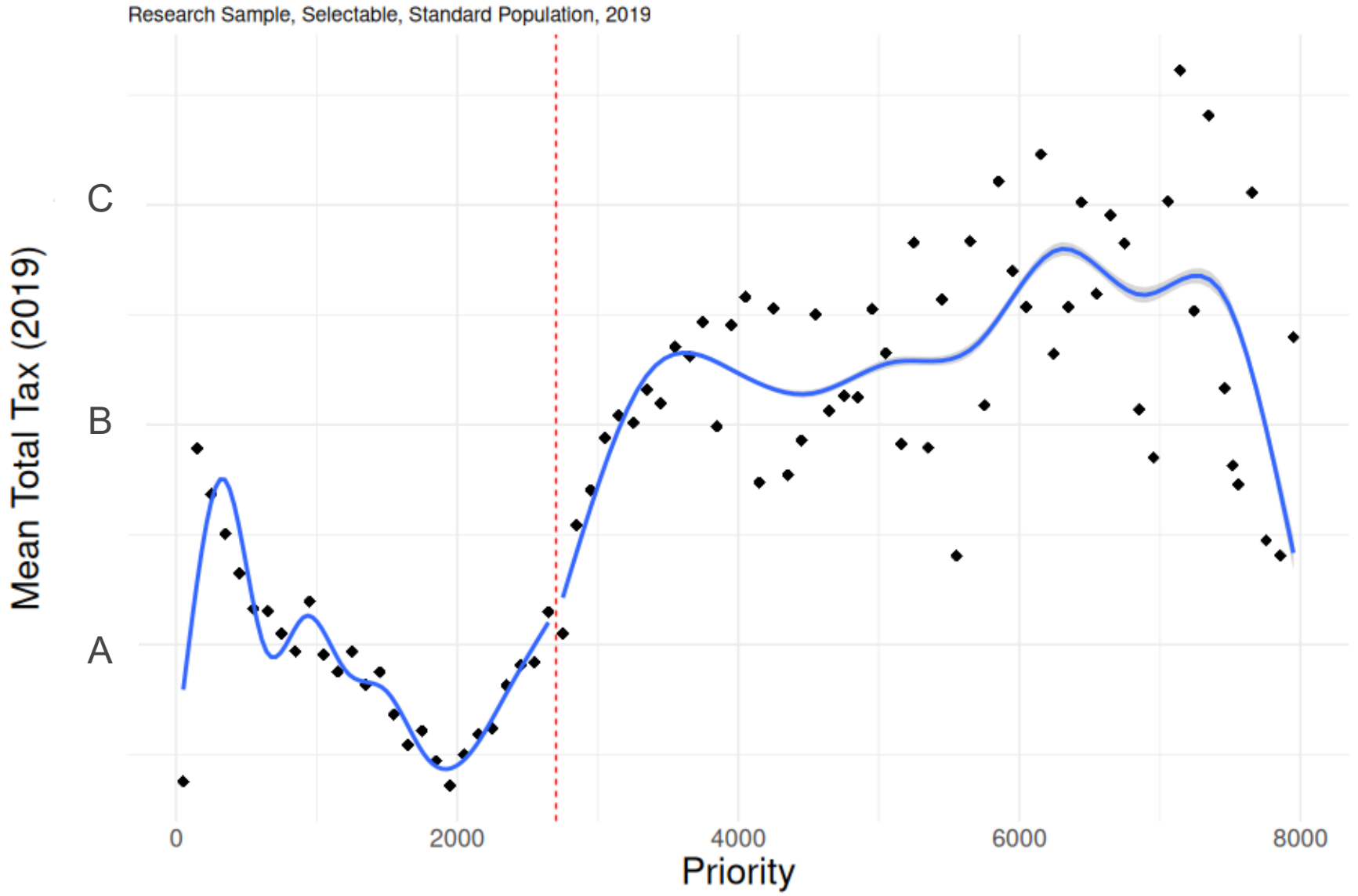
Standard



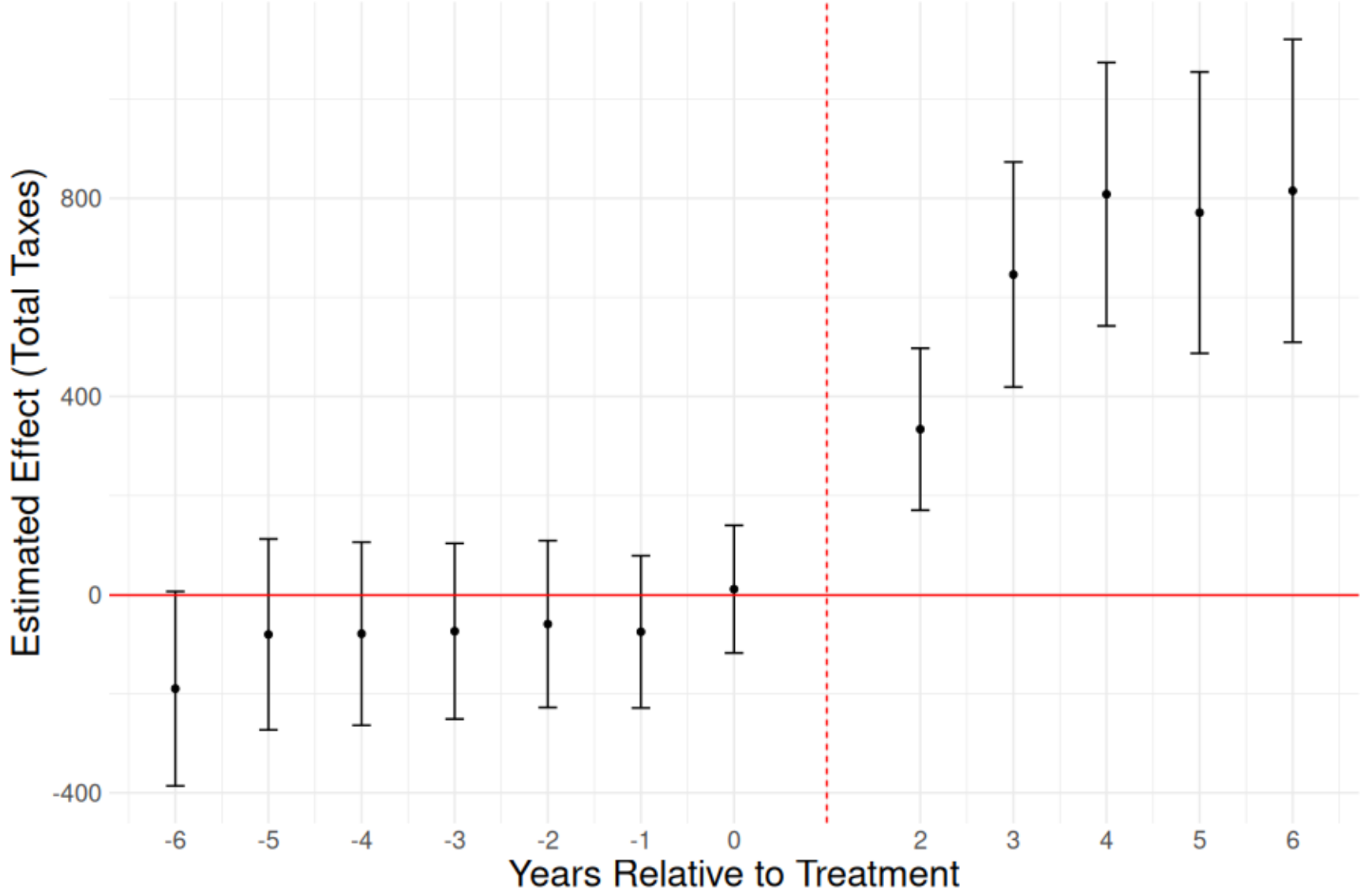
Expanded



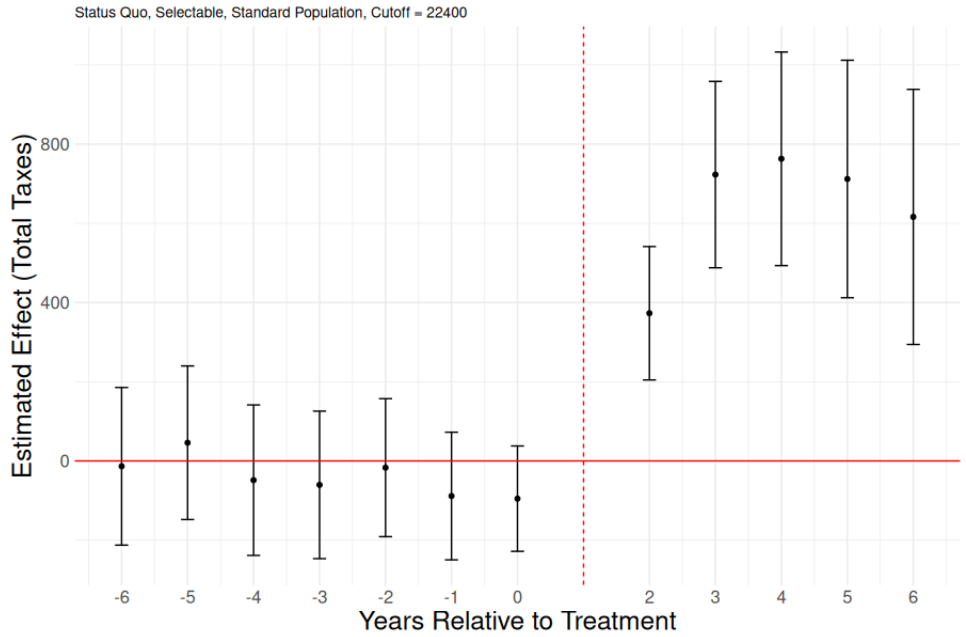
2019: Total Tax Around Cutoff, Research Sample, Selectable, Standard Population

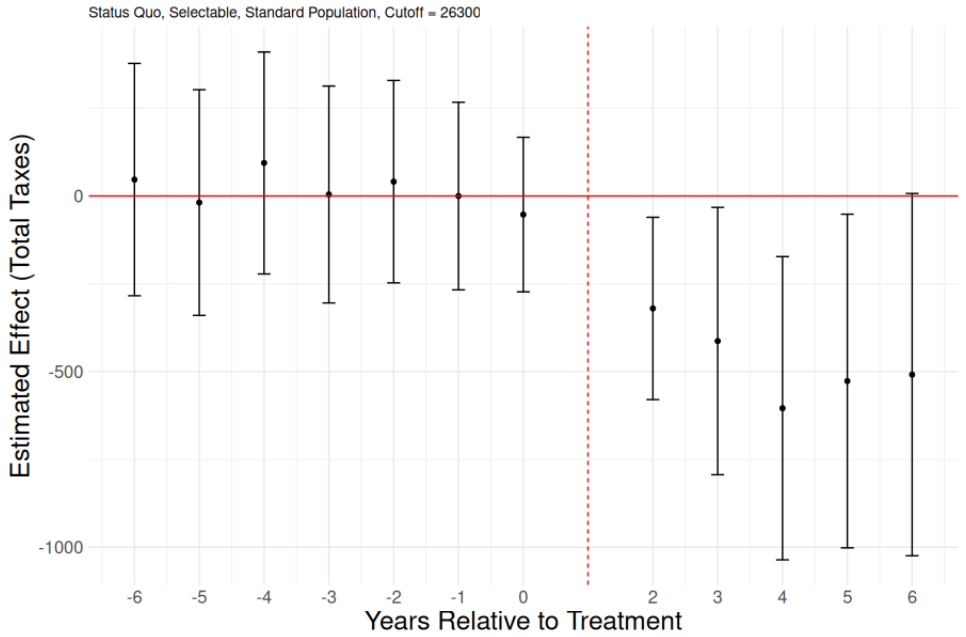


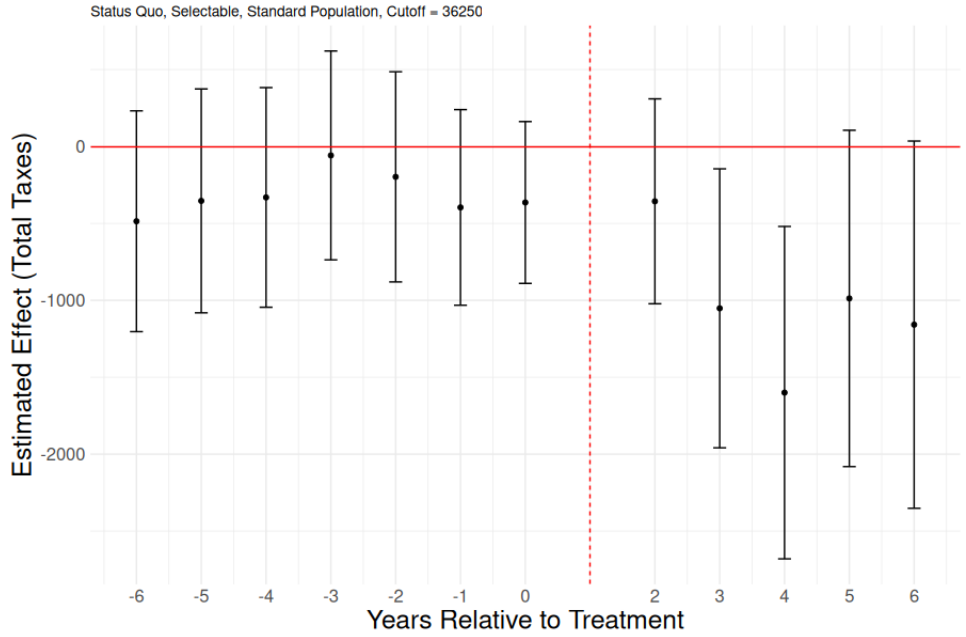
RD-DiD, ITT Research Sample, Selectable, Standard Population 2018

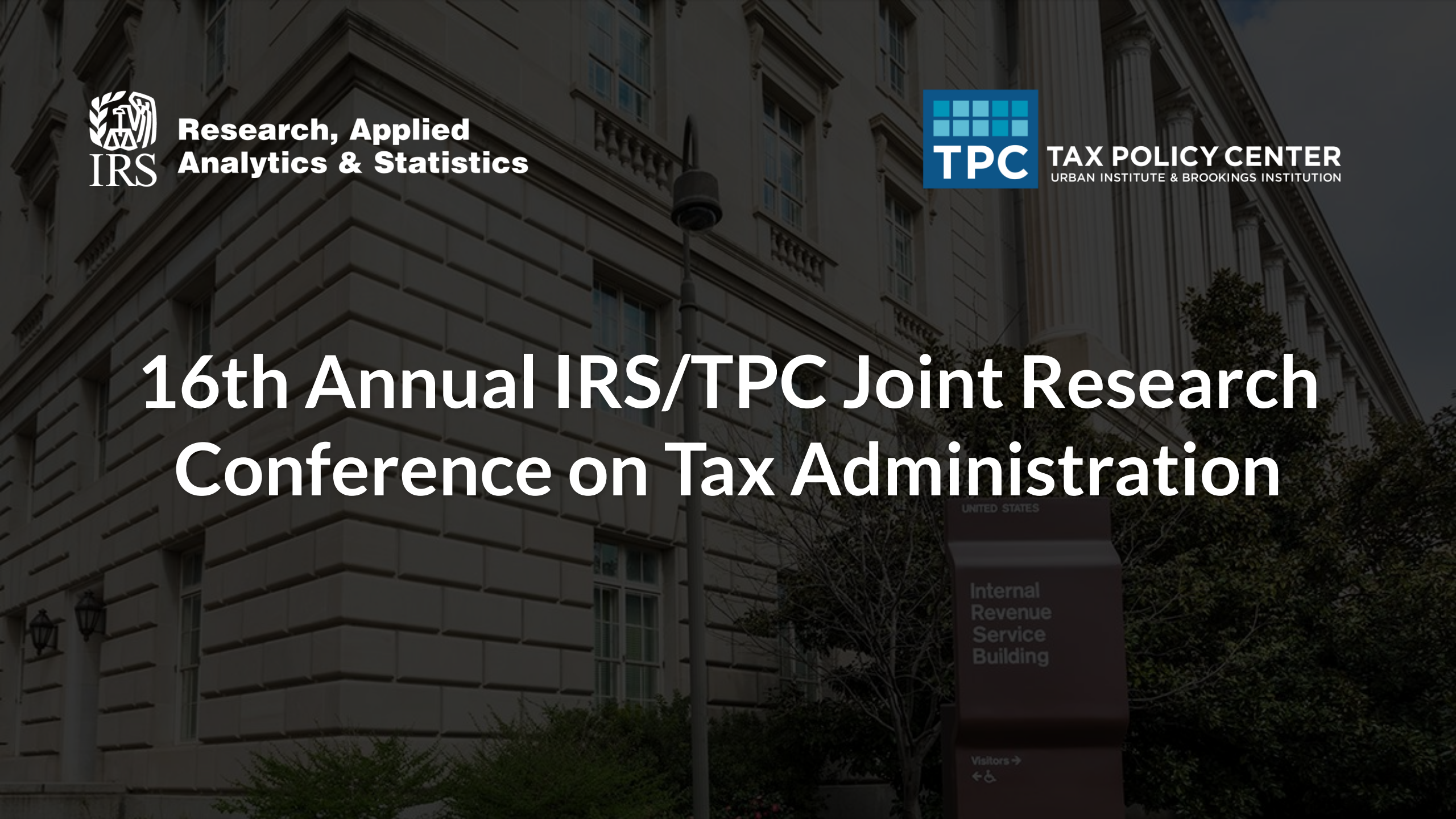


RD-DiD, ITT Status Quo Selectable, Standard Population, 2018









Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

16th Annual IRS/TPC Joint Research Conference on Tax Administration

Deterrence Spillovers through Social Networks:

Evidence from a New York City Tax Enforcement Crackdown

Zizhuo Chen | University of Minnesota | IRS-TPC Research Conference | June 25, 2026

Tax Enforcement Can Generate Network Spillovers Beyond Geography

I study the enforcement spillovers from a tax crackdown event in New York City in the summer of 2015.

Results suggest that those counties more closely connected to New York City subsequently showed higher tax compliance.

These spillover effects cannot be explained by geographic proximity alone and are consistent with deterrence spreading through social connections.

A Broader Approach to Studying Enforcement Spillovers

Tax enforcement is usually evaluated on its local effects.

E.g., Slemrod, Blumenthal, and Christian (2001), Kleven et al. (2011).

Some studies evaluate spillovers through specific channels.

E.g., de Paula and Scheinkman (2010) [trading partner], Paetzold and Winner (2016) [colleague], Boning et al. (2019) [tax preparer], Drago, Mengel, and Traxler (2020) [geographic proximity].

Using a broad measure of social connectedness to identify network spillovers.

E.g., Collins et al. (2025, IRS working paper).

Social connectedness subsumes the explanatory channel associated with geographic proximity.

NYC's 2015 Crackdown: An Unexpected and Salient Tax Enforcement Shock

In the summer of 2015, there was an unexpected major crackdown on electronic "Zappers" used by restaurants to conceal sales in NYC.

[How do "Zappers" work? Where do "Zappers" appear?]

This crackdown received substantial media coverage, e.g.:

Waiter, There's a Tax Dodge in My Soup

'Sales suppression software' in underhanded New York restaurants is lending a new meaning to prix fixe.

By Cyrus R. Vance Jr. And Kathryn Wylde
June 2, 2015 6:53 pm ET

Source: *The Wall Street Journal*

NYC looks to halt use of tech to underreport restaurant sales

By Peter Romeo on Jun. 10, 2015

Source: *Restaurant Business*

Did the Crackdown Affect Tax Compliance in Counties Connected to NYC?

Did counties more socially connected to NYC exhibit larger increases in tax compliance after the 2015 crackdown?

Did social connectedness generate exposure beyond geographic proximity?

How large and how persistent were these spillover effects?

[Theoretical Framework]

$$\max_{e \geq 0} \mathbb{E}[U] = (1-\tilde{p})\, U\left[y(1-\tau)+\tau e\right] + \tilde{p}\, U\left[y(1-\tau)-f\tau e\right]$$

y: real income
e: unreported income
$y-e$: reported income
τ: tax rate
f: penalty rate
$\tilde{p}$: perceived auditing rate

$$\tilde{p}_{i,t} = \hat{p}_{i,t} + \lambda s_i \kappa_t$$

$\hat{p}_{i,t}$: baseline perceived auditing rate
t: year since enforcement
λ: responsiveness of beliefs to enforcement information from social connections
s_i: overall social network exposure to the enforcement event
κ_t: salience of the enforcement information over time
$\forall\ t < 0,\ \kappa_t = 0;\ \forall\ t \geq 0,\ \kappa_t \geq 0;\ \lim_{t \to +\infty} \kappa_t = 0$

Implications:

$$\frac{\partial e^*_{i,t}}{\partial \tilde{p}_{i,t}} < 0, \quad \frac{\partial \tilde{p}_{i,t}}{\partial s_i} = \lambda \kappa_t, \quad \frac{\partial e^*_{i,t}}{\partial s_i} < 0 \text{ when } \kappa_t > 0$$

[Data]

Source	IRS SOI Dataset and Facebook SCI Dataset
Sample Period	2013–2020
Geographic Coverage	U.S. Counties (3104)
Unit of Observation	County-Year

Outcome Variable: Measure Tax Compliance Indirectly

$$\text{TCR}_{i,t} = \frac{\text{Total Reported Partnership and S-corp Net Income}_{i,t}}{\text{Total Reported Salaries and Wages Amount}_{i,t}}$$

The literature confirms that potential "Zapper" users have high cash intensity, low third-party reporting, concentrated ownership, and fewer employees, which are typical features of partnerships and S-corps (Ainsworth, 2008; Vance and Wylde, 2015; Adhikari, Alm, and Harris, 2021).

Divided by salaries and wages to control for local economic fundamentals.

Estimated noncompliance rate: 1% (Kleven et al., 2011; Slemrod, 2019).

To capture within-county changes over time.

Independent Variable: Measure Exposure

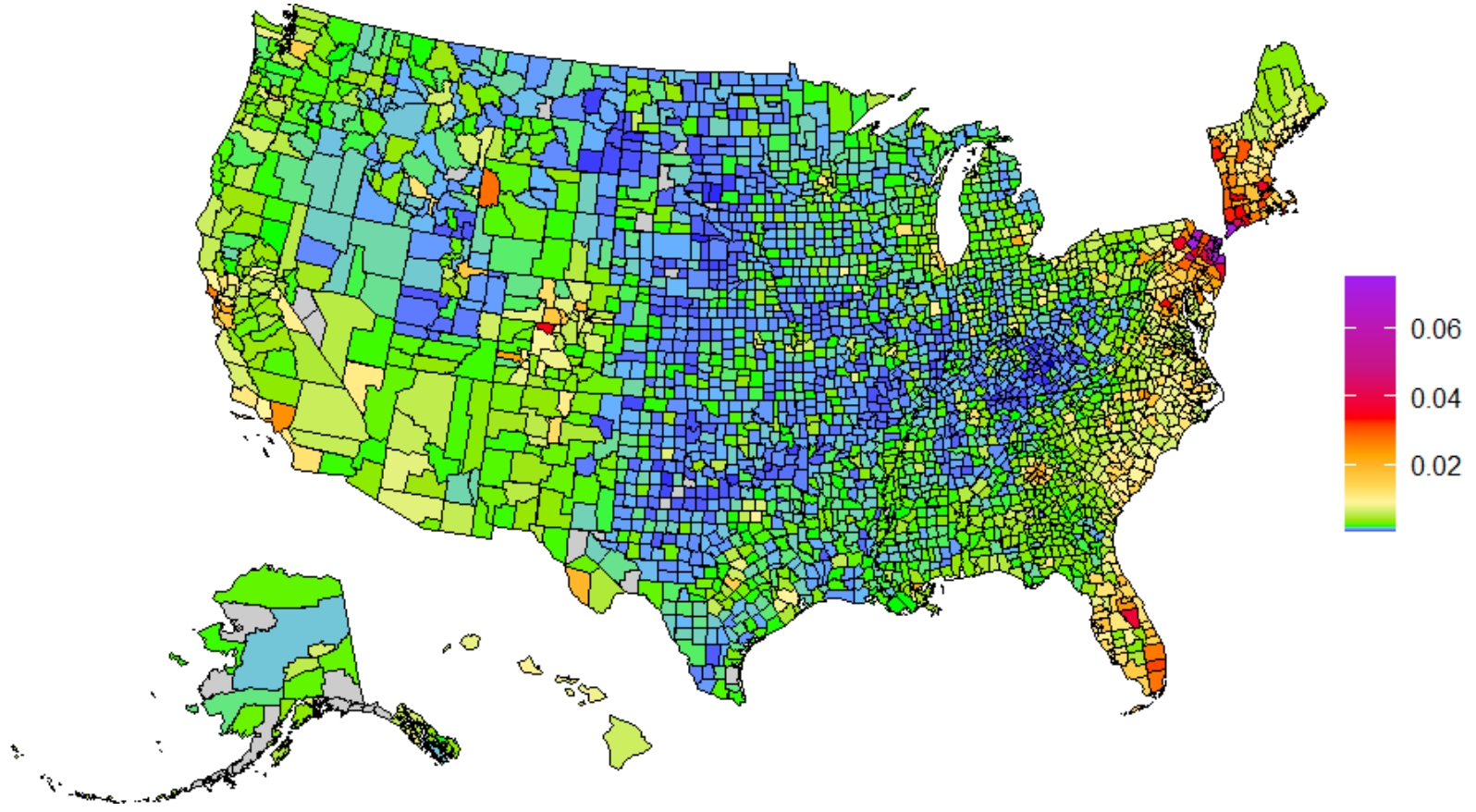
$$\mathrm{RSC}_{i,\mathrm{NYC}} = \frac{\sum_{j \in \mathrm{NYC\ Counties}} \#\mathrm{Friendship}_{i,j}}{\sum_{k \in \mathrm{USA\ Counties}} \#\mathrm{Friendship}_{i,k}}$$

Relative social connectedness to New York City, calculated based on Facebook Social Connectedness Index (Bailey et al., 2018).

The measure proxies for the intensity of potential information flows between counties and New York City through social interactions in general.

Chetty et al. (2022): On Facebook, friendship is two-way; "Facebook friendship network can therefore be interpreted as providing data on people's real-world friends and acquaintances rather than purely online connections."

A Visualization of Relative Social Connectedness to NYC



Counties in New York State were excluded.

[Identification]

Compare More and Less NYC-Connected Counties Year by Year

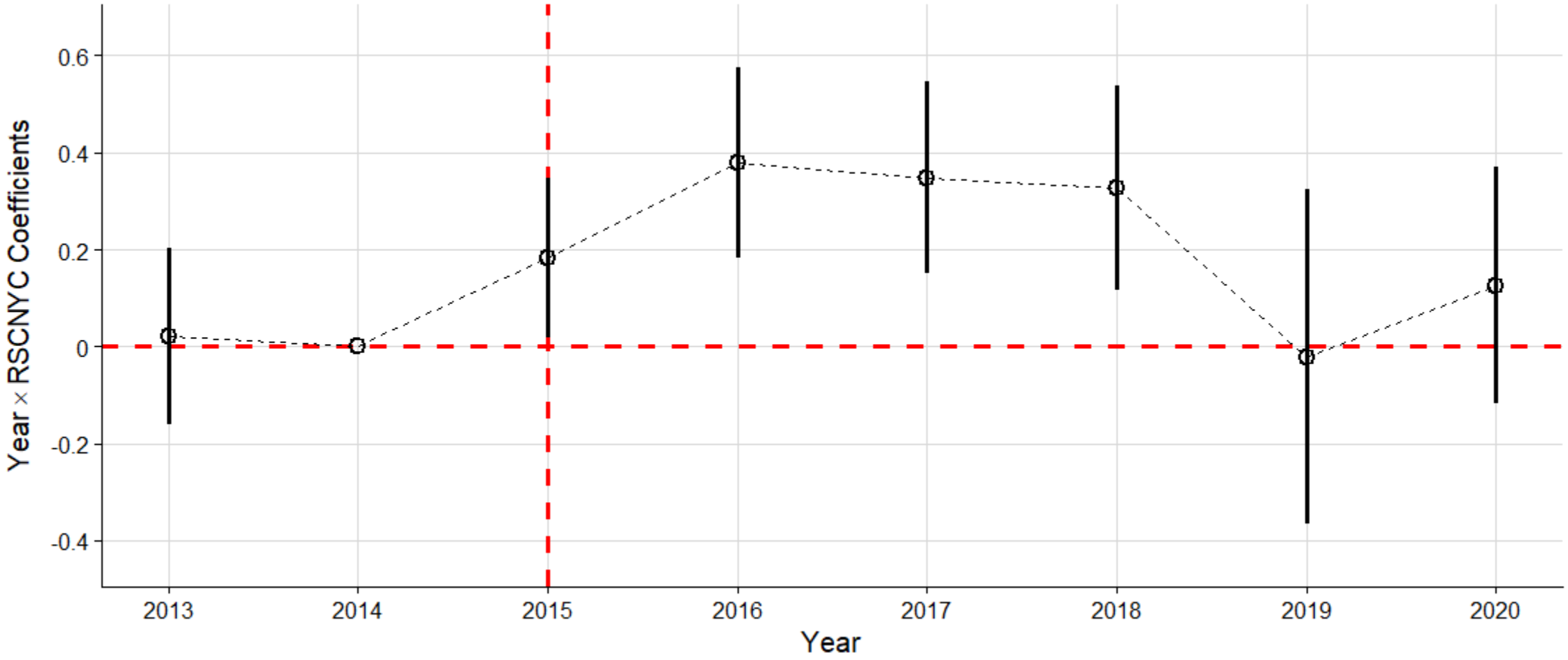
$$\mathrm{TCR}_{i,t} = \sum_{t=2013,\ t\neq 2014}^{2020} \beta_t \cdot \mathrm{Year}_t \cdot \mathrm{RSCNYC}_i + \alpha_i + \theta_{State,t} + \varepsilon_{i,t}$$

If social networks transmit enforcement salience, counties more connected to NYC should exhibit larger compliance changes for several years after 2015.

County fixed effects and state-year fixed effects included.

Standard errors clustered two-way: county and state-year.

Baseline Results: Evidence of Social Network Enforcement Spillovers



Confidence intervals are 90%. Same below.

Geographic Proximity "Absorbed into" Social Network Connectedness

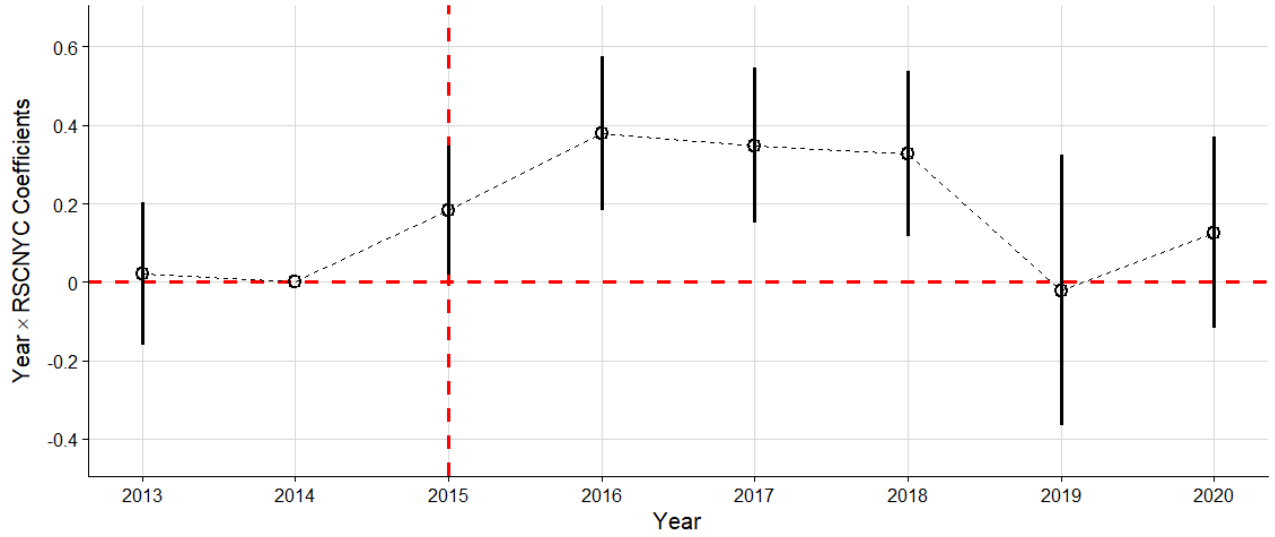
$$\begin{aligned}\mathrm{TCR}_{i,t} = &\sum_{t=2013,\ t\neq 2014}^{2020} \eta_{1,t} \cdot \mathrm{Year}_t \cdot \mathrm{RSCNYC}_i \\ &+ \sum_{t=2013,\ t\neq 2014}^{2020} \eta_{2,t} \cdot \mathrm{Year}_t \cdot \log\left(\mathrm{DistanceNYC}_i\right) \\ &+ \alpha_i + \theta_{State,t} + \varepsilon_{i,t}\end{aligned}$$

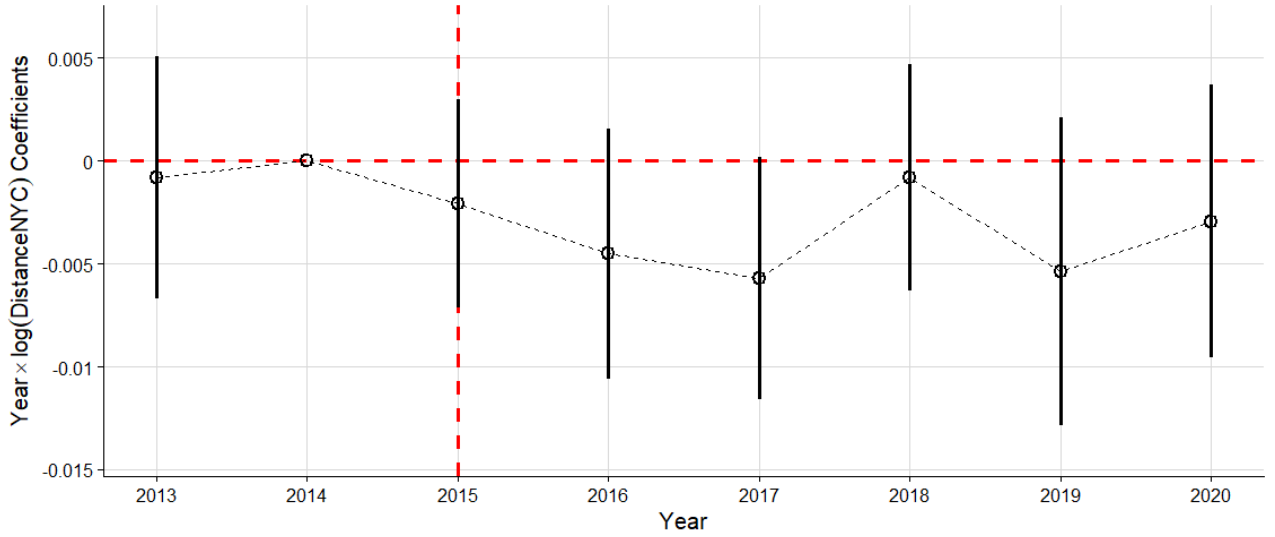
To see whether geographic-distance spillovers are contained in social-network spillovers.

Include the interaction term of geographic distance and year and check whether they attenuate the social-network coefficients.

[Empirical Results II]

Spillover Effects via Geographic Proximity Superseded





Left Graph: Social-network spillover estimates remain similar.

Right Graph: Geographic-distance coefficients are not statistically significant.

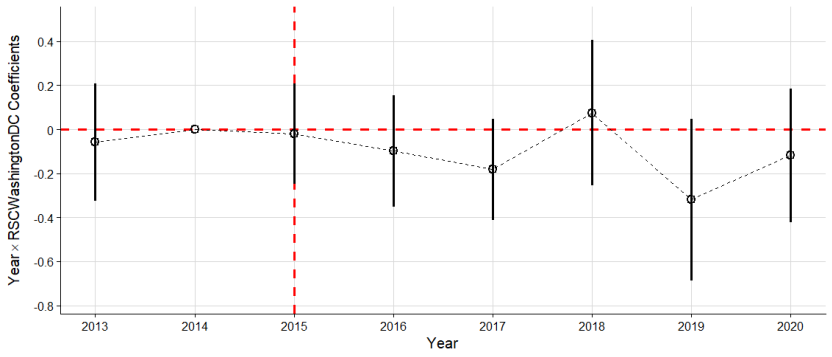
[Robustness: Placebo Tests]

Consider "Imaginary" Crackdown Events in Other Cities

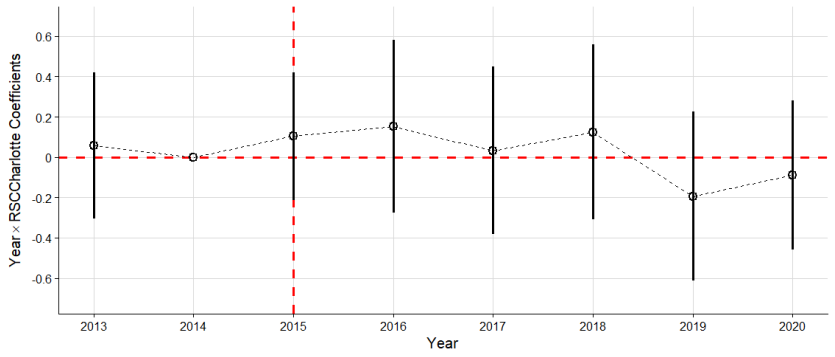
The results are not general responses to social exposure to large metropolitan areas.

Using alternative metropolitan areas without comparable crackdowns for placebo analyses (Washington, D.C., Charlotte, and Philadelphia).

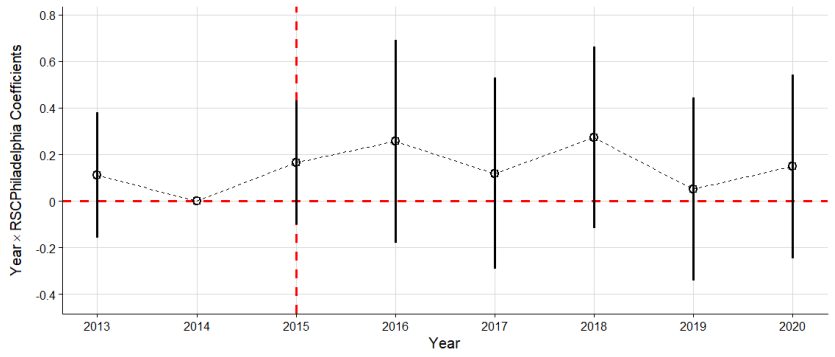
No spillover effect from these "imaginary" crackdowns.



Washington, D.C.



Charlotte



Philadelphia

[Robustness: Falsification Test]

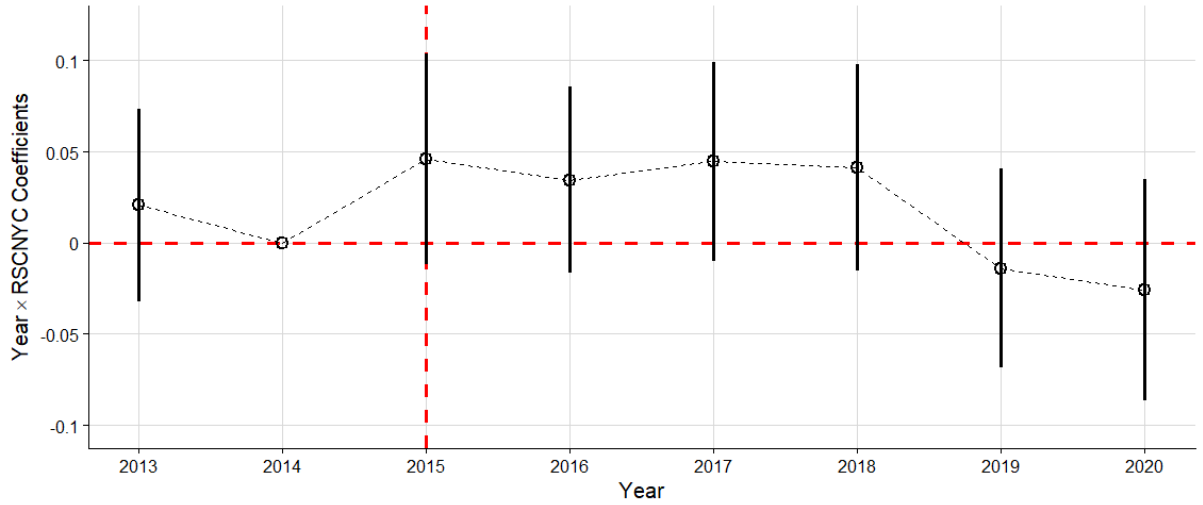
Consider Tax Compliance for Another Tax Schedule

$$\text{TCR\_Fal}_{i,t} = \frac{\text{Total Reported Business and Professional Net Income}_{i,t}}{\text{Total Reported Salaries and Wages Amount}_{i,t}}$$

The results are not merely a reflection of broader tax trends.

Using reported business or professional net income (sole proprietors, contractors, etc.) instead of partnership/S-Corp net income.

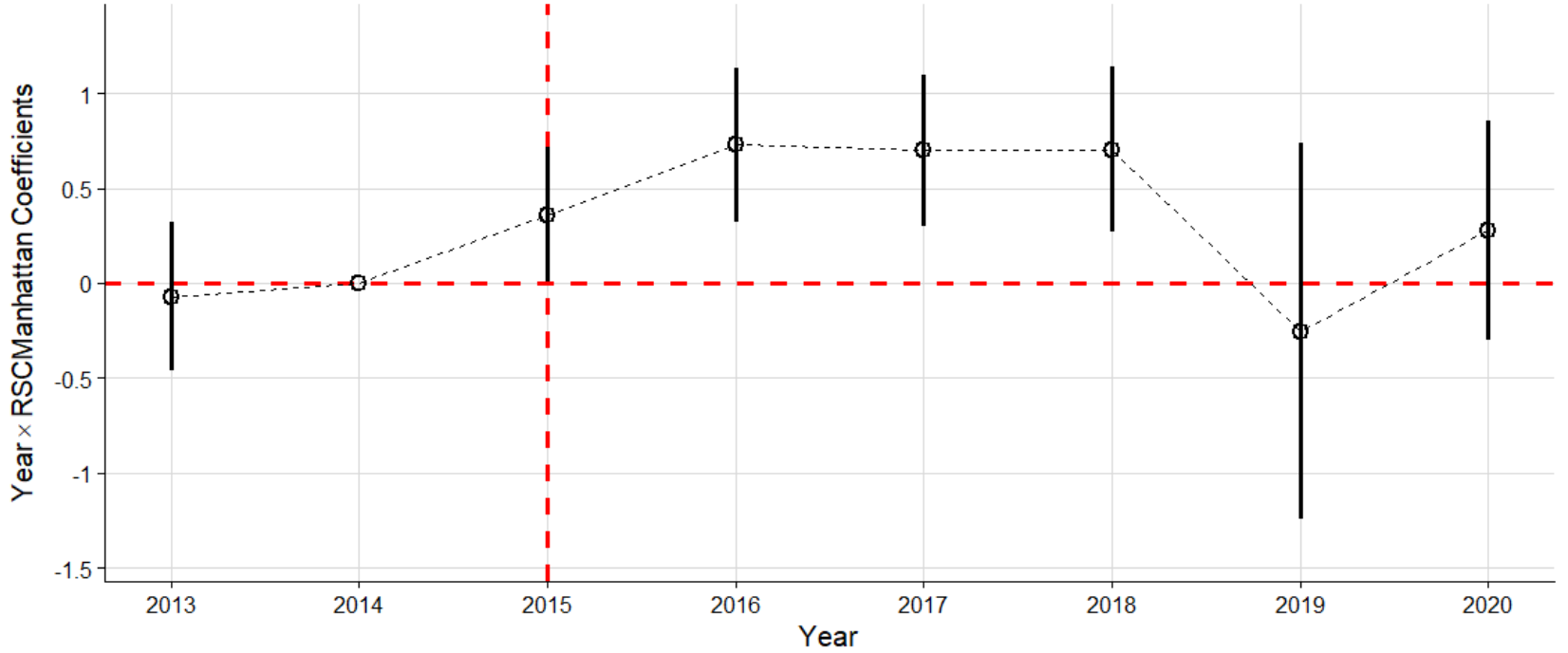
No spillover effect identified here.



Consider Social Connections to Manhattan

In news reports, the crackdown was described and perceived as an event of New York City. But it was initiated by Manhattan District Attorney Cyrus Vance Jr.

Running the core regression with RSCManhattan instead of RSCNYC yields similar patterns.



[Implications]

Evaluations focusing only on the jurisdiction in which enforcement occurs may **understate** the total effect of a tax-enforcement policy.

Considering **social-network** coverage when evaluating enforcement actions.

A question for future research is whether the **network position** and **public salience** of enforcement actions shape their aggregate deterrence effects.

[Limitations]

1. “Total reported net income of partnerships and S-corps” is not an ideal proxy for “total reported net income of potential ‘Zapper’ users”.

[Micro-level data for small restaurants and retailers might be better.]

2. The assessment of pre-trends is based only on 2013 and 2014.

[If SOI can extend further back, it might be better.]

3. Facebook data may not perfectly represent real-world social networks.

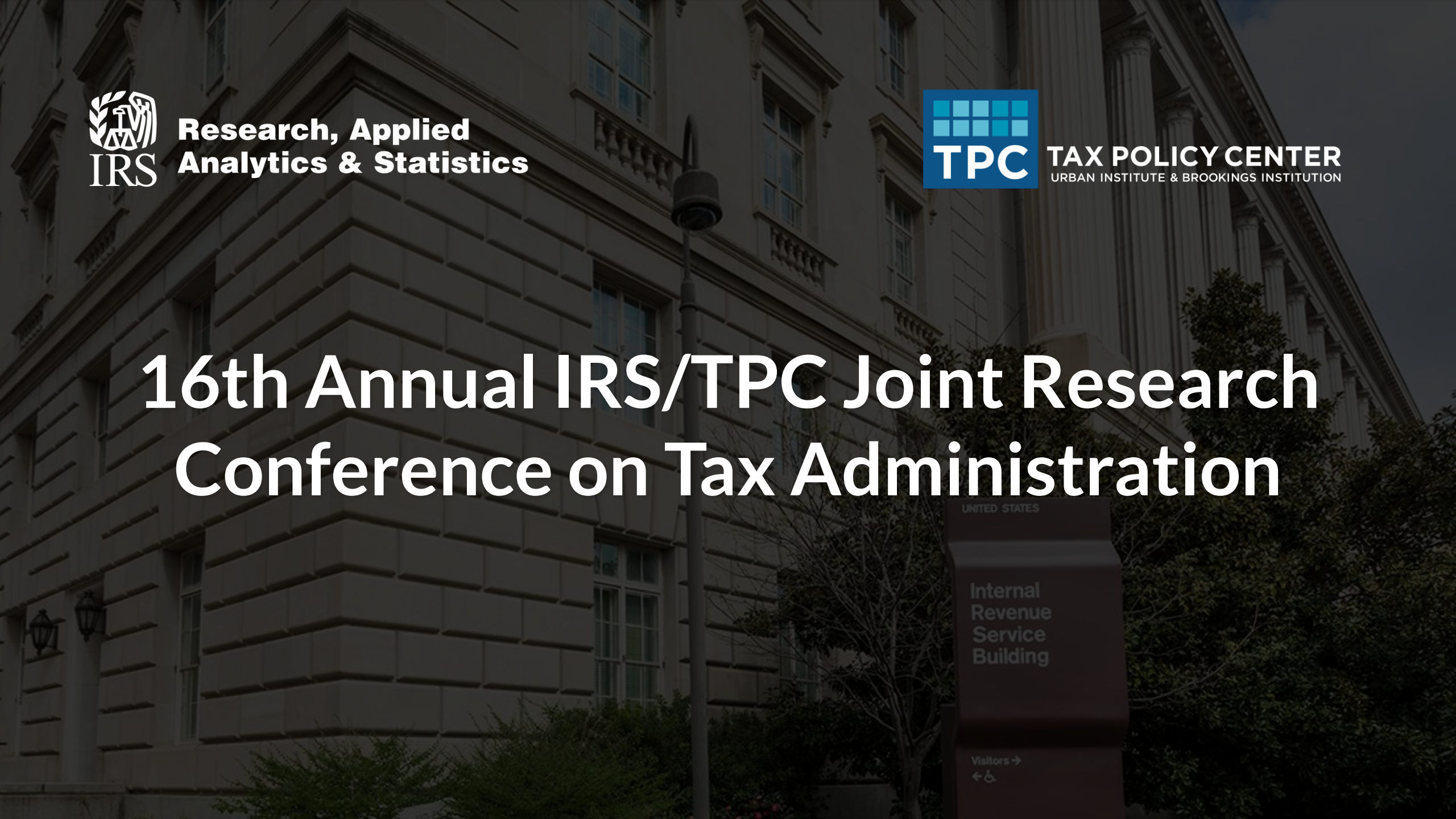
[Not uniformly distributed across age groups.]

Deterrence Spillovers through Social Networks:

Evidence from a New York City Tax Enforcement Crackdown

Thank you for listening!

Zizhuo Chen | University of Minnesota | IRS-TPC Research Conference | June 25, 2026



Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

16th Annual IRS/TPC Joint Research Conference on Tax Administration

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

The Specific Indirect Effect of IRS' Automated Underreporting (AUR) Enforcement

Tom Hertz (IRS-RAAS) and Miguel Sarzosa (MITRE)

2026 IRS-TPC Research Conference

June 2026

Presenter: Tom Hertz

Disclaimers

- The views expressed here are those of the authors or MITRE and are not necessarily in accordance with the views of the Government.
- Approved for Public Release; Distribution Unlimited. Public Release Case Number 26-0908

NOTICE

This (software/technical data) was produced for the U. S. Government under Contract Number TIRNO-99-D-00005, and is subject to Federal Acquisition Regulation Clause 52.227-14, Rights in Data—General, Alt. I, II, III and IV (MAY 2014) [Reference 27.409(a)].

No use other than that granted to the U. S. Government, or to those acting on behalf of the U. S. Government under that Clause is authorized without the express written permission of The MITRE Corporation.

For further information, please contact The MITRE Corporation, Contracts Management Office, 7515 Colshire Drive, McLean, VA 22102-7539, (703) 983-6000.

Introduction

- Revenue agencies increasingly rely on automated systems to detect discrepancies in taxpayer reporting (IMF, 2015; OECD, 2015).
- ***IRS' Automated Underreporter (AUR)*** identifies discrepancies between individual tax returns and third-party information submitted to the IRS.
- AUR enforcement revenue: $4 billion in 2023. ROI = 18:1
- The AUR data can be used to study behavioral responses of taxpayers to enforcement interventions in a large-scale quasi-experimental analysis
- We leverage *sharp discontinuities in the probability of AUR treatment* to identify the causal effect of receiving an AUR notification on subsequent tax reporting
- Note: We look only at AUR's W2-related enforcement, not at all workstreams

AUR Case Selection Process

- Each year, AUR compares approximately 150 million individual income tax returns to over 4.8 billion information documents (e.g. W2s, 1099s, etc)
- The identification of a line-item discrepancy triggers the creation of an AUR case, but not all cases are subject to further action.
- Those flagged for further action are selected with three goals in mind:
    1. Maximize the dollar-yield of the cases processed
    2. Address repeat offenders
    3. Process a broad range of types of noncompliance
- Case selection occurs in batches called “correlations”

AUR Case Selection Process, Cont'd

- **Filter 1**: Drop unproductive/unworkable cases according to a set of year-specific predetermined business rules (GNS & NON)
- **Filter 2**: Cases ranked based on the Estimated Potential Assessment (EPA) and selected in descending order; volume determined by staff availability.
  - The EPA metric is an estimate of expected yield, based on historical AUR data.
  - The cases that fall below the corresponding EPA threshold comprise the "untreated" pool of cases (UNT).
  - Most of the remaining cases receive treatment (i.e., CP2000 letters). These are the selected cases (SEL).

Data, cont'd

	TY2013 Count	TY2013 Mean EPA	TY2014 Count	TY2014 Mean EPA
Overall Counts				
Total Observations	2,842,956	377	3,254,315	226
Unique TINs	1,657,015		1,864,837	
By Correlation				
1	1,513,516	443	1,742,190	319
2	1,319,781	304	1,494,487	116
By Repeater Code				
First Time Underreporters (OT)	1,643,049	268	2,091,067	221
Repeater Not Worked (RN)	832,338	472	841,611	418
Repeater Worked (RW)	367,522	648	321,637	166
By Treatment Code				
NON	1,601,976	290	2,076,631	72
SEL	376,191	1113	308,374	1350
UNT	864,742	218	869,310	195

Empirical Strategy: Regression Discontinuity Diff-in-Diff

- Case selection in descending order of EPA + IRS' capacity constraints:
  - Discontinuity in EPA creates a drastic change in treatment probability at $\widetilde{EPA}$
  - Identify the causal effect of AUR by comparing the change in outcomes over time for taxpayers just above and just below the threshold
- RD-DiD exploits the cut-off-based treatment in a panel data context.
- The reduced form model can be specified as:

$$Y_{it} = \alpha + \beta \text{Post}_t + \gamma \mathbf{1}\left[\text{EPA}_i \geq \widetilde{EPA}\right] + \tau \text{Post}_t \mathbf{1}\left[\text{EPA}_i \geq \widetilde{EPA}\right] + f_0(\text{EPA}_i) + f_1(\text{EPA}_i) + \mu_t + \varepsilon_{it}$$

  where $\text{Post}_t = 1$ when $t \geq \tilde{t}$ ($\tilde{t}$ is time of treatment), $f.(\text{EPA}_i)$ are flexible functions of the EPA score, and μ_t is the tax year fixed effect.
- Identifying assumption: in the absence of treatment, the outcome trends for units just above and just below $\widetilde{EPA}$ would have evolved in parallel over time.

Empirical Strategy: Fuzzy RD-DiD

- Probability of treatment does not perfectly jump from 0 to 1 as EPA rises above threshold
- This calls for "Fuzzy RD-DiD": We instrument treatment with the exogenous discontinuity using a two-stage least squares estimator.
- First-stage regression:

$$D_i = \varphi + \psi \mathbf{1}\left[\text{EPA}_i \geq \widetilde{EPA}\right] + g_0(\text{EPA}_i) + g_1(\text{EPA}_i) + \nu_i$$

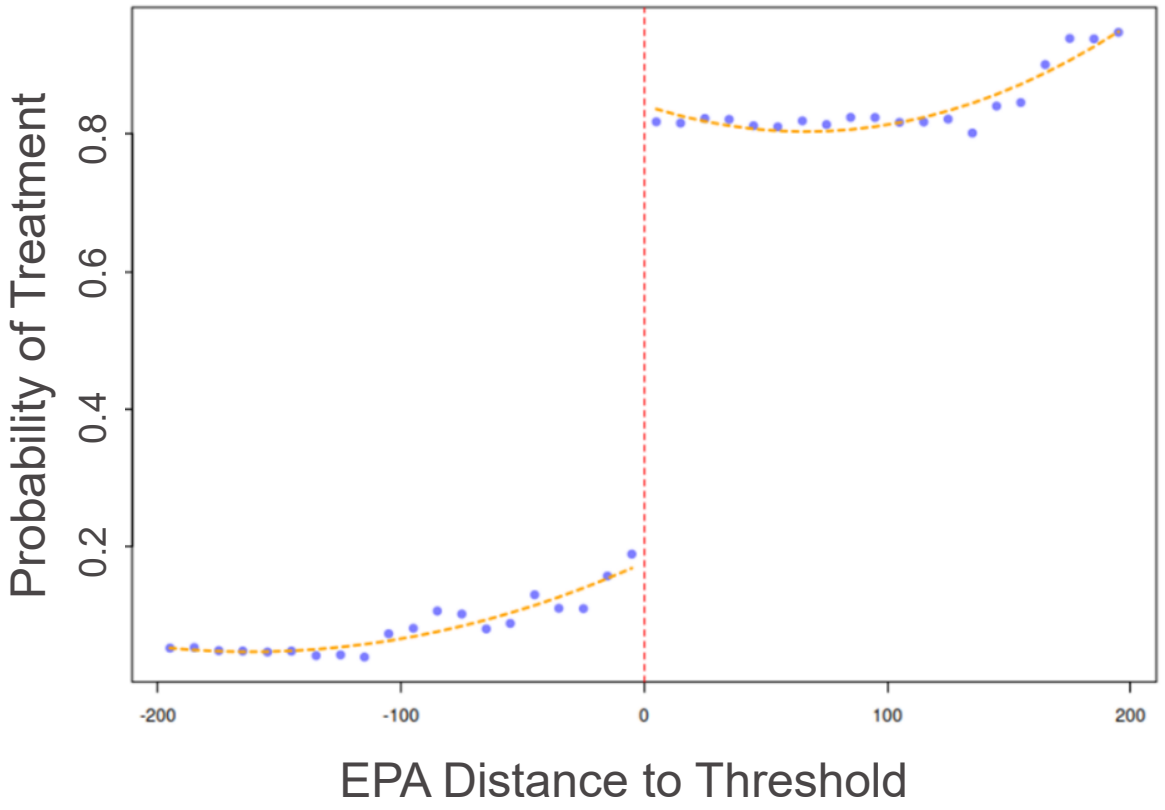
  where D_i is the binary treatment indicator

- Second-stage regression includes the predicted treatment probability $\widehat{D_i}$:

$$Y_{it} = \tilde{\alpha} + \tilde{\beta}\text{Post}_t + \tilde{\gamma}\widehat{D_i} + \lambda\text{Post}_t\widehat{D_i} + f_0(\text{EPA}_i) + f_1(\text{EPA}_i) + \mu_t + \varepsilon_i$$

- λ is the LATE for those whose treatment status is changed by the cutoff.

Results: RD First Stage AUR 2013/2014



	Dependent Var: Case Selected Vs Untreated		
	(1)	(2)	(3)
1[EPA>Threshold]	0.756***	0.709***	0.711***
	(0.008)	(0.008)	(0.008)
Year-Corr FE		X	Did
Repeater Code FE			X
Observations	30,343	30,343	30,343
R2	0.645	0.664	0.671

Note: All regressions include two quadratic functions of EPA distance to the threshold (EPA=500), one for the negative side and another one on the positive range. Year-Corr FE stands for treatment year and correlation cycle fixed effects. The sample includes SEL/UNT Correlation 1 cases from 2013 and 2014 within 39.47 EPA points around the threshold (optimal bandwidth by Calonico (2014)), for whom data for tax years 2010 to 2019 could be found, and excludes taxpayers with tax credits above $C in treatment year and egregious repeaters. Standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01

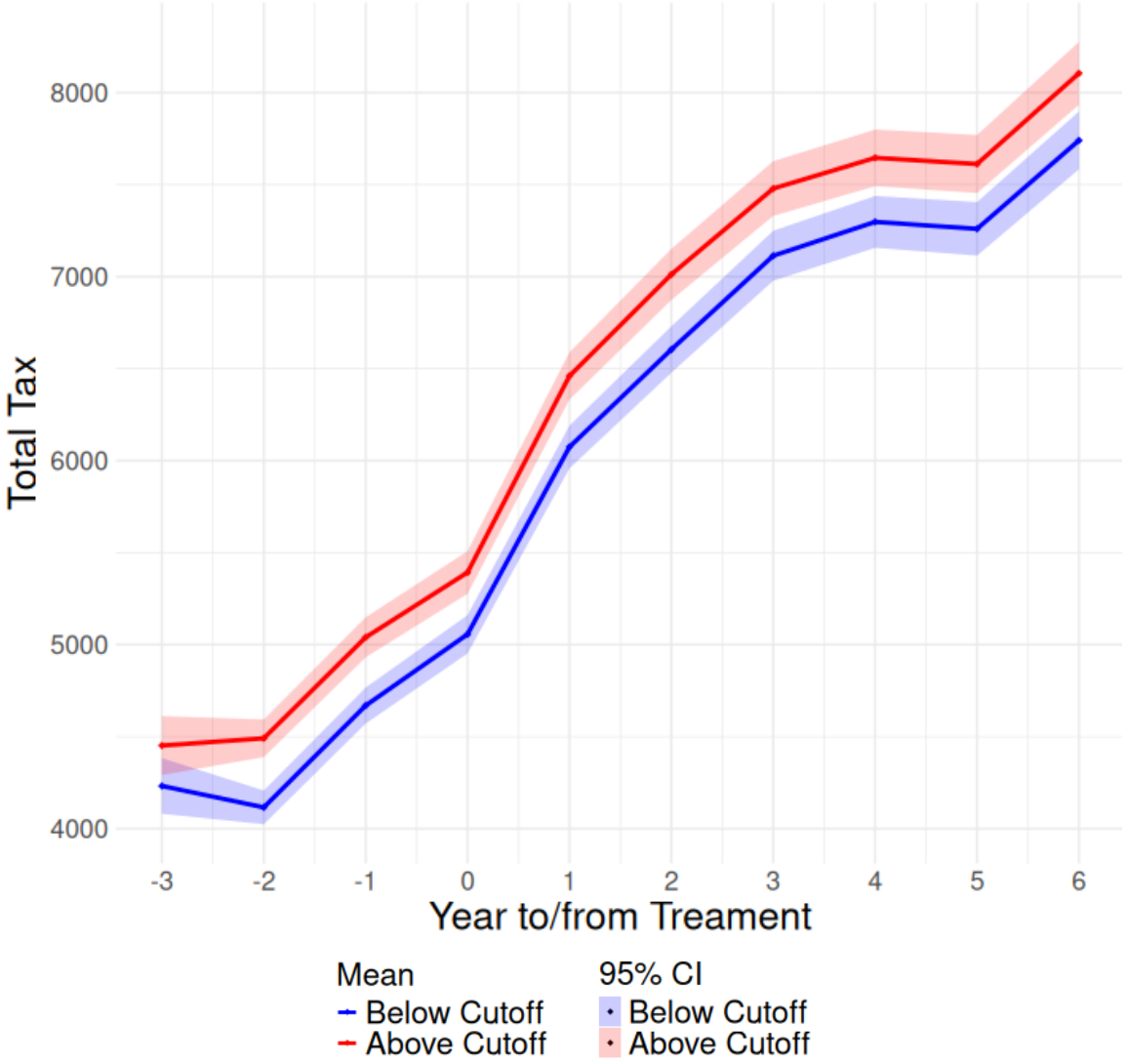
Average Total Tax by Group Relative to Threshold

Check for parallel trends in total tax prior to treatment

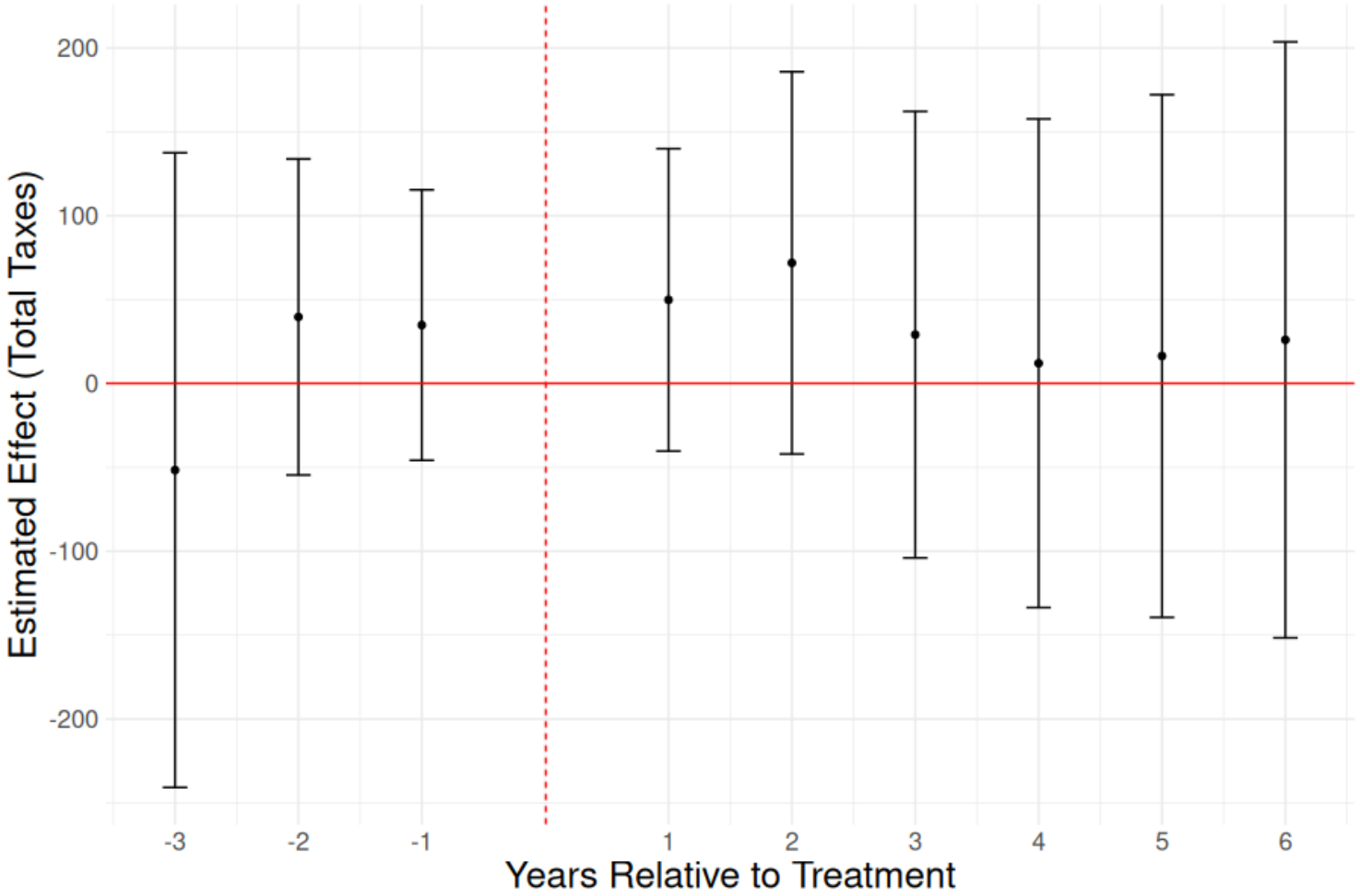
- People with $\text{EPA}_i \geq \widetilde{EPA}$ report higher taxes due to positive correlation between EPA and taxes.
- This would be a problem for a static RD
- RD-DiD deals with this by asking whether that gap grows over time after treatment (assumes the drivers of the gap difference are time-invariant).

But this plot also implies no treatment effect!

- Total taxes for those just above (red) and just below the cutoff (blue) continue their parallel trend after treatment.



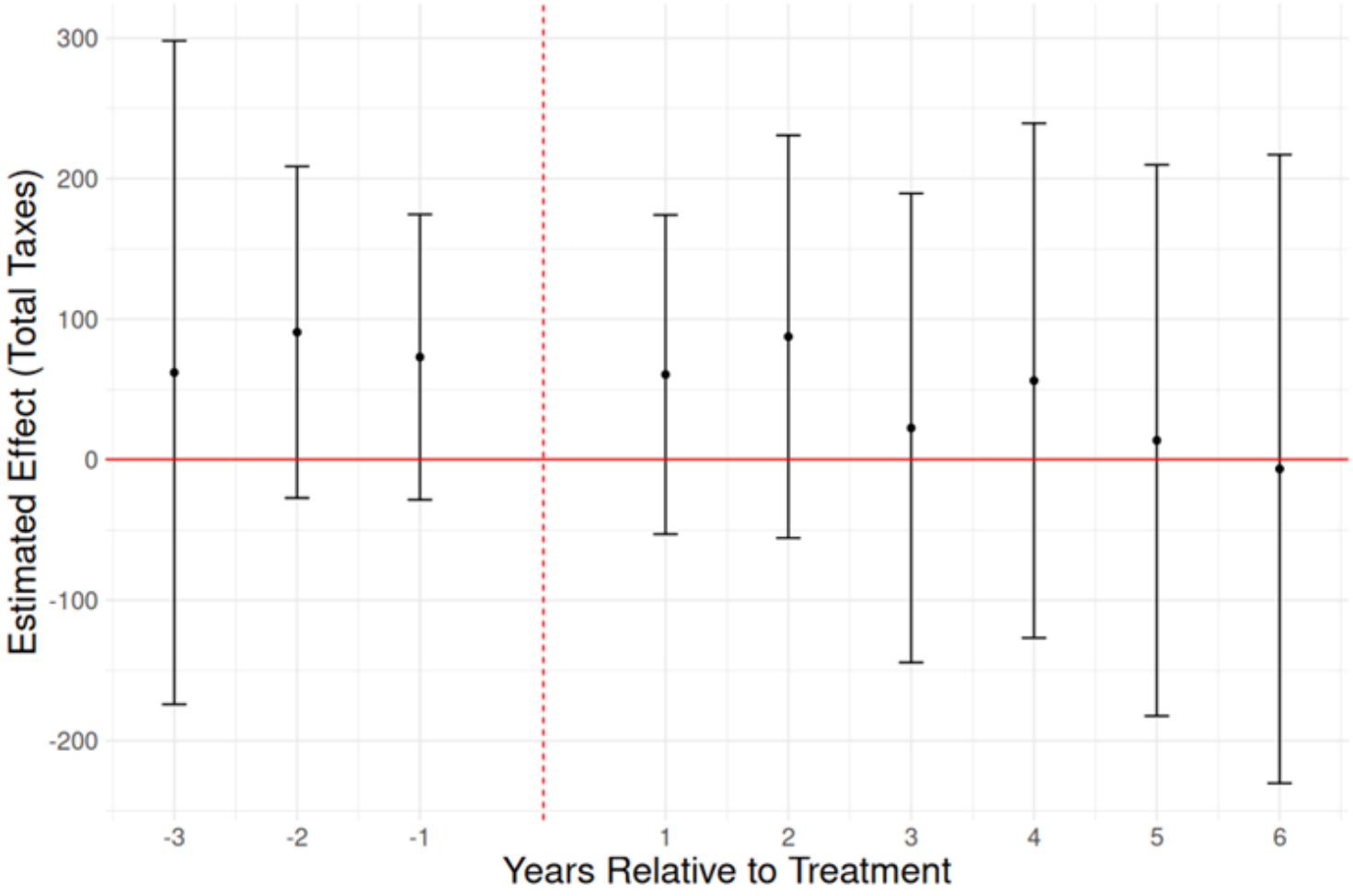
Reduced-Form Estimates (ITT) of AUR on Total Tax



- Pre-treatment trends assumption holds.
- Being significantly more likely be treated (by lying just above the threshold) does not cause changes in the total taxes reported in subsequent years.
- Additional estimates show that these findings also hold within repeater code groups.

Note: The sample includes SEL/UNT Correlation 1 cases from 2013 and 2014 within Calonico (2014) optimal bandwidth around the EPA threshold for whom data for tax years 2010 to 2019 could be found and excludes taxpayers with tax credits above $C in the treatment year and egregious repeaters. Standard errors clustered at the taxpayer level.

IV Estimates (LATE) of AUR on Total Tax



- Consistent with our null ITT effects, we find null LATEs
- Null LATEs also within repeater code groups.
- Interpreted locally, these results indicate that taxpayers in the vicinity of the EPA threshold who did not receive an AUR notice would not have different subsequent tax-reporting behavior if they had received the AUR notice.

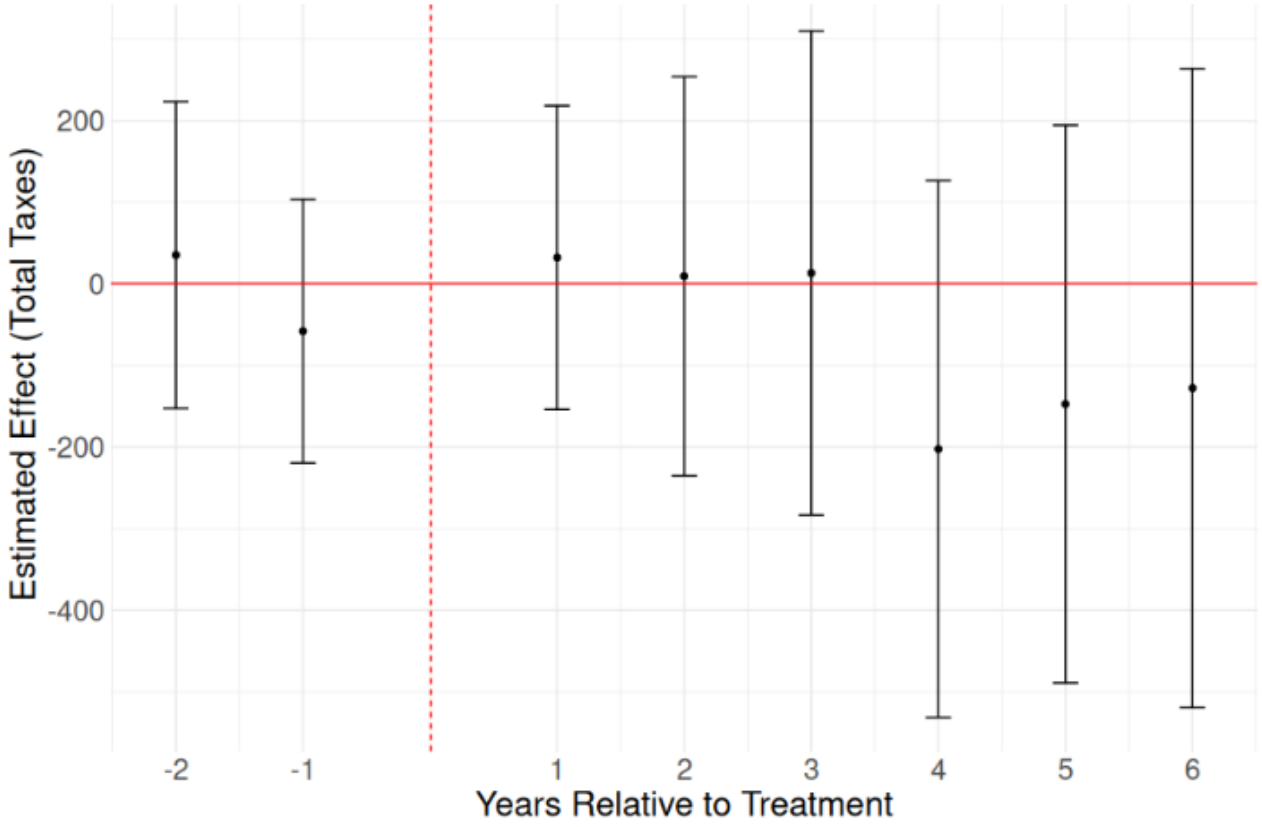
Note: The sample includes SEL/UNT Correlation 1 cases from 2013 and 2014 within Calonico (2014) optimal bandwidth around the EPA threshold for whom data for tax years 2010 to 2019 could be found and excludes taxpayers with tax credits above $C in the treatment year and egregious repeaters. Standard errors clustered at the taxpayer level.

Heterogeneous Effects

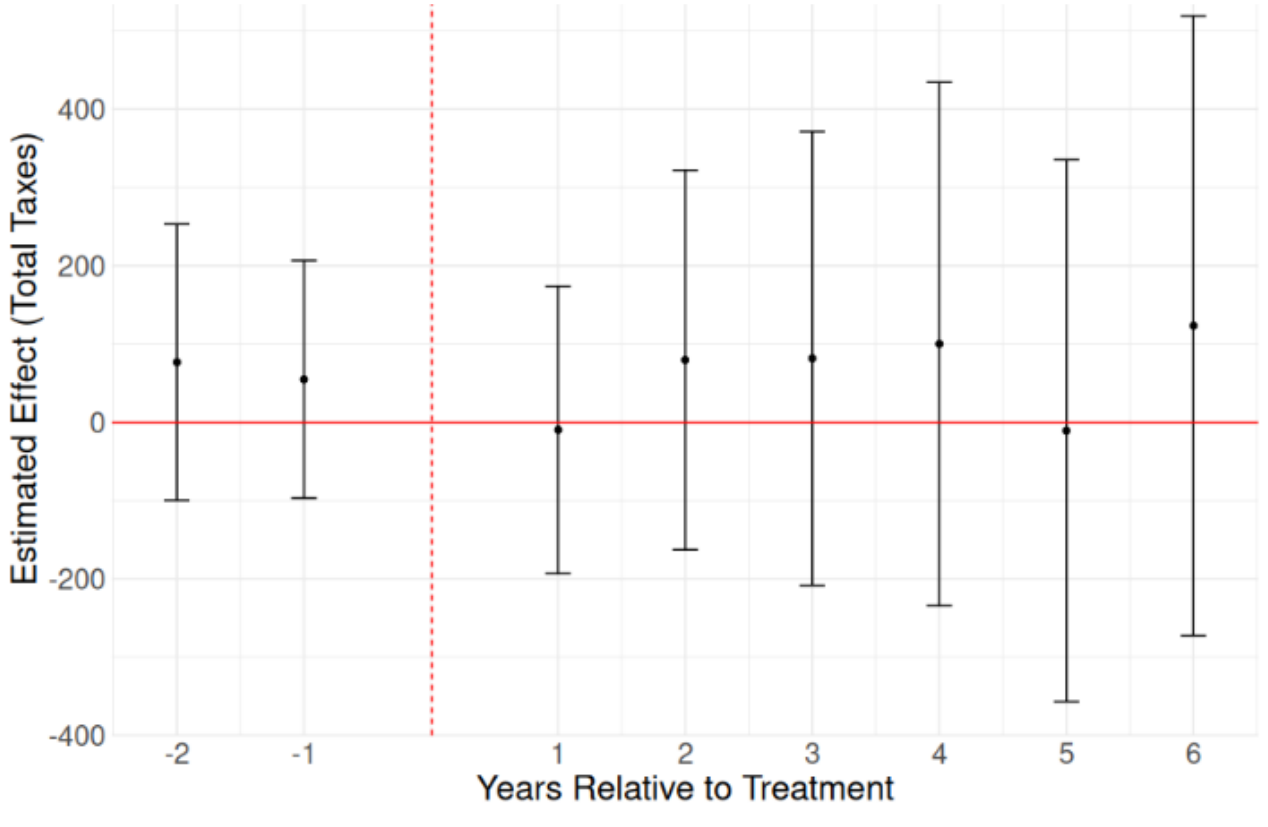
- Our findings of null LATEs on subsequent taxpayer compliance might mask meaningful heterogeneity in behavioral responses.
- Taxpayers who faced larger discrepancies may be more likely to adjust their future reporting behavior.
- We analyze heterogeneous effects two ways:
    1. By the size of the discrepancy identified by the AUR program,
    2. By the amount the taxpayer paid due to AUR (AUR Enforcement Revenue).

Heterogeneous Effects by AUR Discrepancy

AUR Discrepancy Quartile 3



AUR Discrepancy Quartile 4

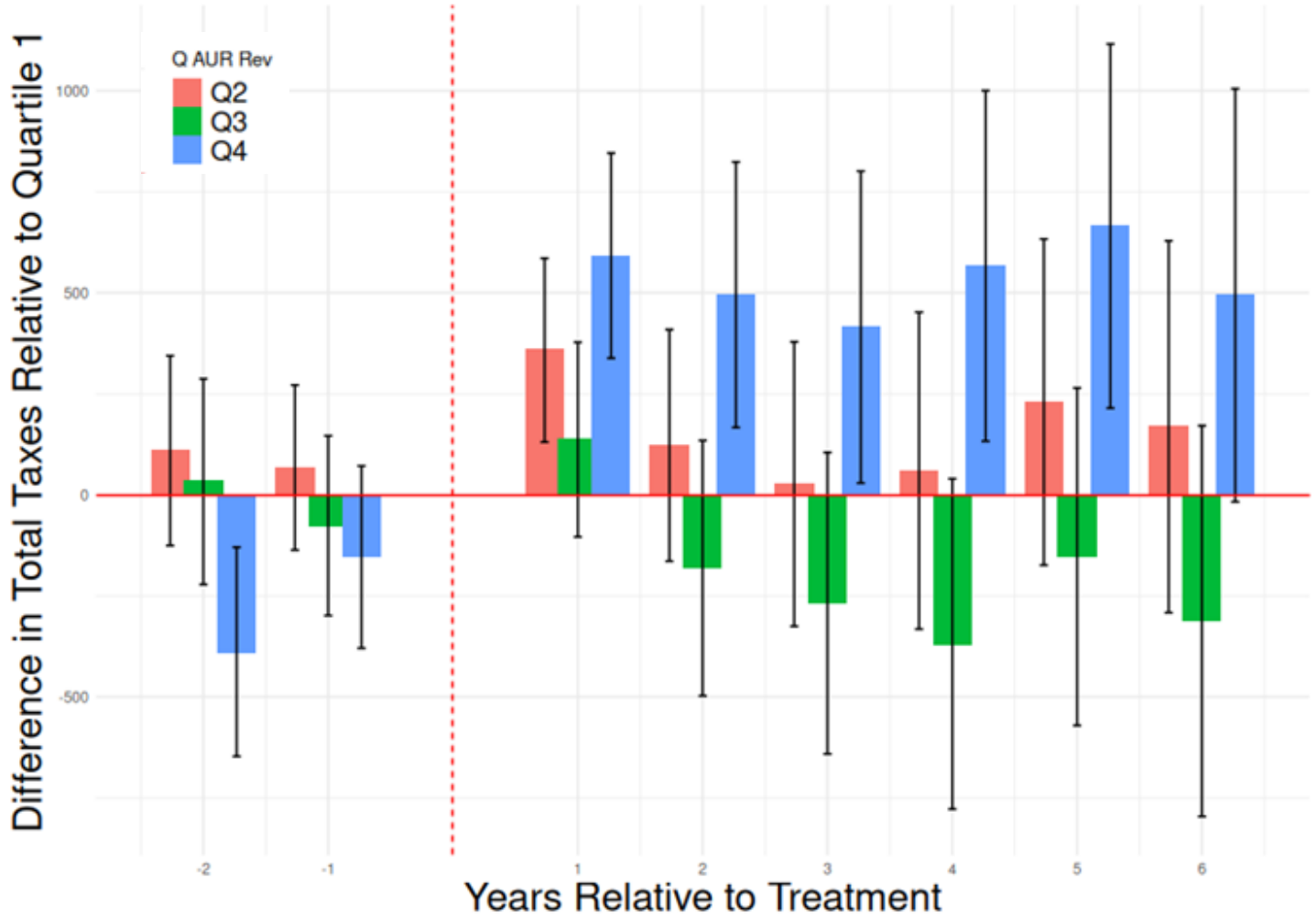


AUR treatment had no significant effect on the future tax compliance, even if they faced large discrepancies.

Heterogeneous Effects by Enforcement Revenue

- We split our treated sample in quartiles of AUR enforcement revenue.
- Two data limitations:
  - Enforcement Revenue is only observable for the treated.
  - No enforcement revenue recorded for cases with EPA < another threshold
- We implement a DiD approach capturing the diff between the total taxes paid by the average taxpayer in quartile Q and the average taxpayer in $Q1$ at $t \neq 0$, relative to the size of that difference at $t = 0$

Heterogeneous Effects by Enforcement Revenue



- AUR's indirect effect on those who had to submit large enforcement payments (mean $947) is significantly larger than the effect experienced by those with smaller payments (mean $144).
- Relative to those in the first quartile of enforcement revenue, those in the top quartile reported a yearly average of $538 more taxes.

Note: SEL Correlation 1 cases from 2013 within optimal bandwidth around the EPA threshold and excludes taxpayers with tax credits above $C in the treatment year and ER. AUR revenue quartiles: Q_1∈(0,384), Q_2∈(384,536), Q_3∈(536,713), and Q_4∈(713,6676). t=0 and Quartile 1 are the omitted categories. Height of bars obtained through a regression interacting time and quartile dummies that include repeater code fixed-effects and a cubic polynomial of EPA. Cases with missing revenue data treated as a separate category in the regression, but estimates are omitted. Standard errors clustered at the taxpayer level

Conclusions

- We estimate the causal effect of W2-related AUR notices on future tax payments
- Despite the scale of the AUR program, the analysis finds no detectable effect overall of receiving a W2-related AUR notification on subsequent tax reporting behavior.
- However, we do find evidence that taxpayers who paid the largest amounts of AUR assessments in the treatment year *did* pay more taxes in subsequent years in relation to taxpayers who had the smallest AUR assessments.
- Both of these findings suggest that “salience” matters:
    1. People appear to take more notice of more costly W2-related assessments
    2. The average size of tax assessments made by AUR’s W2 program is small compared to IRS exam programs, for which we *do* find significant specific indirect effects
- Maybe misreporting of W2s is not subject to the same kind of behavioral dynamics as other forms of noncompliance? Errors related to one-time changes of job?
- Next: Examine later years, other document-matching workstreams.

Thank You

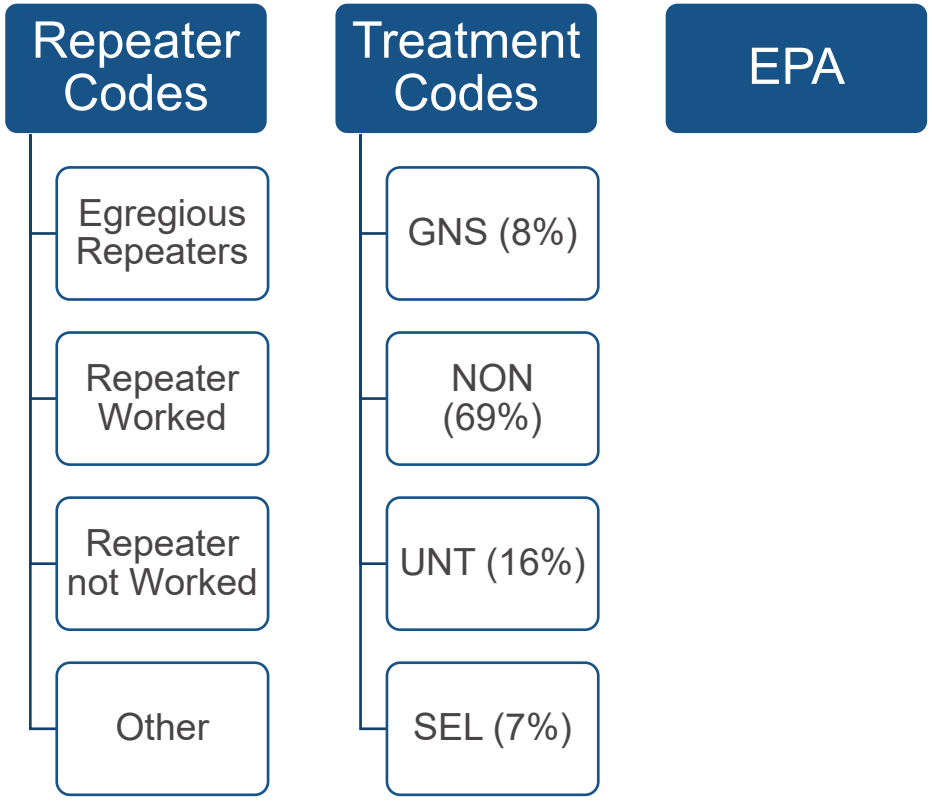
Tom Hertz
IRS-RAAS
Thomas.N.Hertz@irs.gov

Miguel Sarzosa
MITRE
msarzosa@mitre.org

Data

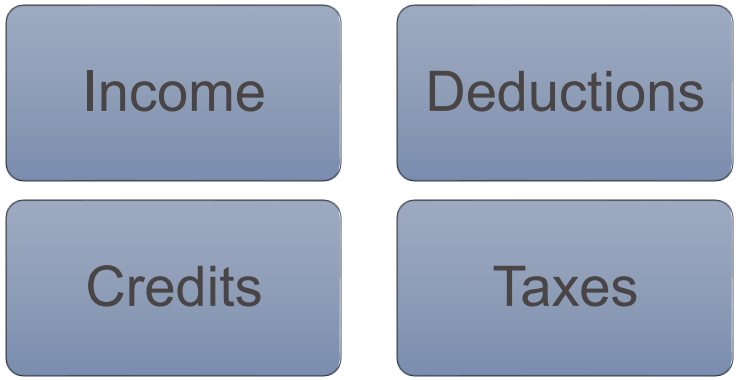
IRS's Business Objectives Enterprise system

Identify AUR eligibility and detect AUR treatment.



IRS's administrative data on F1040s.

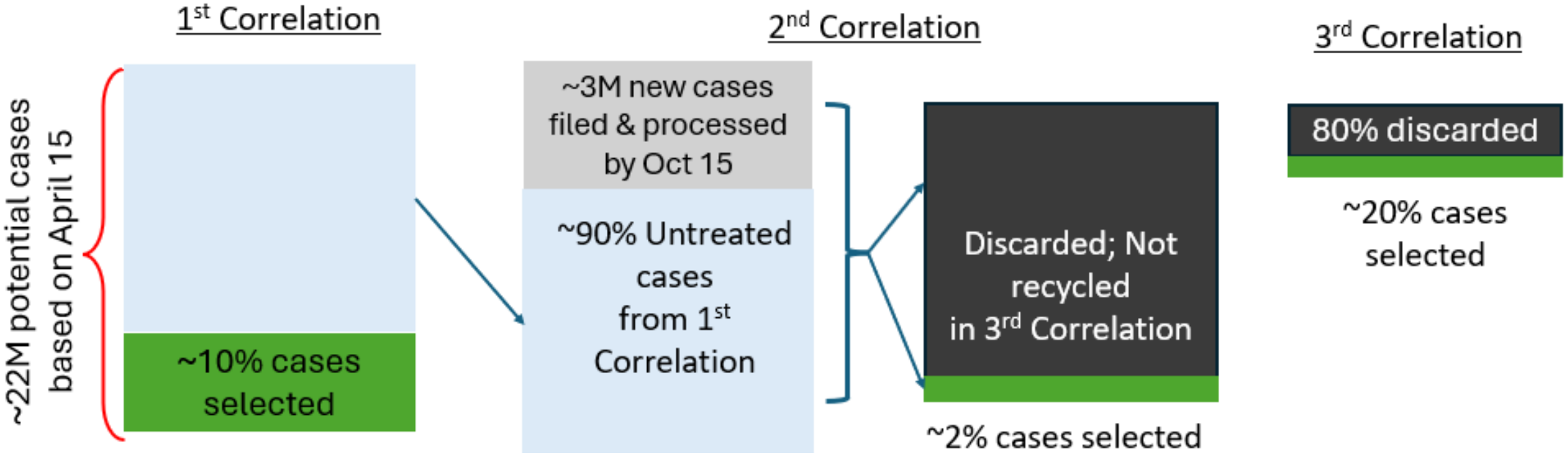
Track all the tax returns of each taxpayer across time. Create a panel data set of reported total tax liability for taxpayer i in year t.



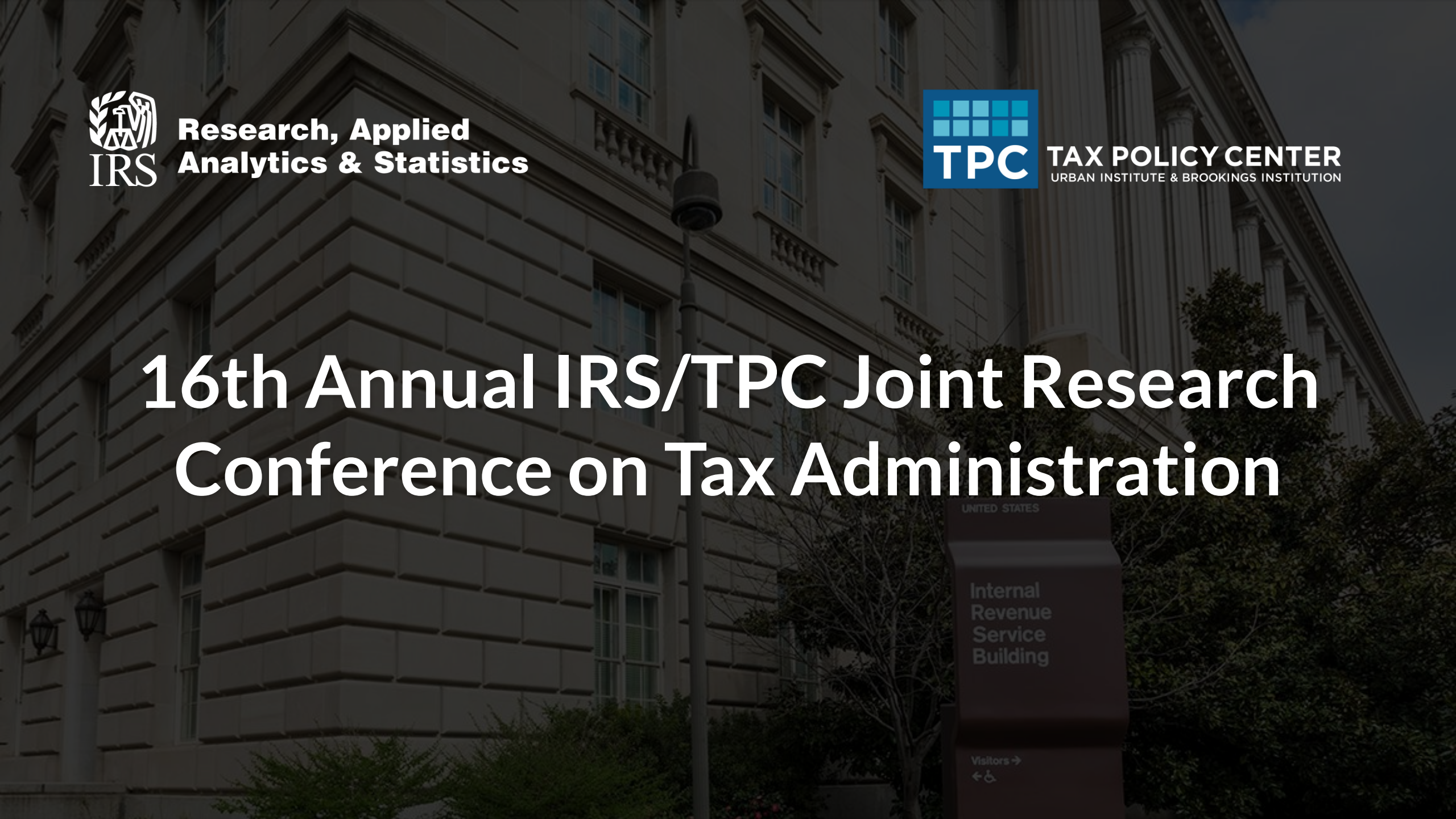
Analysis datasets

- Cases identified by AUR for TY2013 or TY2014
- Follow their tax reporting behavior for 6 years.

W2-Related Case Selection in Batches



- The first batch analyzes returns filed by April 15 yielding ~22 million potential AUR cases. 10% are selected.
- The remaining cases are combined with ~3 million potential cases from those filed by Oct 15. Of them, about 2% are selected. The remaining cases are discarded.
- There is a small third batch of late filers which is not considered in this paper.
- Of 25 million potential AUR cases in a typical year, around 2.5 million are selected



Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

16th Annual IRS/TPC Joint Research Conference on Tax Administration

Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

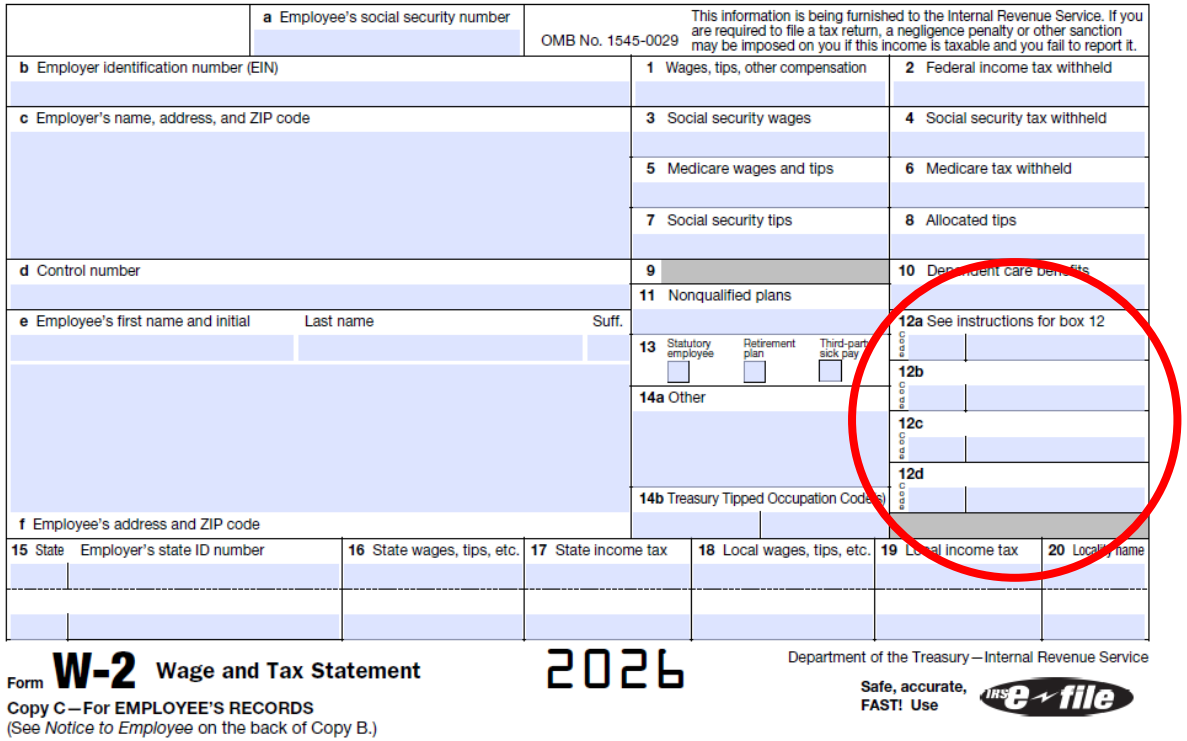
Session 4

June 25, 2026

Unpacking the Box 12: Imputing Deferred Compensation Categories on Population-Level Form W-2 Data Using SOI Transcription and Machine Learning

Rachel Geiger, Derek Gutierrez, Victoria Bryant

Box 12: Elective Deferred Compensation



	a Employee's social security number	OMB No. 1545-0029	This information is being furnished to the Internal Revenue Service. If you are required to file a tax return, a negligence penalty or other sanction may be imposed on you if this income is taxable and you fail to report it.
b Employer identification number (EIN)		1 Wages, tips, other compensation	2 Federal income tax withheld
c Employer's name, address, and ZIP code		3 Social security wages	4 Social security tax withheld
		5 Medicare wages and tips	6 Medicare tax withheld
		7 Social security tips	8 Allocated tips
d Control number		9	10 Dependent care benefits
		11 Nonqualified plans	
e Employee's first name and initial	Last name Suff.	13 Statutory employee / Retirement plan / Third-party sick pay	12a See instructions for box 12
		14a Other	12b
			12c
		14b Treasury Tipped Occupation Code(s)	12d
f Employee's address and ZIP code			

15 State	Employer's state ID number	16 State wages, tips, etc.	17 State income tax	18 Local wages, tips, etc.	19 Local income tax	20 Locality name

Form **W-2** Wage and Tax Statement 2026 Department of the Treasury—Internal Revenue Service

Copy C—For EMPLOYEE'S RECORDS
(See *Notice to Employee* on the back of Copy B.)

Safe, accurate, FAST! Use IRS e-file

Elective Deferral Codes:

D - 401k
E - 403(b)
F - 408 (k)(6) SEP
G - 457(b)
H - 501(c)(18)(D)

Background

Why do tax researchers care about retirement deferral amounts?

Areas of Impact:

- Taxable income analysis
- Federal revenue estimates
- Retirement security
- Compliance research
- Retirement incentive evaluation

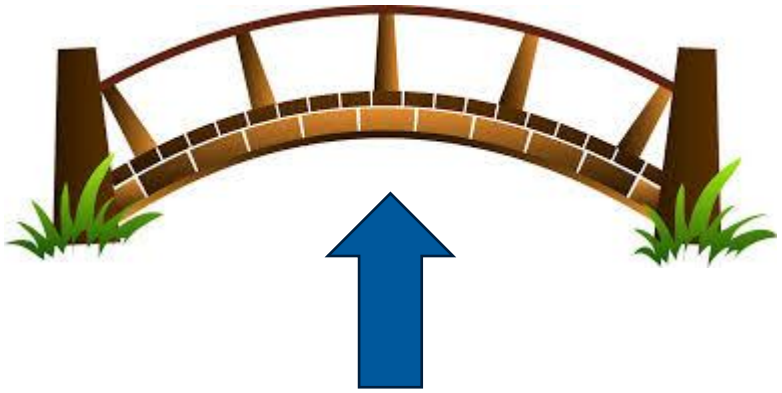
The Research Problem

- Population-level W-2 data that is obtained by the IRS from the Social Security Administration (SSA) is aggregated into a single field during transfer, removing any distinction between codes
- IRS Statistics of Income (SOI) has sample-level data where these Box 12 codes are observed but is not population-level
- Can we accurately impute these values? Yes.

Leveraging a “Bridge”

Box 12 Sample-Level Data from IRS SOI

Issue: It’s only sample-level



Use machine learning to train a model using the sample and apply it to population data to impute values and codes

Box 12 Population-Level Data from SSA

Issue: All codes aggregated into a single amount during data transfer removing any distinctions between codes

Data

- SOI's sample-level data is a merged dataset of SOI's INSOLE file + SOI's Information Returns Processing (IRP) W-2 data
- After merging the datasets using SOI record identifiers, the resulting file contains 522,699 observations and 42 variables, allowing direct comparison between detailed and aggregated deferred compensation information.
- The merged dataset (COMB W-2) was reduced to a modeling dataset consisting of key W-2 variables
- Following filtering and preprocessing, including restricting records to Box 12 codes D–H and applying one-hot encoding, the final analytical dataset contains 132,071 observations and 23 variables
- 70/30 training/test split was applied to the derived dataset where the training and testing datasets consists of 92,449 and 39,688 rows

Methodology - Model Exploration

- CART (Classification and Regression Trees)
  - Caret (R)
  - Tidymodels (R)
  - Sci-kit (Python)
- GBM (Gradient Boosting Model)
  - Caret (R)
  - Gbm (R)
  - Sci-kit (Python)

Baseline Performance

- Measured with RMSE which is the average difference between the model's predicted and actual values

Table 1. Baseline Performance						
Algorithm	CART			GBM		
Package/Module	Caret	Tidymodels	Sci-kit	Caret	Gbm	Sci-kit
Test RMSE	$11,911.49	$5,595.66	$5,770.09	$10,816.37	$10,908.39	$10,806.75

Understanding the Baseline Error

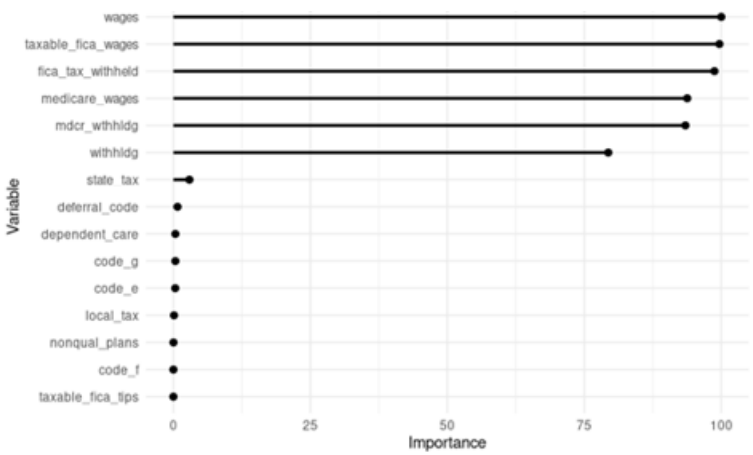
Bias Variance Decomposition – Breaks down a machine learning model's prediction error into squared bias, variance, and irreducible error.

(Estimated*) Bias-Variance Decomposition						
Algorithm	CART			GBM		
Package/Module	Caret	Tidymodels	Sci-kit	Caret	Gbm	Sci-kit
Squared Bias	140,424,745	19,540,393	25,635,827	118,507,114	118,585,616	115,371,903
Variance	1,574,993	11,849,873	17,980,324	964,213	332,492	1,117,745
MSE	141,999,738	31,390,266	43,616,151	119,471,327	118,918,108	116,489,648

*Bias-Variance Decomposition values derived from bootstrapping

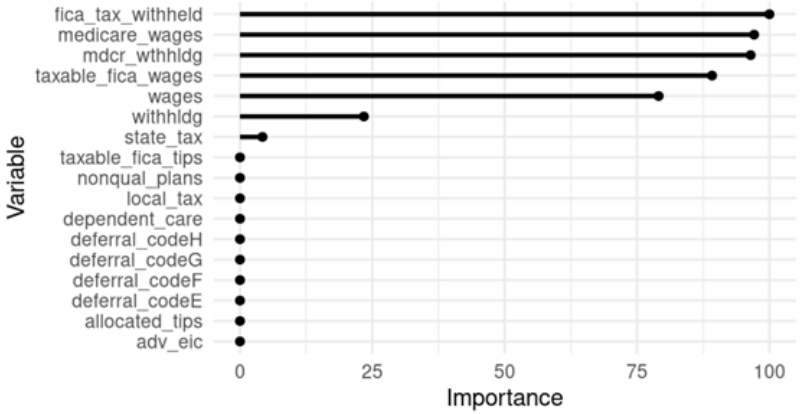
What predicts retirement deferrals?

Variable Importance Plots for CART Models



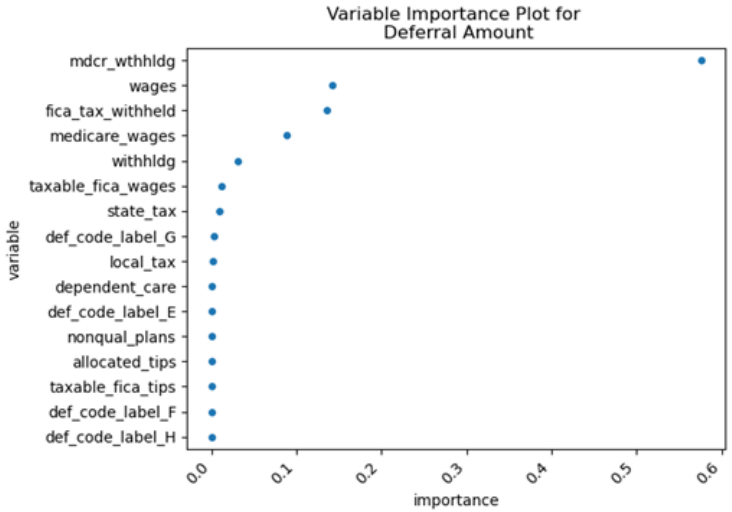
RStudio: Tidymodels

1. Wages
2. Taxable FICA wages
3. FICA tax withheld
4. Medicare wages
5. Withholding



RStudio: Caret

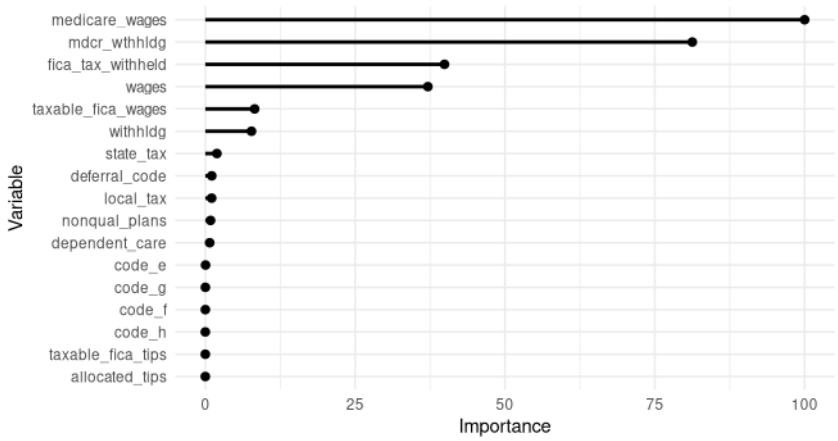
1. FICA tax withheld
2. Medicare wages
3. Medicare withholding
4. Taxable FICA wages
5. Wages

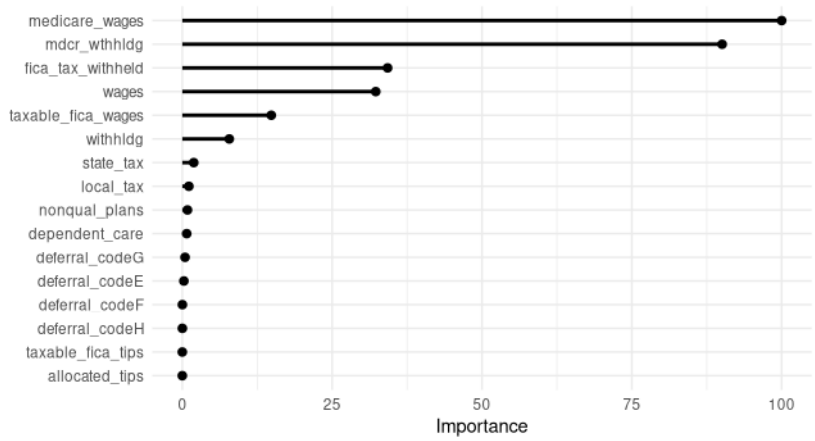


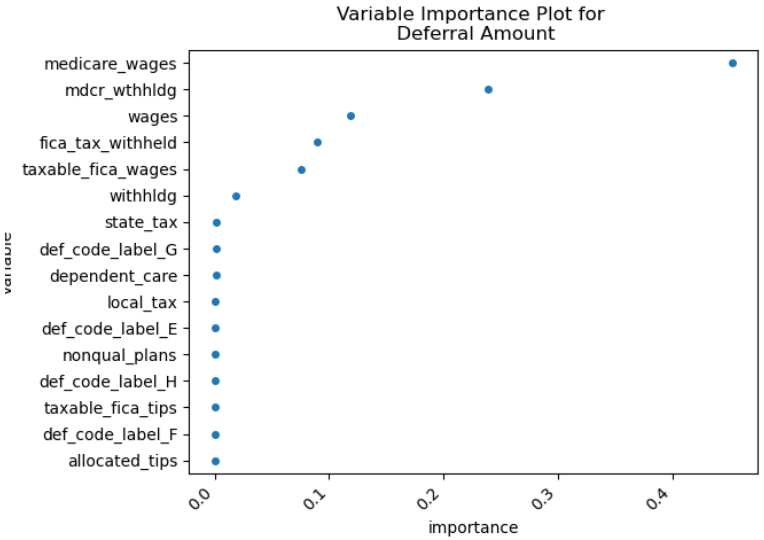
Python: Sci-kit

1. Medicare withholding
2. Wages
3. FICA tax withheld
4. Medicare wages
5. Withholding

Variable Importance Plots for GBM Models







RStudio: GBM

1. Medicare wages
2. Medicare withholding
3. FICA tax withheld
4. Wages
5. Taxable FICA wages

RStudio: Caret

1. Medicare wages
2. Medicare withholding
3. FICA tax withheld
4. Wages
5. Taxable FICA wages

Python: Sci-kit

1. Medicare wages
2. Medicare withholding
3. Wages
4. FICA tax withheld
5. Taxable FICA wages

Hyperparameter Tuning the Model

Parameters tuned:

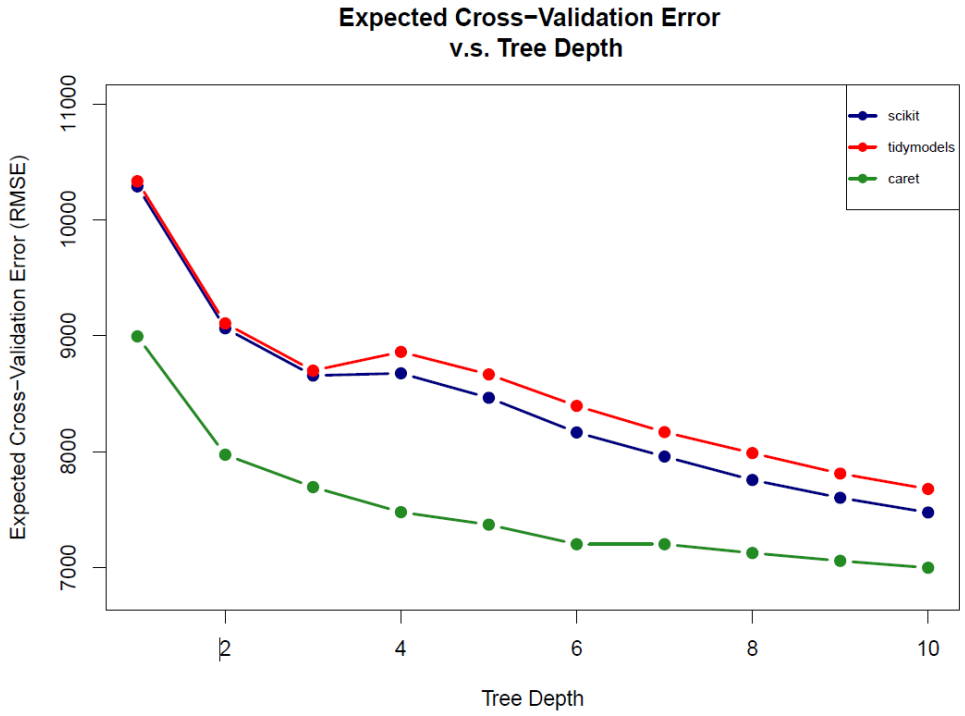
CART

- Tree depth
- Terminal node size

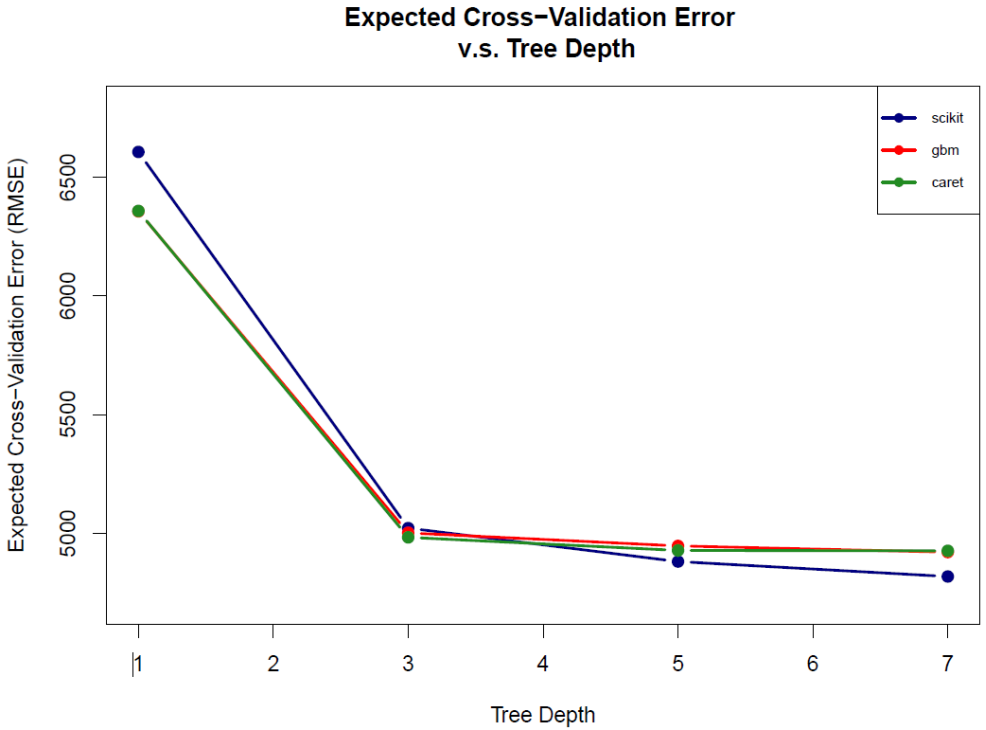
GBM

- Number of trees
- Tree depth
- Learning rate
- Terminal node size

Expected Cross-Validation Error vs. Tree Depth



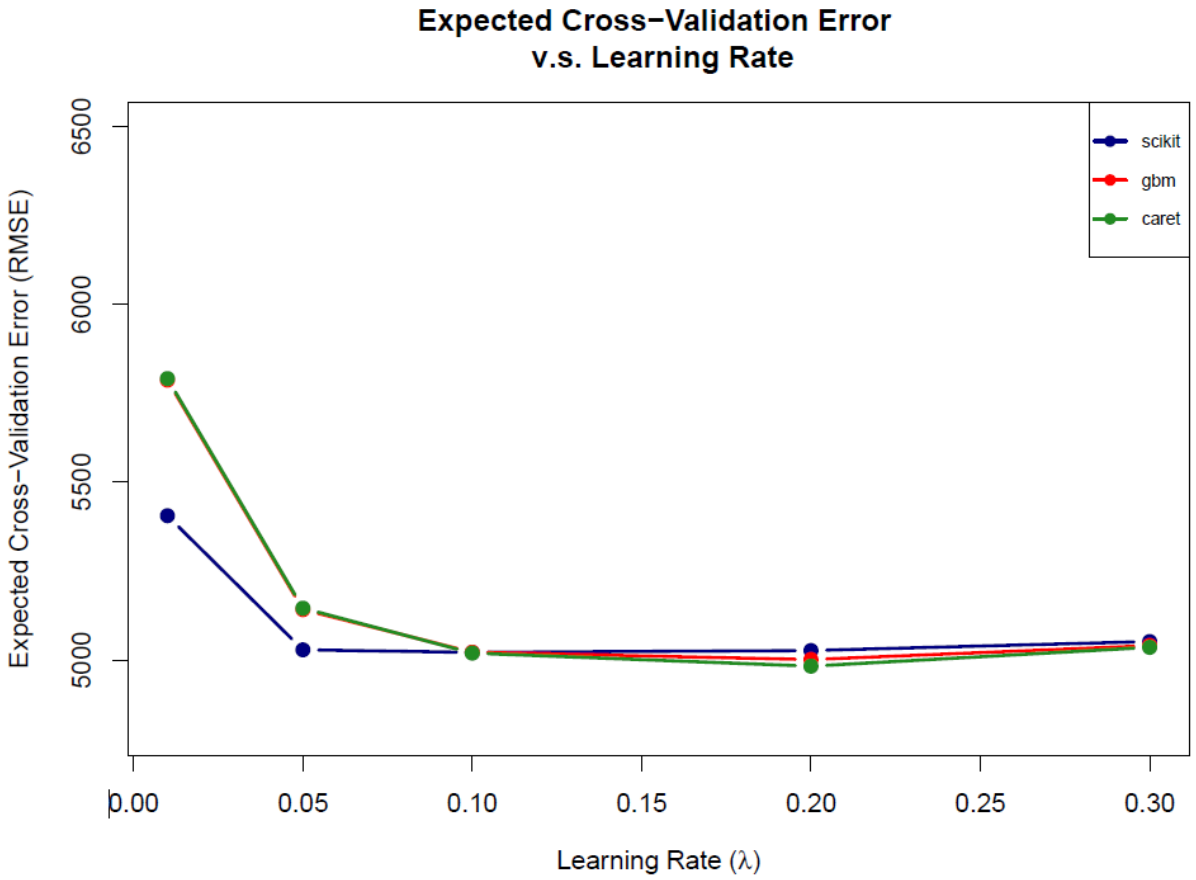
CART Model



GBM Model

GBM Learning Rate Results

- The scikit package has a lower cross-validation error initially but conjoins with gbm and caret when the learning rate is 0.1.
- Higher learning rate leads to a higher risk of overfitting because it requires fewer trees to converge
- There is a tradeoff between a faster and higher learning rate and a lower fine-grained learning rate



Final Model Performance

Table 2. Final Model Performance						
Algorithm	CART			GBM		
Package/Module	Caret	Tidymodels	Sci-kit	Caret	Gbm	Sci-kit
Test RMSE	$12,127.67	$7,452.84	$7,746.27	$11,477.86	$10,841.85	$10,691.53
(Estimated*) Bias-Variance Decomposition						
Algorithm	CART			GBM		
Package/Module	Caret	Tidymodels	Sci-kit	Caret	Gbm	Sci-kit
Squared Bias	146,645,324	55,379,162	59,619,105	131,889,524	116,224,678	114,136,557
Variance	596,177	1,557,897	392,398	295,401	1,185,090	1,264,772
MSE	147,241,501	56,937,059	60,011,503	132,184,926	117,409,768	115,401,329

*Bias-Variance Decomposition values derived from bootstrapping

Applying The Model to the Data

- Each model is then applied to the population-level Box 12 data to get the following results
- Discrepancies between model types are mostly due to sampling differences between programming packages

	CART			GBM			
	Amount			Amount			
Deferral Code	Tidymodels	Caret	Sci-kit	Caret	gbm	Sci-kit	SOI Aggregated Population Estimates
D	343,717,283	343,168,253	351,482,719	276,671,044	275,489,600	274,328,467	259,466,350
E	50,841,504	50,743,617	51,451,361	39,009,650	39,084,689	39,577,262	38,344,871
F	337,764	331,491	330,951	325,662	319,453	299,118	280,519
G	26,408,794	26,334,233	26,524,494	18,048,805	18,022,616	18,508,968	17,906,253
H	87,407	87,675	87,675	11,808	13,736	47,843	19,390

Note: Amounts are in thousands of dollars

Next Steps

- This is an ongoing study that will be applied to tax years beyond 2019
- Refine predictor selection
- Complete final model comparison
- Validate model across additional tax years
- Impute deferral codes

Tax Administration Application

- New analytical possibilities with distinctions between deferral codes at the population level
- Potential uses for:
  - Compliance studies
  - Revenue estimation
  - Distributional analysis
  - Retirement policy evaluation

Takeaways

- We can accurately impute W-2 Box 12 data using machine learning and will apply this knowledge to future tax years to recover disaggregated deferred compensation amounts
- There is variation between using different R studio packages and coding languages, comparison is included in the study
- This model is applied to 2019 data and can be applied to additional years for further deferred compensation analysis

Thank you!

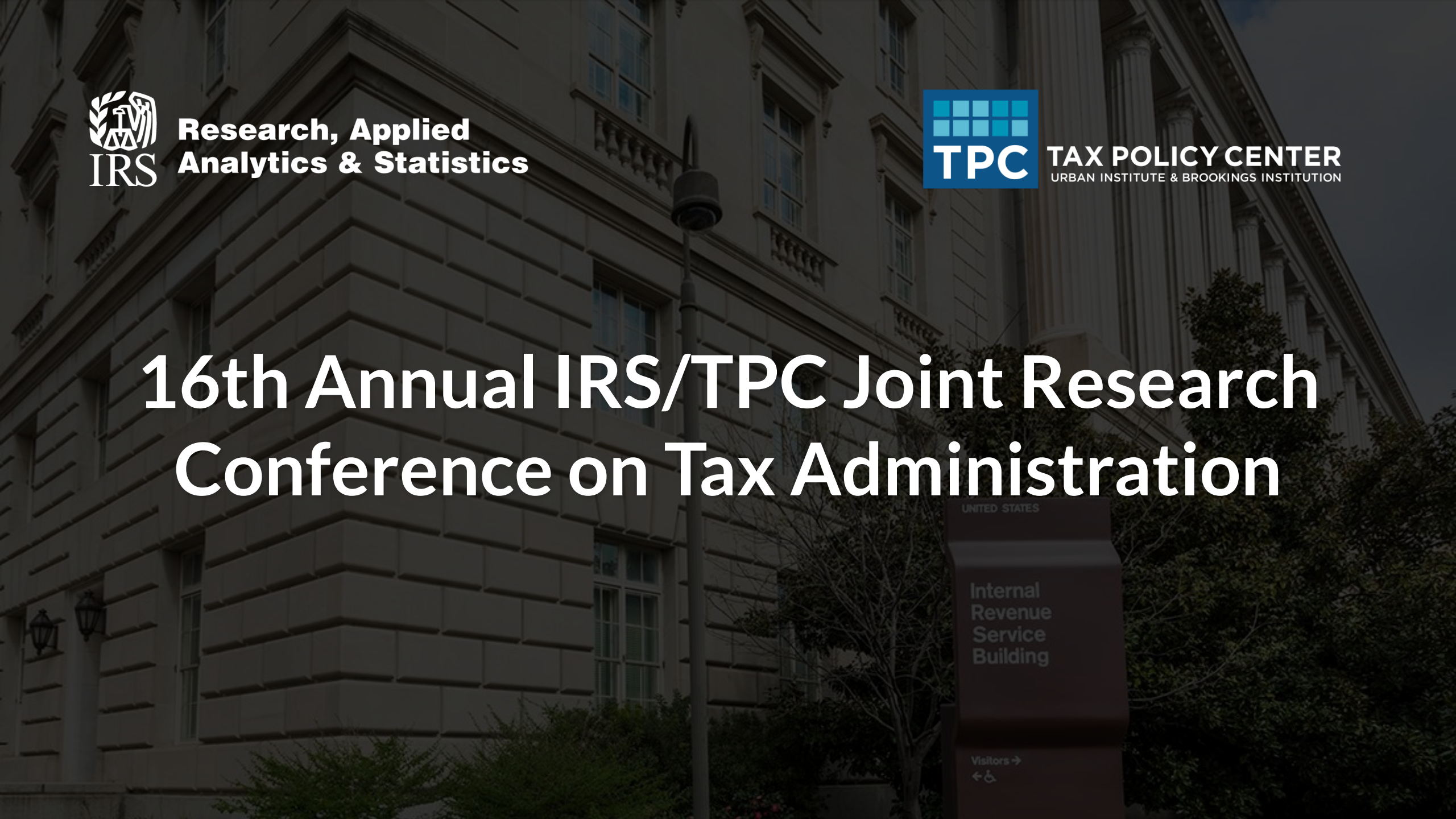
Contact:

Rachel Geiger

rachel.p.geiger@irs.gov

Derek Gutierrez

derek.a.gutierrez@irs.gov



Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

16th Annual IRS/TPC Joint Research Conference on Tax Administration

Wage Misreporting on Individual Income Tax Returns: Evidence from W-2 Mismatches

Will Boning
Lucas Goodman
Emily Lin

June 2026

Is Information Reporting Enough?

- Tax gap studies: 1% net under-reporting rate for wages
- This paper is about that 1% (about $100 billion):
  - Why doesn't information reporting prevent it?
  - Has it changed?
  - How much could it cost (before enforcement)?
- Out of scope: wages paid under-the-table (no W-2)

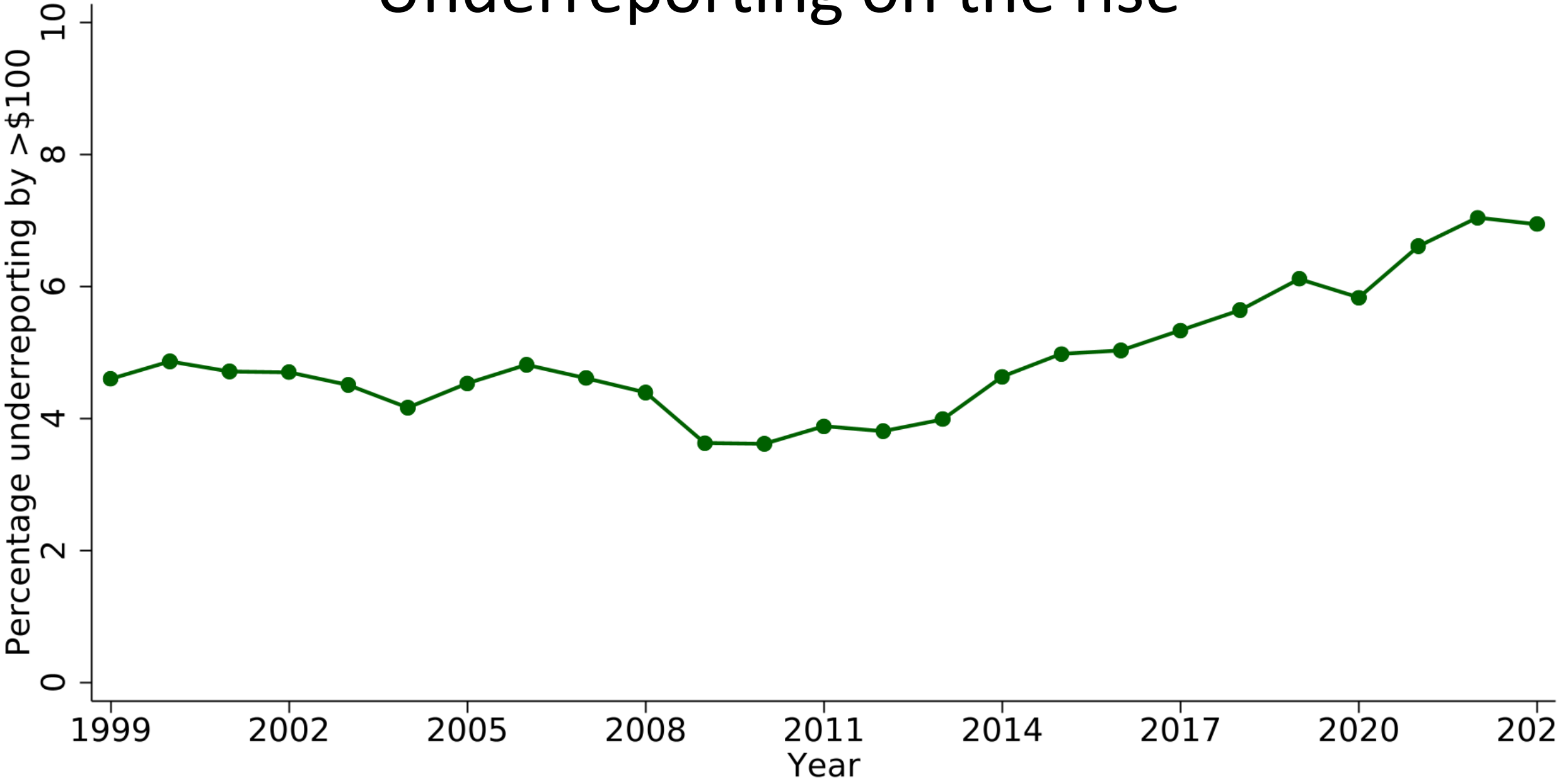
Data and Measurement

- IRS admin data (1999–2023, focal sample 2022-2023)
- Match 1040 wages to W-2s employers send to SSA, forwarded to IRS
  - 1040 universe: primary (and if applicable secondary) have SSNs.
- Where needed: limit to e-filed returns + “attached” W-2s

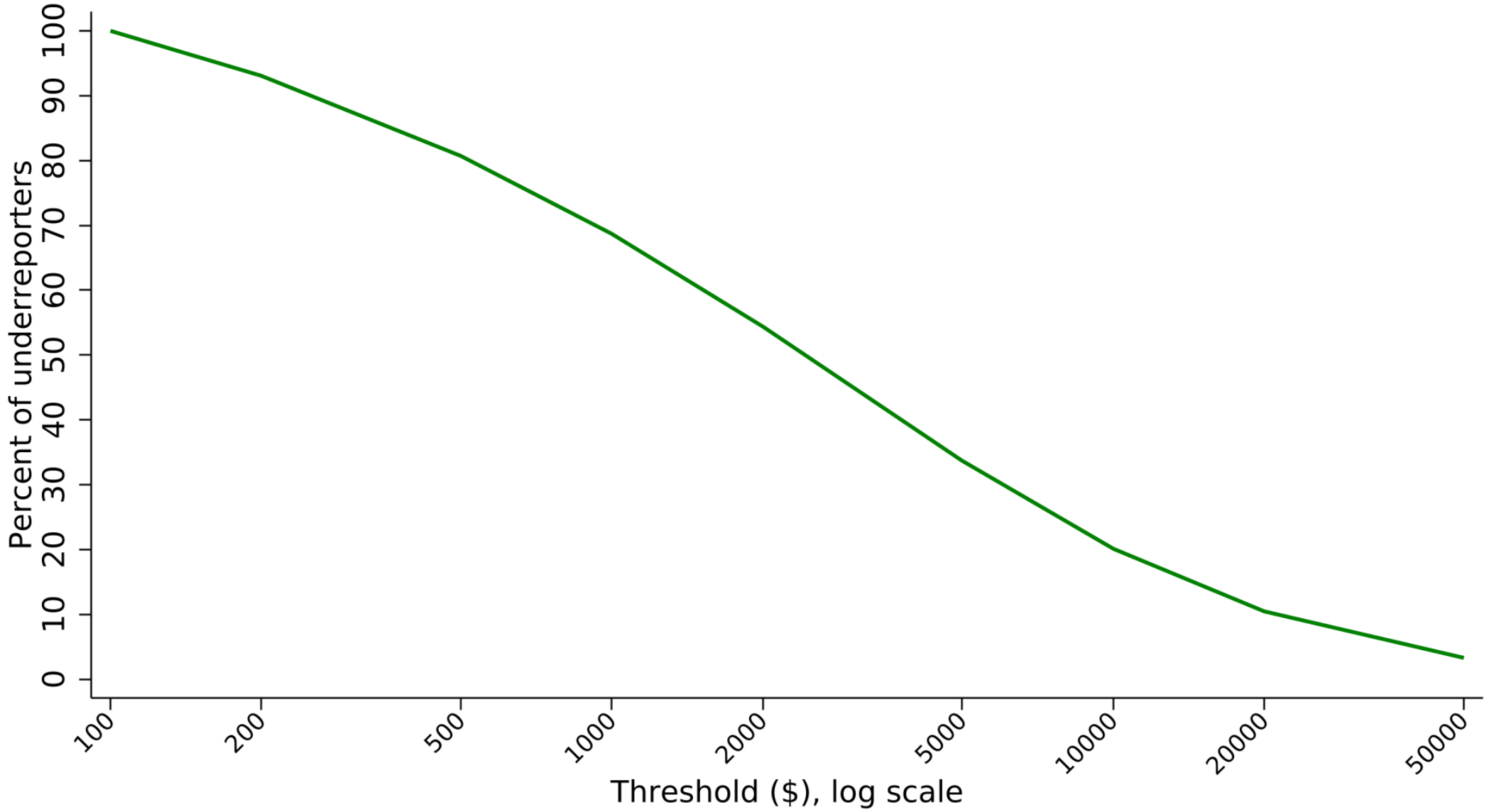
Data and Measurement

- Form 1040:
  - 2022 onwards: Line 1a contains W-2 wages.
  - Other wage-like items on Lines 1b-1h, summed to Line 1.
- Underreporting: 1040 Line 1 < W-2 – $100
- Overreporting: 1040 Line 1a > W-2 + $100
- “Conservative” definition of W-2 wages in case of duplicates

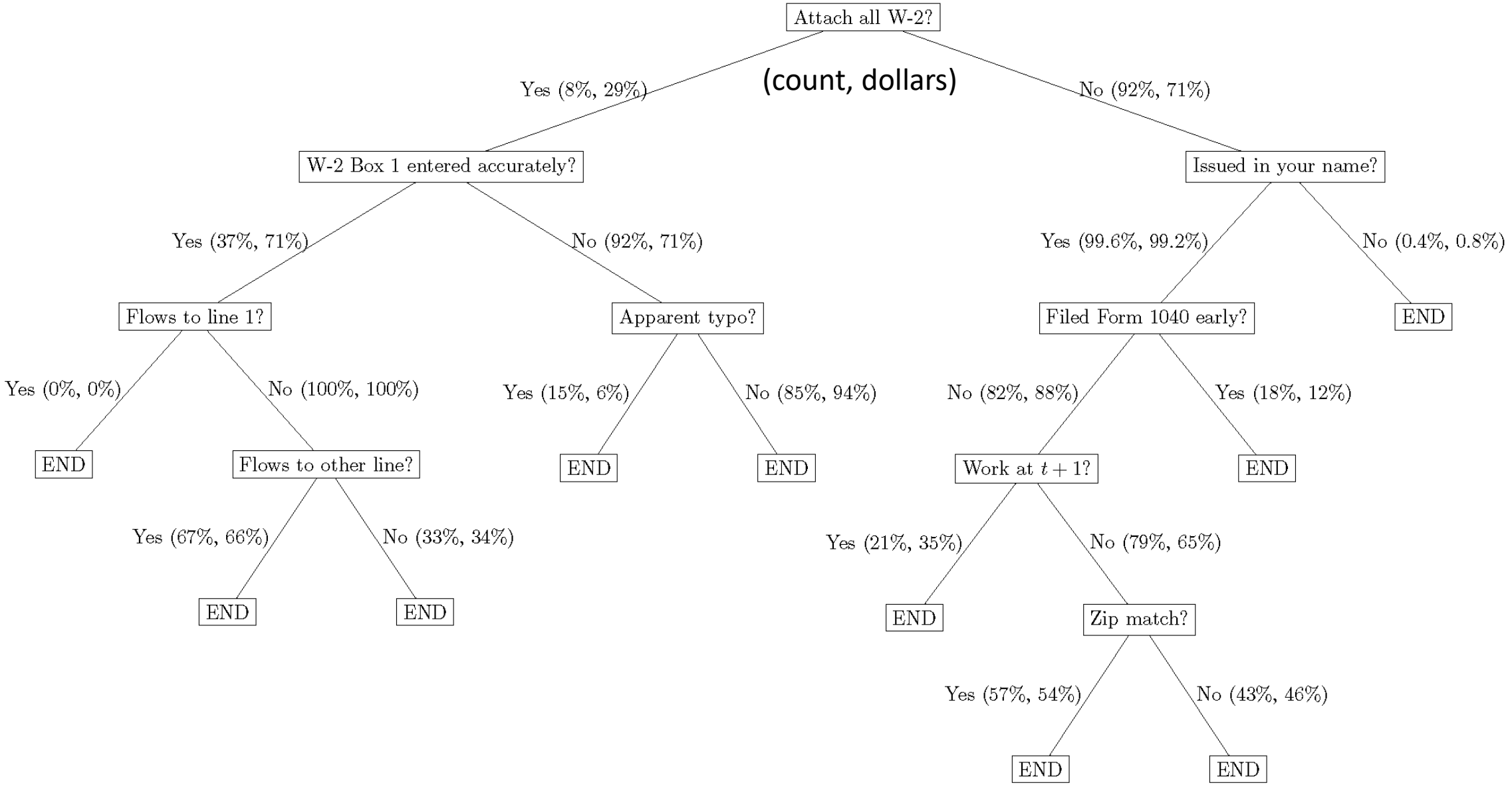
Underreporting on the rise



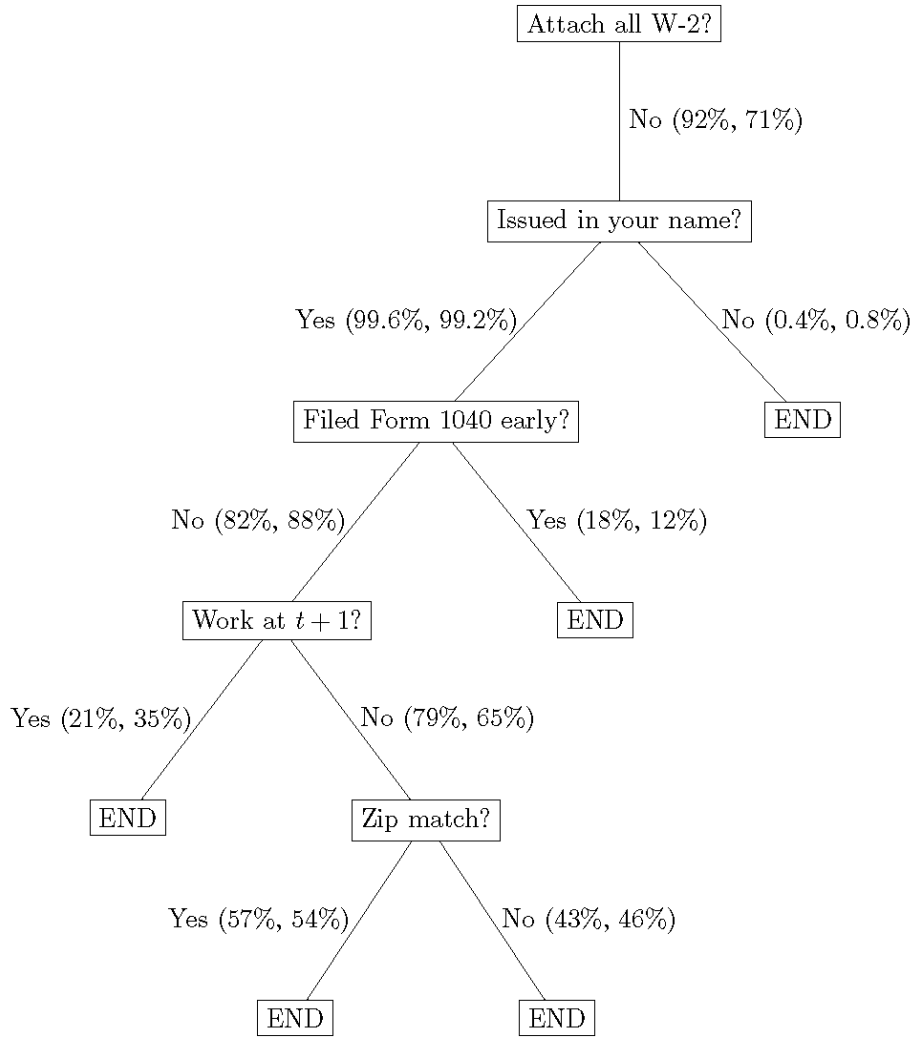
Mostly small amounts, but a long tail



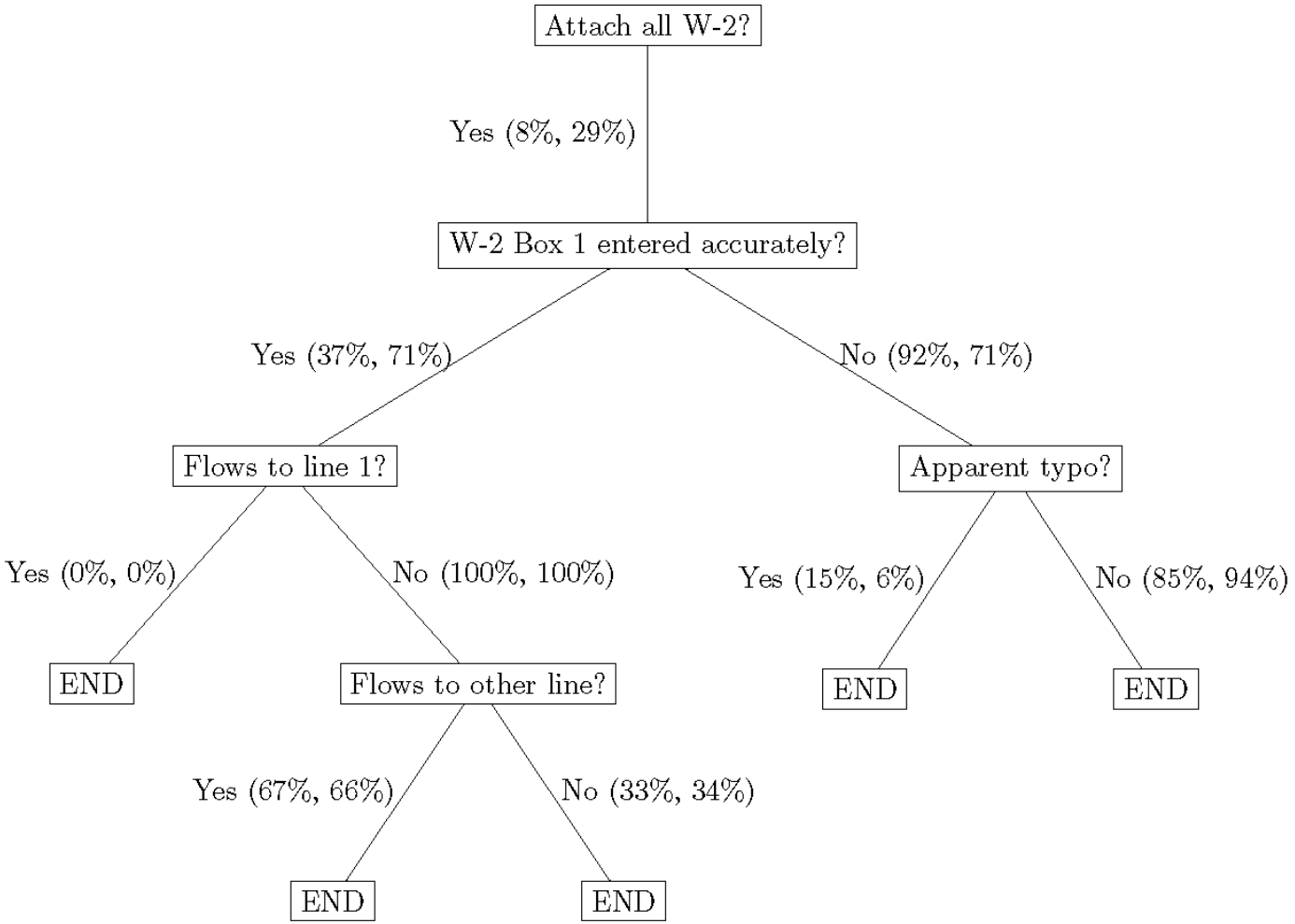
What explains underreporting?



What explains underreporting (right half)?



What explains underreporting (left half)?

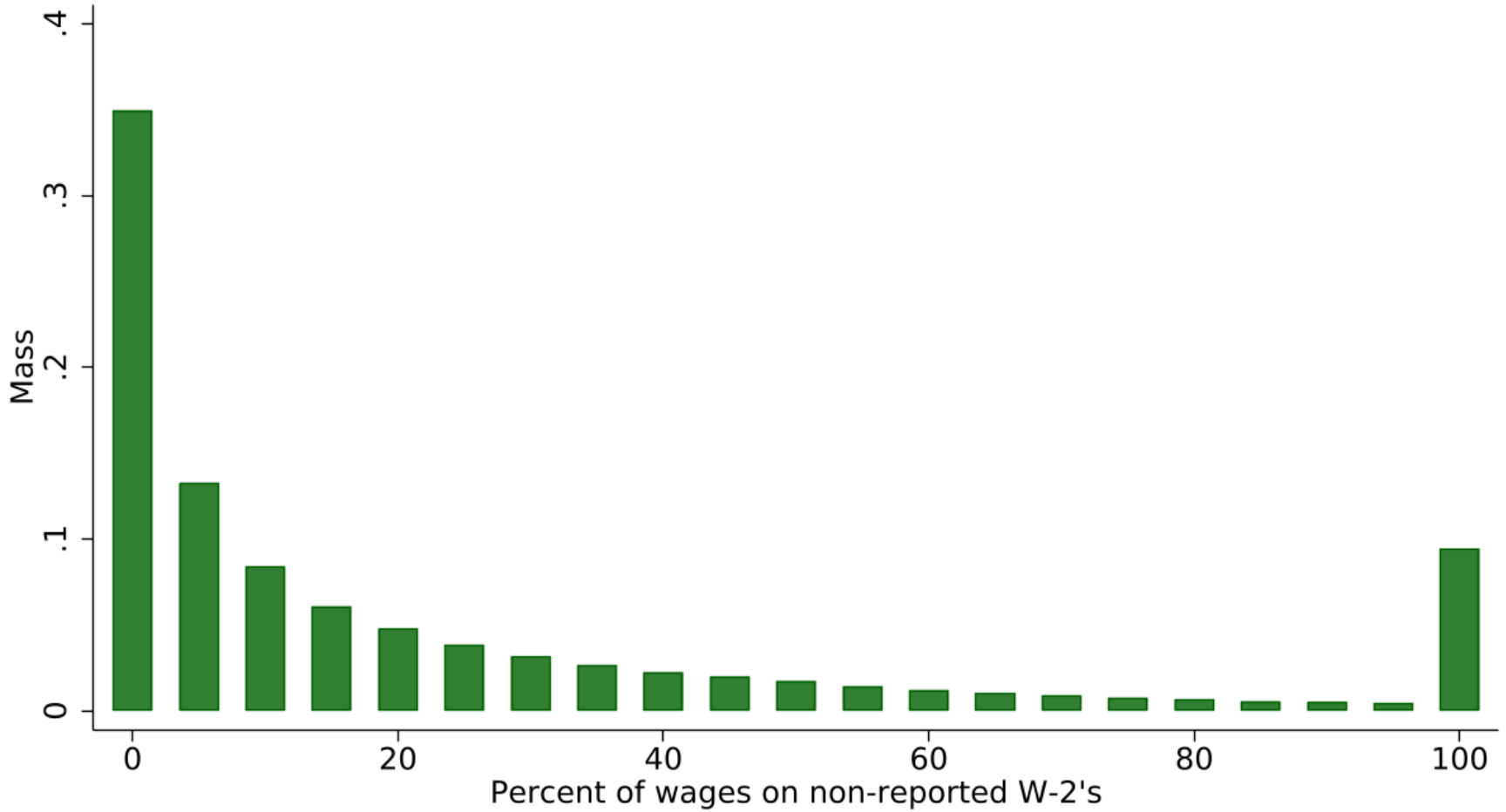


92% of underreporters fail to attach at least one W-2. Why?

- File early or W-2 issued late
- Separated from employer and/or changed address
- Small/secondary job
- Other explanations?

Secondary Jobs

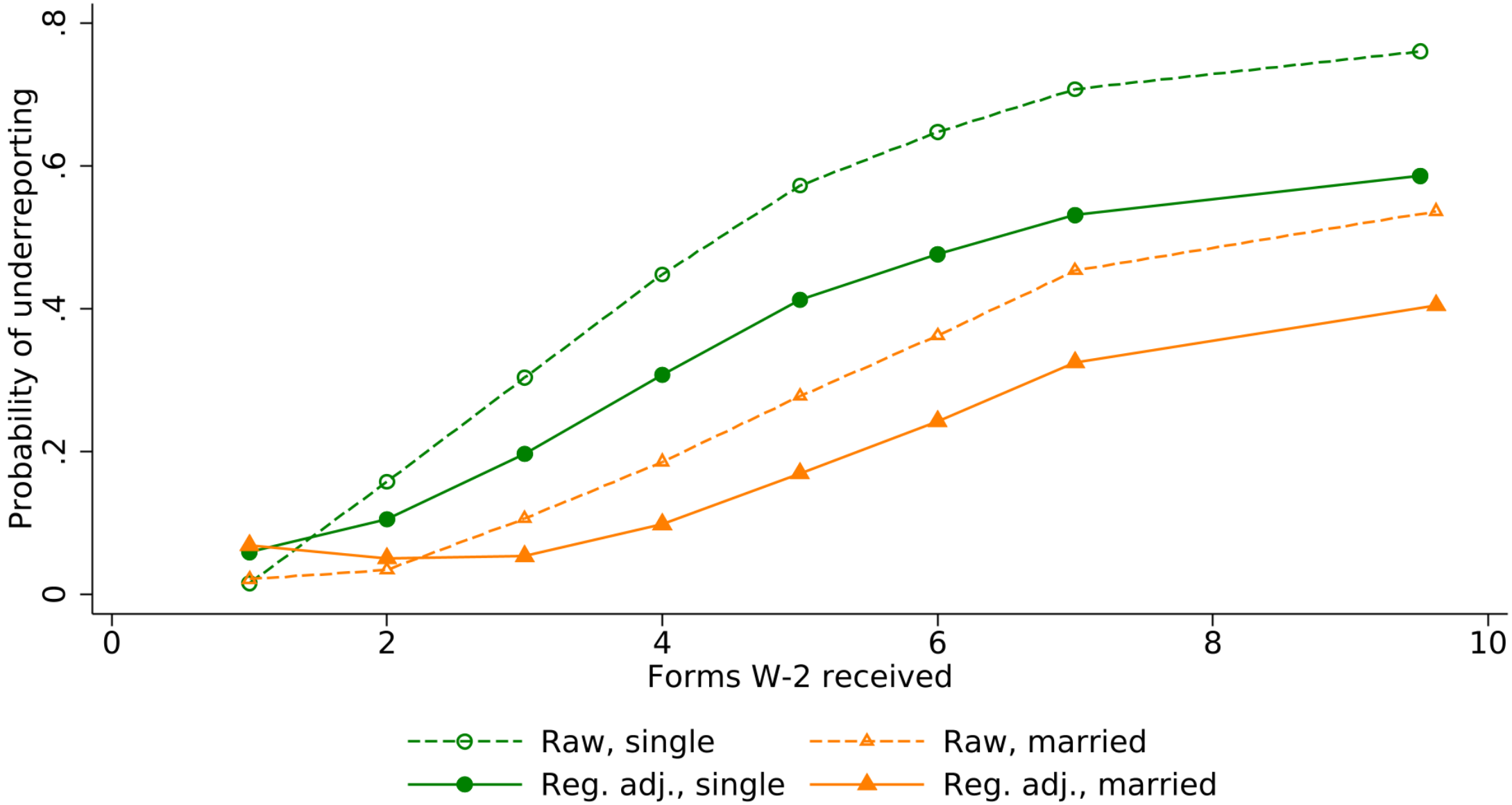
- Omitted W-2s are mostly for secondary jobs
- Most underreporters omit <20% of wages; some omit them all



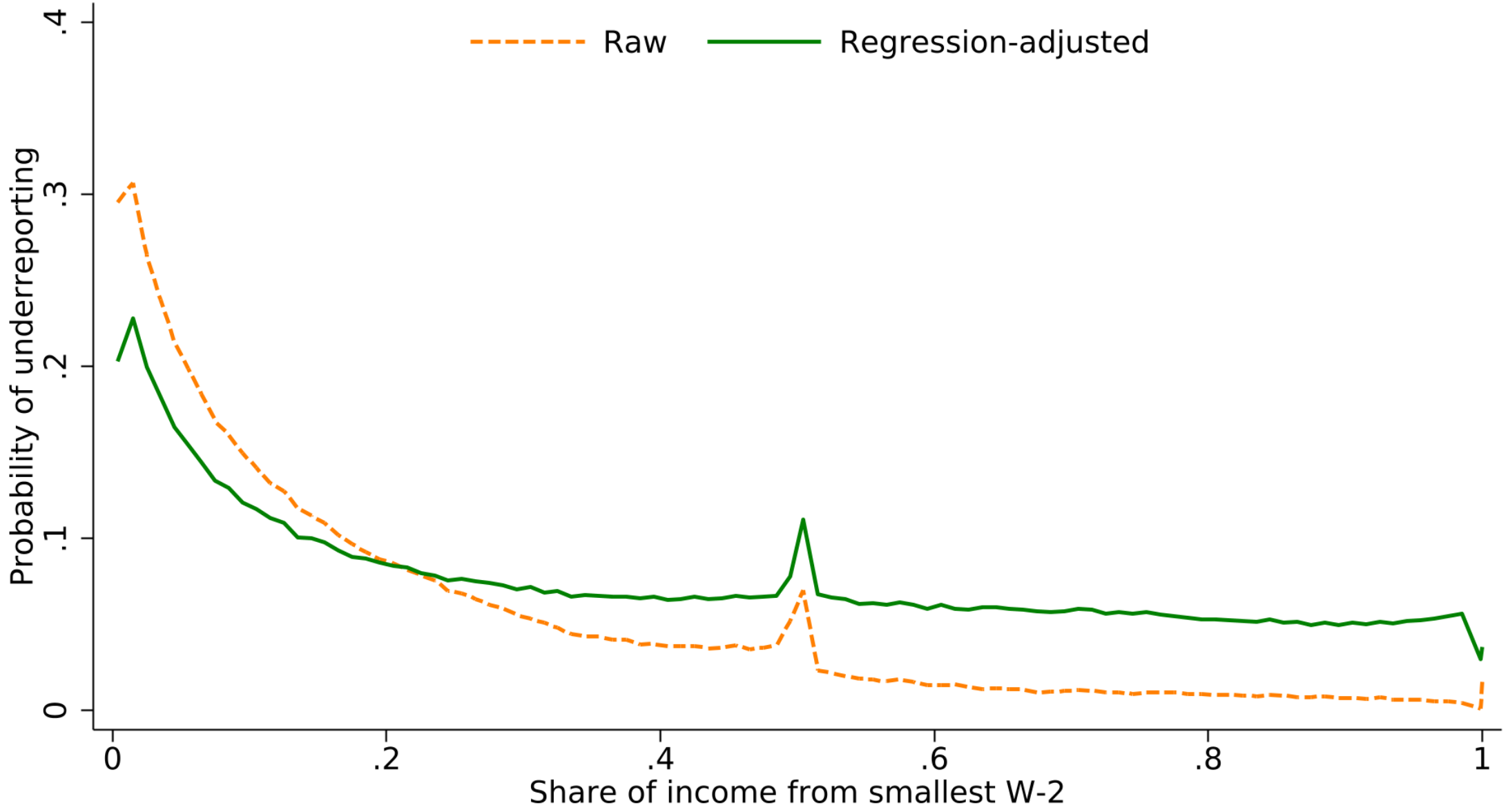
Predicting underreporting

- We regress an indictor for underreporting on:
  - Fixed effects for:
    - **W-2 count x joint filing status**
    - **Size of smallest W-2**
    - Week of filing
    - Income bins
    - Age bins
  - Lower-dimensional covariates:
    - Filed in prior year, Moved, Changed jobs, Moved and changed jobs, Any interest/dividends/capital gains, E-filed, Used paid preparer, Years of post-secondary education, Received a duplicate W-2, Worked at an employer that sent many duplicate W-2's.

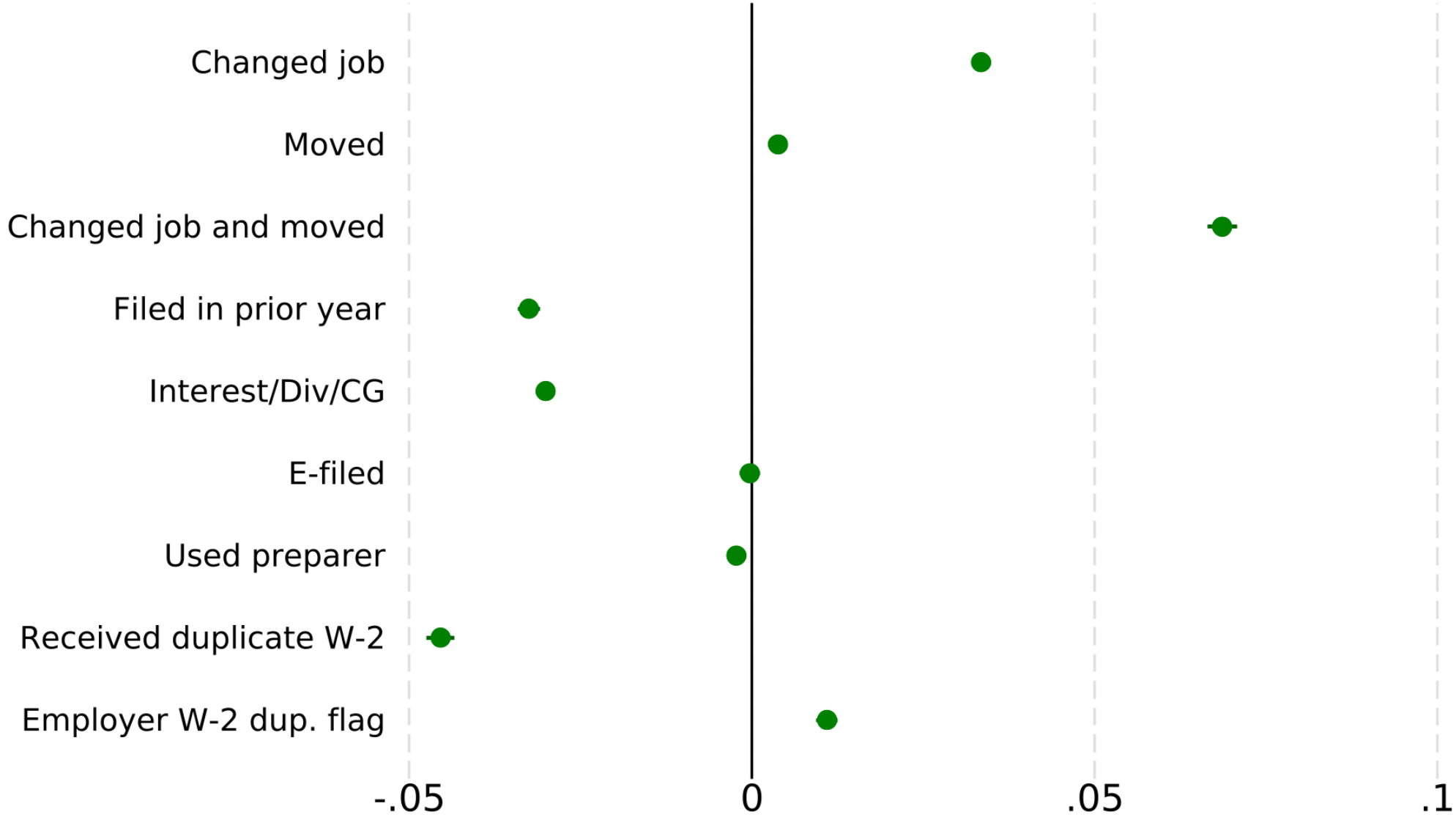
Predictors: Number of jobs



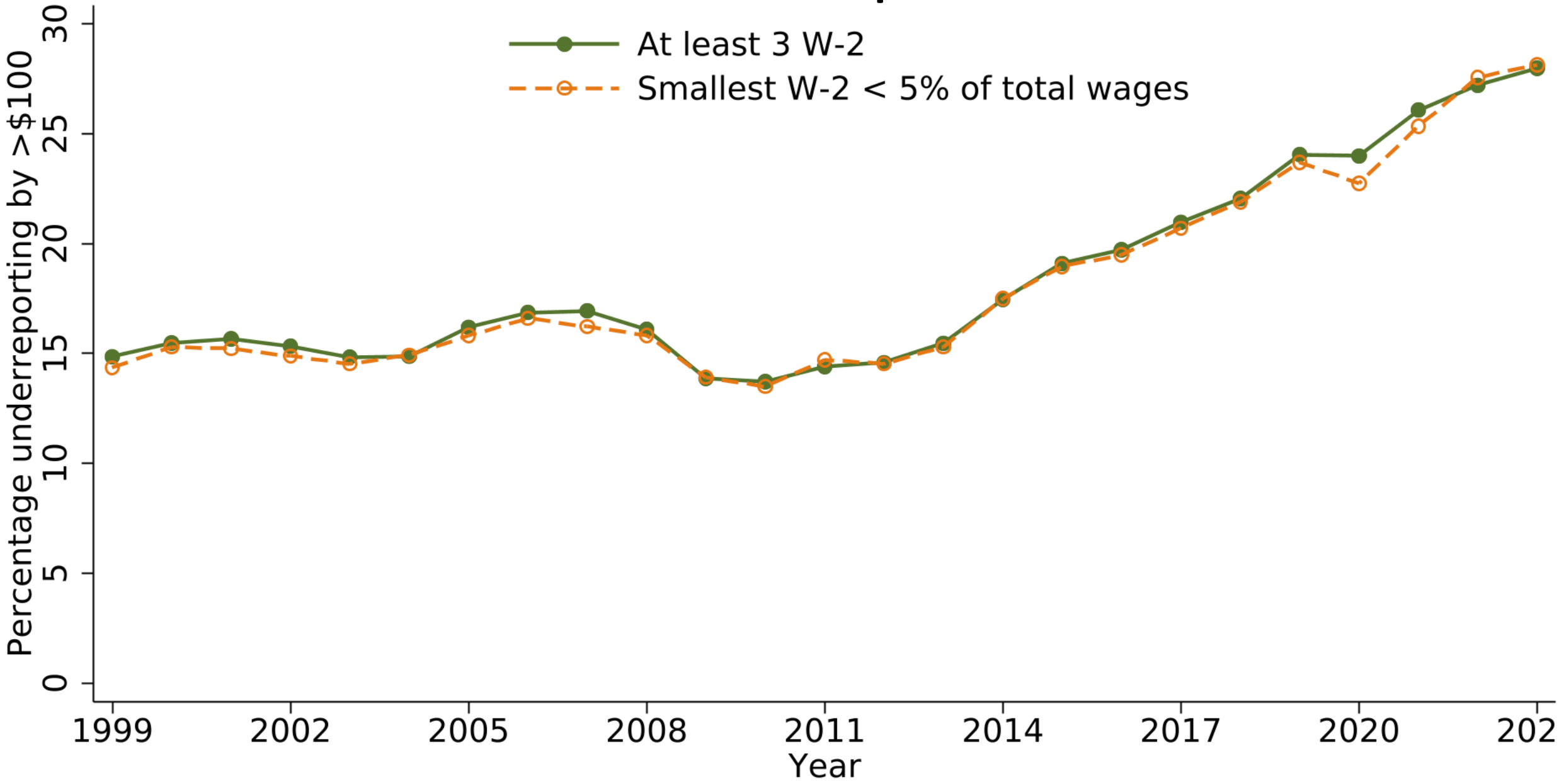
Predictors: Share of income from smallest W-2



Predictors: Sophistication and changes



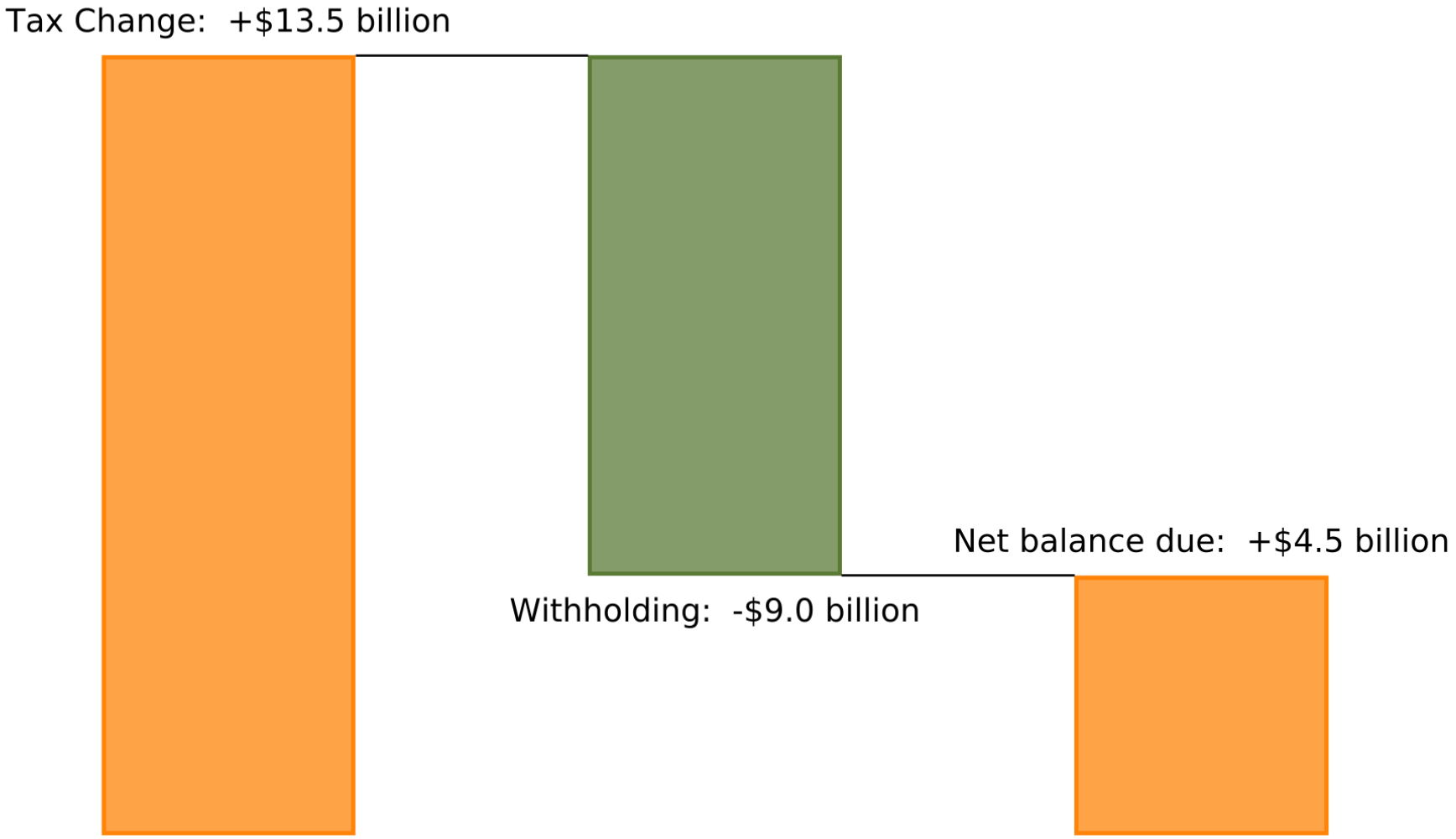
Observables don’t explain the trend



How would balances due change without underreporting?

- Focus on the case (92% of underreporters) that involve omitting at least one W-2.
- Failing to attach the W-2 affects gross tax liability *but also reported withholding*.

How would balances due change without underreporting?



(prior to any enforcement)

How would balances due change without underreporting?

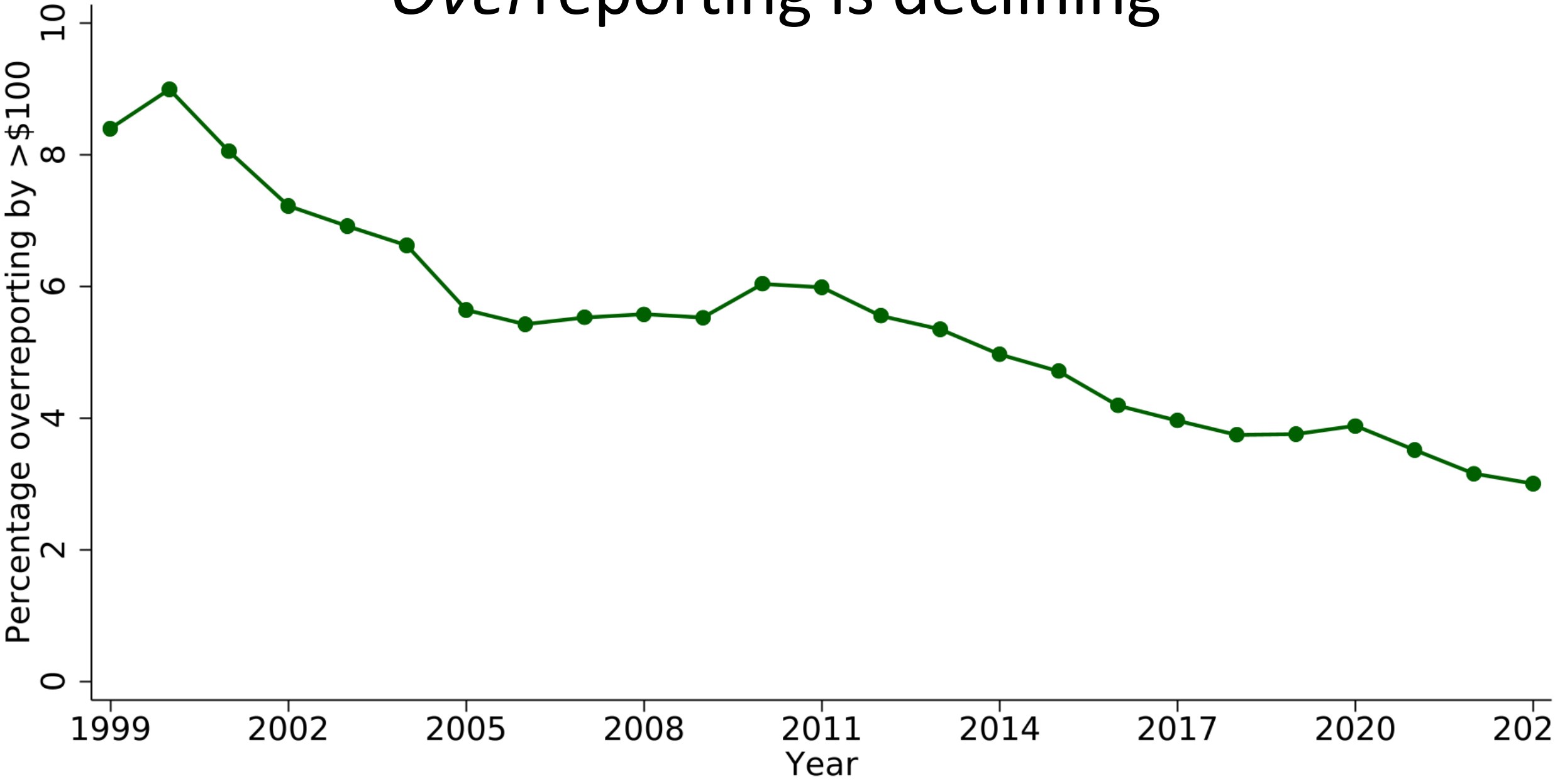
Change in balance due	Share of sample	Average tax rate on underreported wages	Average withholding rate on underreported wages	Change in liability net of withholding
Less than -$2,000	3%	6%	20%	-$1.8B
-$2,000 under $0	23%	2%	9%	-$0.6B
$0	2%	2%	2%	$0B
$0 under $2,000	65%	14%	6%	$2.2B
$2,000 under $11,000	7%	20%	7%	$2.7B
$11,000 and over	1%	30%	16%	$1.9B
Total	100%	15%	10%	$4.5B

OVERREPORTING

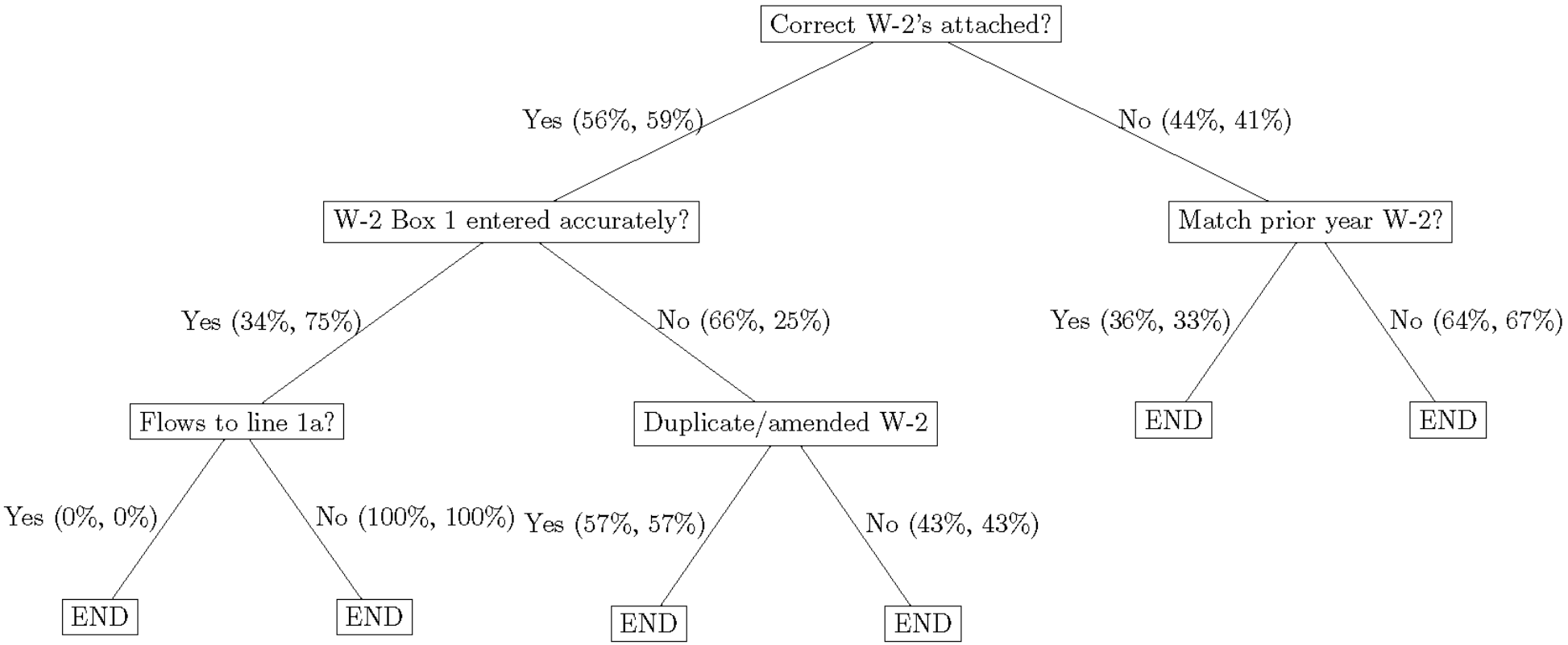
Overreporting: overview

- In 2023:
  - Line 1a definition: 2.1% overreport.
- But, for comparability across time, we also use a Line 1 version.
  - Line 1 can be greater than W-2 wages even with perfect reporting.
  - Under this definition, 3.0% overreport in 2023.

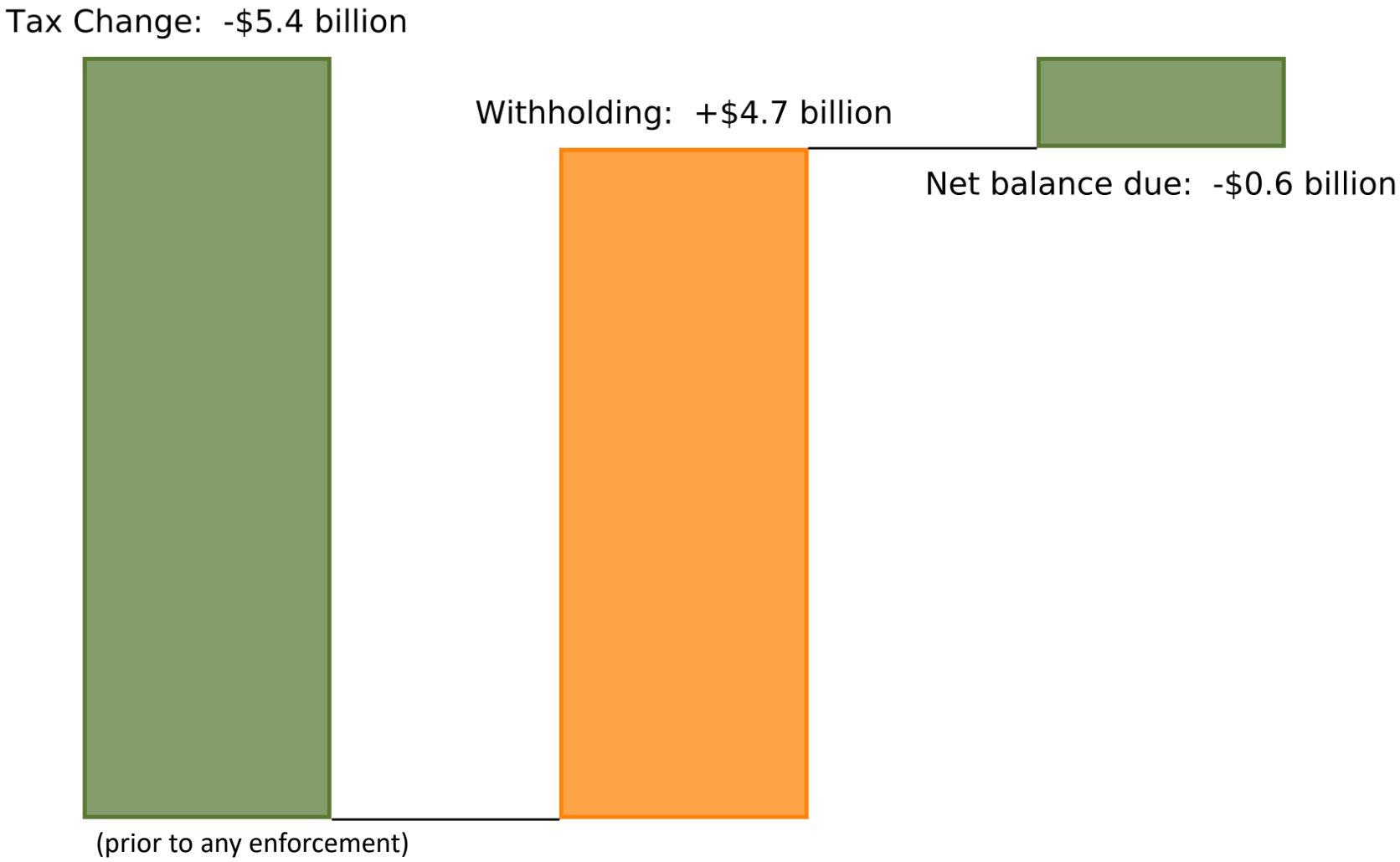
*Over*reporting is declining



What might explain overreporting?



How would balances due change without overreporting?

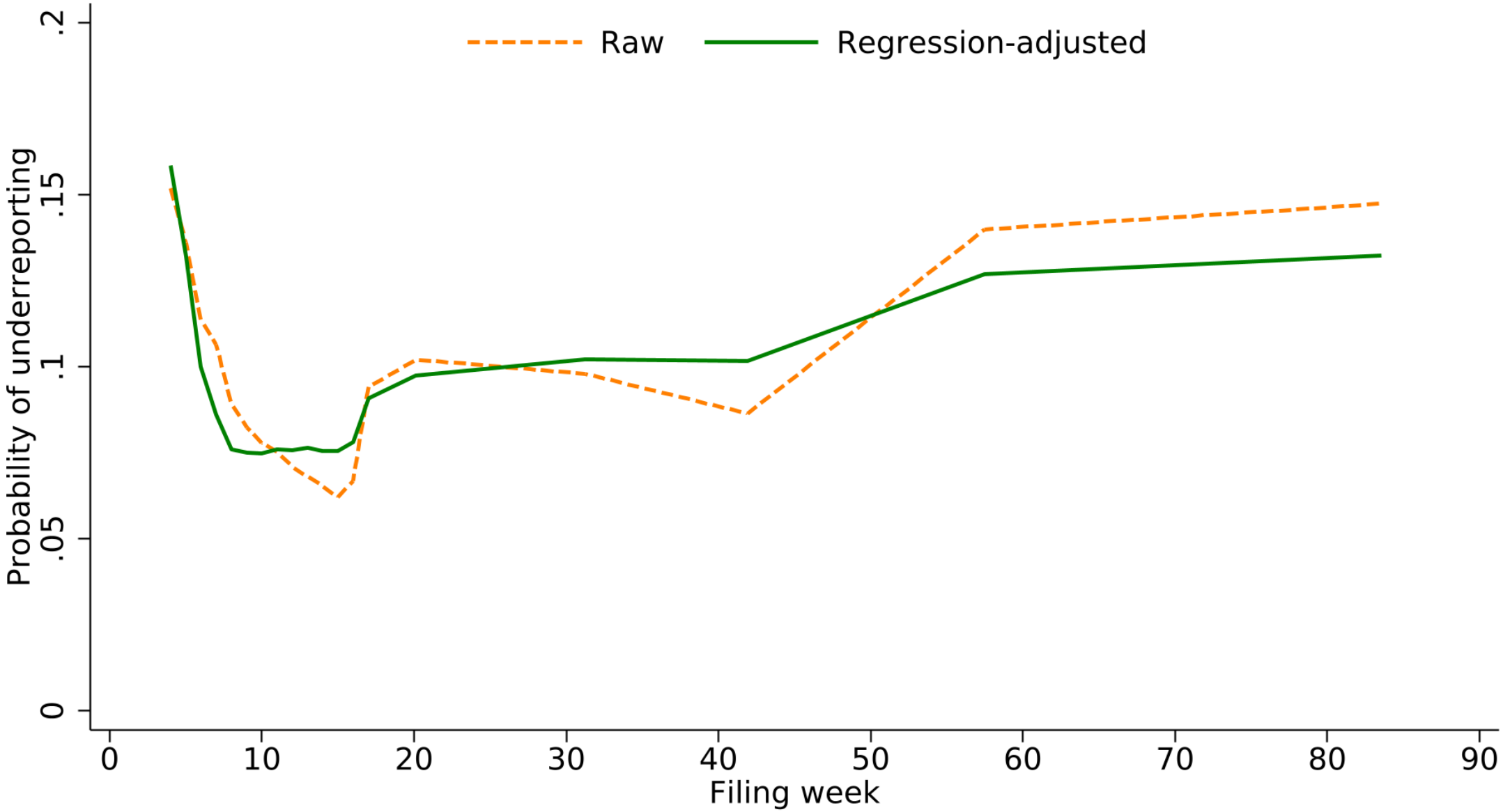


Conclusion

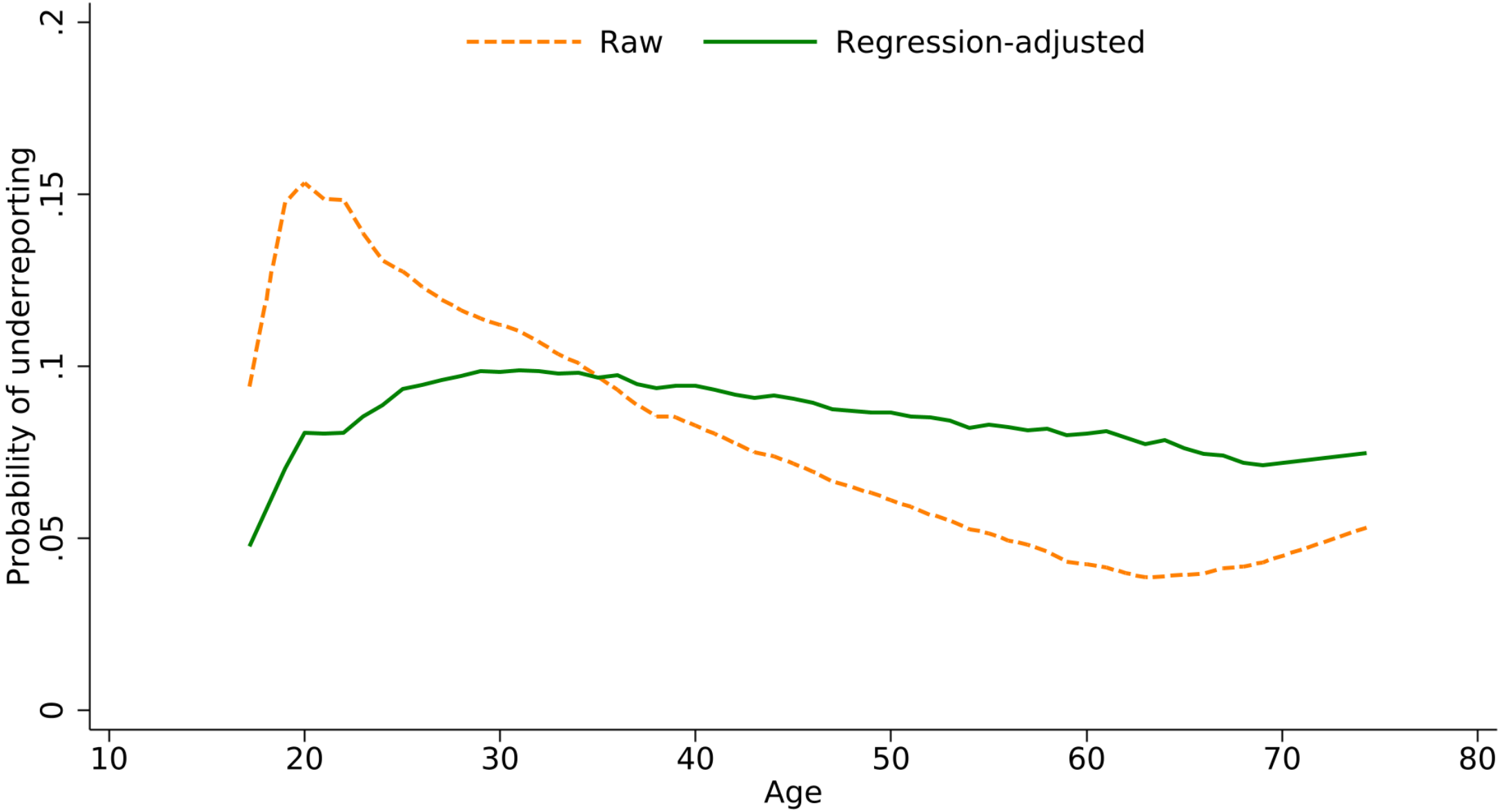
- Misreporting of employer-reported wages is mostly the result of failing to attach at least one W-2 to the tax return.
- Largely in circumstances like multiple jobs, small jobs
- Misreporting is rising; this is mostly NOT explained by observables.
- Costs to enforce, pain and hassle for taxpayers, but not a lot of lost revenue
  - Withholding is key

APPENDIX

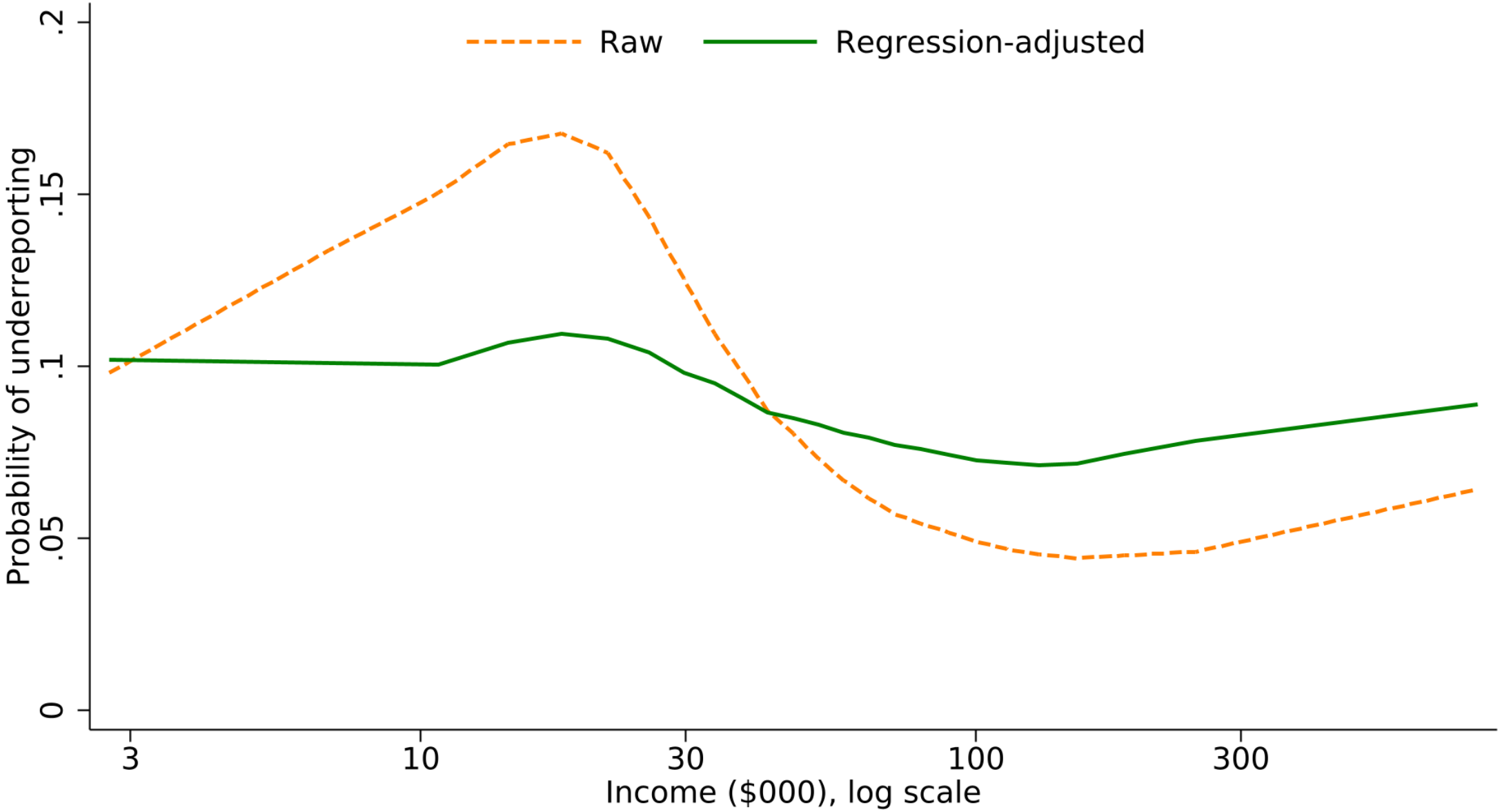
Predictors: Time of filing



Predictors: Age



Predictors: Income

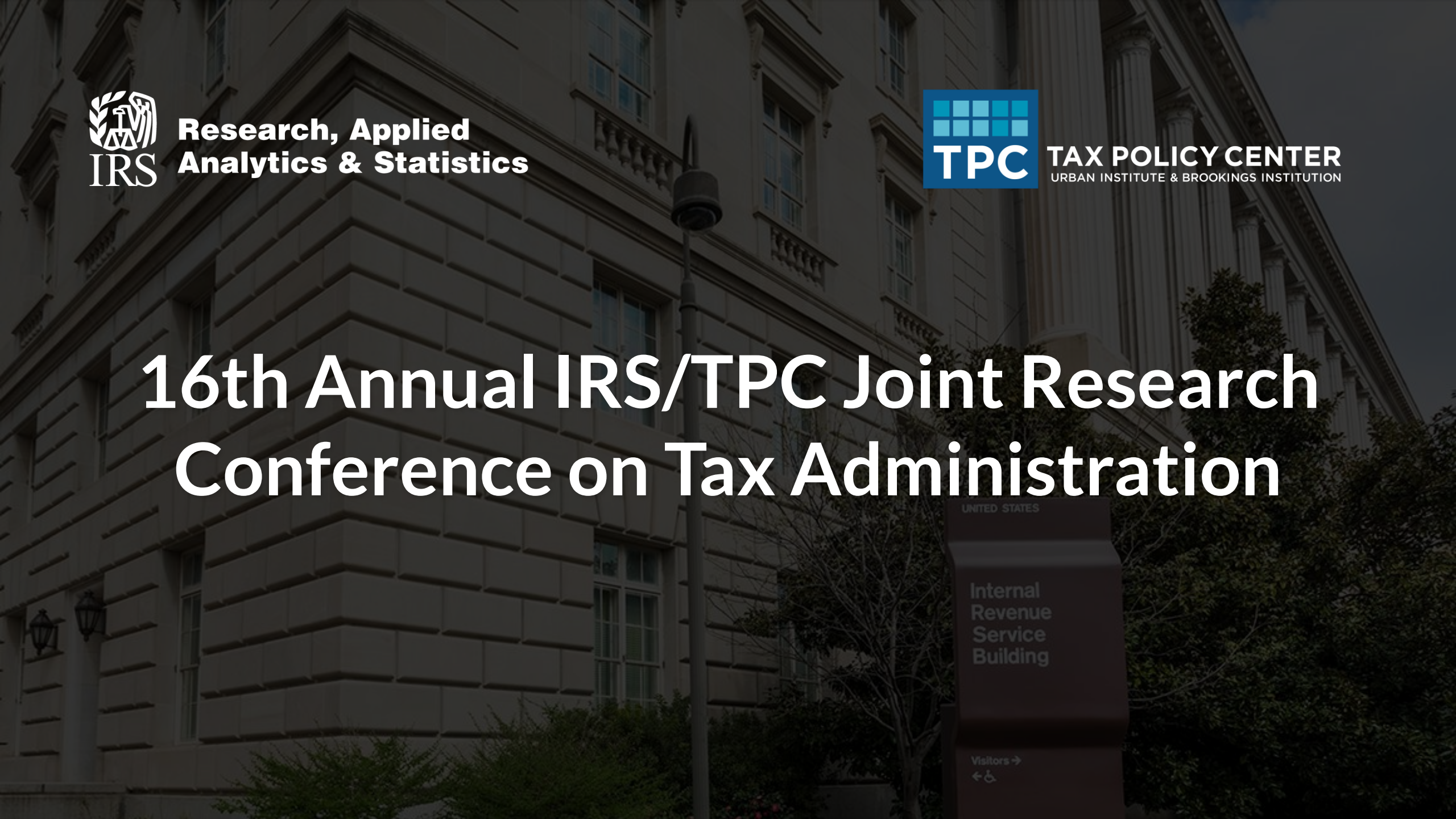


Effects on Earned Income Credit

As a result of misreporting, taxpayers...	Underreporters	Overreporters
Newly obtain EIC	8.5%	9.7%
Increase EIC	20.2%	11.2%
Decrease EIC	11.7%	9.9%
Lose EIC	2.1%	10.8%
	More mass in phaseout range	More evenly split

Amended returns

	Raw share	Regression-adjusted
Underreporters	0.042	0.026
Overreporters	0.030	0.012
All else	0.017	--



Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

16th Annual IRS/TPC Joint Research Conference on Tax Administration

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

Out of Sight, Out of Tax?

Strategic Payment Routing and the Limits of Third-Party Reporting

Riddha Basu • Omri Even-Tov • Ben Lourie • Chenqi Zhu

Presented by Riddha Basu
George Washington University
June 2026

Roadmap

INTRODUCTION

HYPOTHESIS DEVELOPMENT

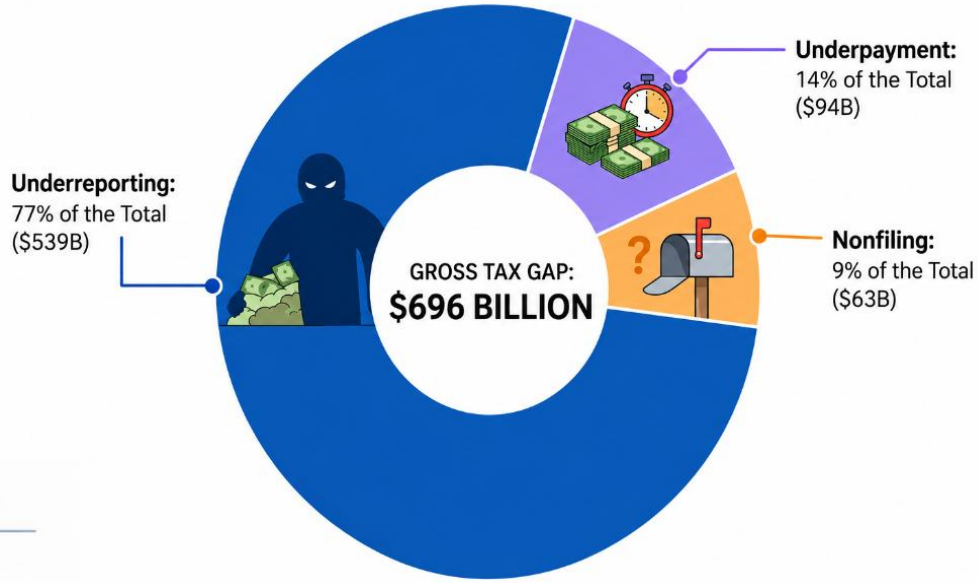
SAMPLE & DATA

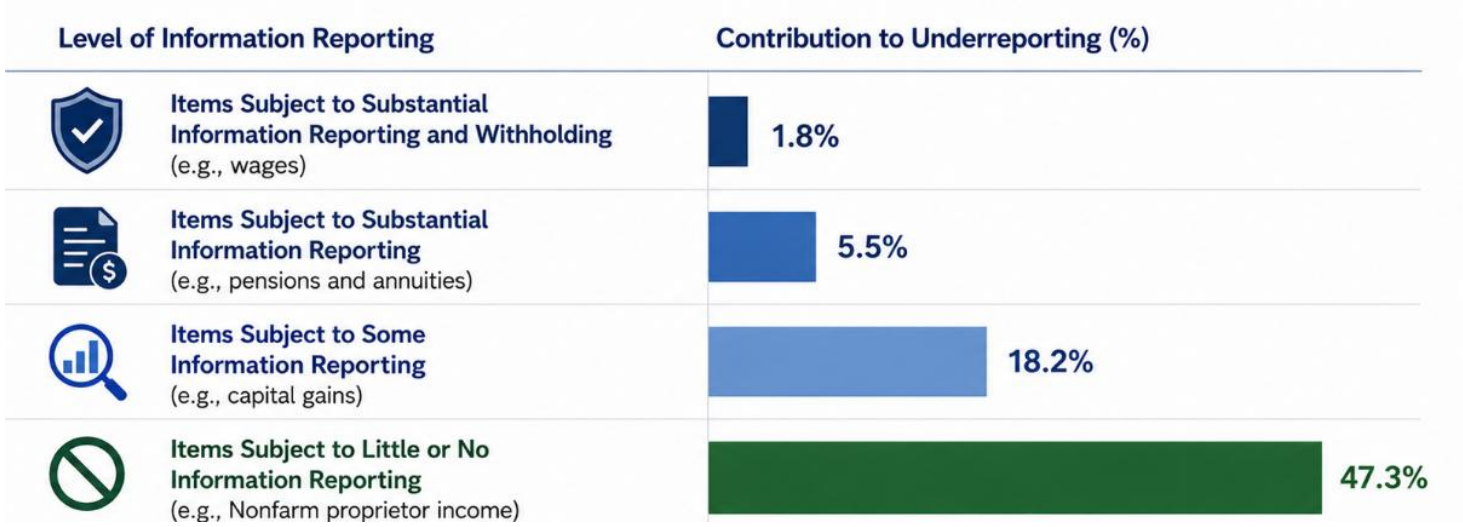
EMPIRICAL RESULTS

CONCLUSION

Motivation: Tax Underreporting

Underreporting is the largest source of U.S. tax noncompliance **($539B, 77% of the gap)** — and is concentrated where income is hard to verify independently.



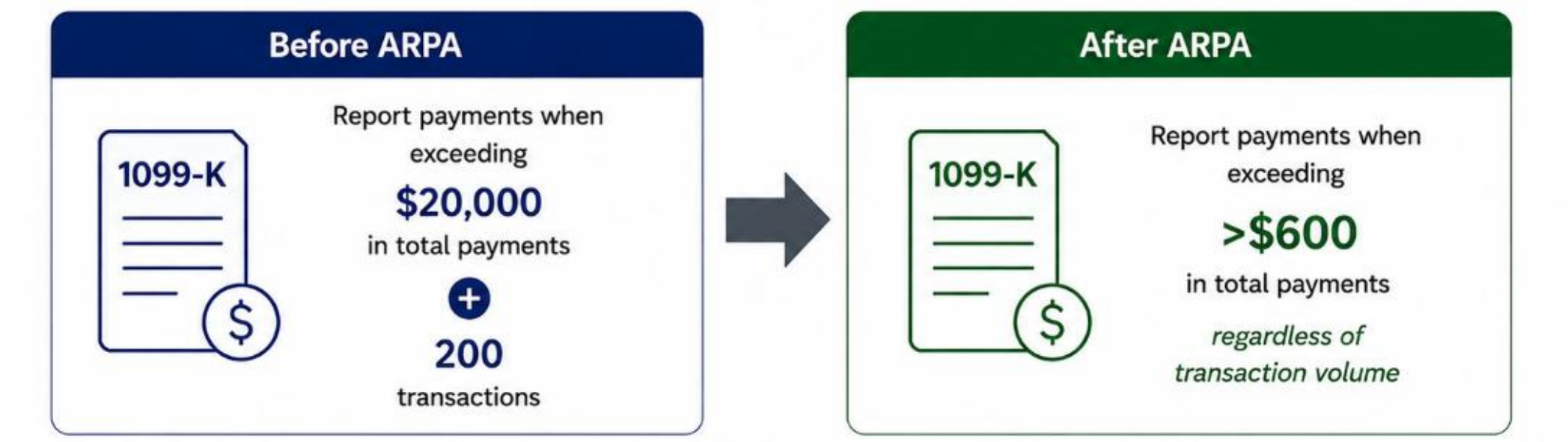


Source: IRS Research, Applied Analytics & Statistics (2024). Tax gap projections for tax year 2022 (Publication 5869, Rev. 10-2024). U.S. Department of the Treasury.

Motivation: Third-Party Reporting and Small Businesses

Motivation: ARPA of 2021

ARPA was enacted **March 11, 2021**; the lower reporting threshold was scheduled to take effect **January 1, 2022**.



Applies to covered platforms (Venmo, PayPal, Cash App, Apple Pay, Google Pay, Stripe, …)

Bank-to-bank; not a TPSO under § 6050W

✕ **Does not apply**
Zelle

Research Question

Does expanded third-party information reporting lead small businesses to change how they route customer payments?

- ARPA of 2021
- Peer-to-peer payment platforms (TPSOs vs. Zelle)

Hypothesis Development 1

Allingham–Sandmo–Yitzhaki framework: compliance increases with detection probability and penalty severity.

Third-party reporting raises the expected cost of underreporting, shifting behavior toward greater compliance (Slemrod & Yitzhaki 2002; Kleven et al. 2011; Slemrod 2007, 2019; Slemrod et al. 2017; Adhikari et al. 2021, 2022).

Strategic response: taxpayers may respond to reporting regimes by shifting activity toward alternatives with lower reporting visibility (De Simone et al. 2020; Ferguson 2023).

H1. Small businesses shift receipts from covered payment apps to Zelle after ARPA, **increasing Zelle's share.**

Hypothesis Development 2

Alternatively, expanded 1099-K reporting may be **absorbed through other margins** or blocked by frictions — without any change in payment routing.

Frictions that limit routing changes

- **Awareness:** may not know platform rules differ
- **Switching costs:** requires merchant and customer coordination
- **Platform differences:** speed, irreversibility, buyer protection
- **Implementation complexity:** delays and evolving guidance may slow response

Predictions

H1 (Strategic Routing). Small businesses' Zelle share of payment-app receipts increases following the announcement of the ARPA Form 1099-K reporting expansion, with the increase concentrated among firms that have *greater ex-ante incentives or capacity* to operate in lower-visibility environments.

Weaker tax compliance

Low_Tax, Inmate_Tie

Kleven et al. (2011); Slemrod et al. (2017)

Lower risk aversion

Robinhood, Crypto, Gambling

Allingham-Sandmo-Yitzhaki; Slemrod (2007)

Tighter financial constraints

Debt obligations, Overdraft

Edwards et al. (2016); Law & Mills (2015)

Less transparent info. env.

Cash deposits, Payroll usage

Kerr (2019); De Simone et al. (2020)

Sample Construction

Data source

Transaction-level bank and credit-card data from a major financial data aggregator (deHaan et al. 2024).

Each transaction: amount, date, direction, counterparty, vendor-tagged category, anonymized description.

Anonymous user IDs link transactions across accounts and cards.

Business identification

Retailers retained if, in any pre-period tax year (2019–2021), they made either (a) ≥3 IRS payments and ≥1 IRS refund, or (b) ≥4 IRS payments. Captures quarterly Form 1040-ES estimated-tax behavior characteristic of small businesses.

Sample-selection waterfall

	Initial bank-panel observations	57,008,193
−	Less: Firms not meeting pre-period activity threshold	53,413,625
=	Sub-total (commercially active retailers)	3,594,568
−	Less: Annual app income ≤ $10,000	2,138,219
=	**Final sample (firm-months)**	**1,456,349**
	Unique small businesses	**26,445**

Descriptive Evidence: Zelle Share Over Time

Mean Zelle share, by year

Year	Zelle share	Δ vs. prior
2020	0.432	
2021	0.480	+0.048
2022	0.531	**+0.051**
2023	0.567	+0.036
2024*	0.579	+0.012

** 2024 observation includes January through August.*

Key takeaways

+14.7 pp
Zelle share growth from 2020 to 2024

34%
Increase relative to 2020 baseline

Largest single-year jump: 2021 → 2022
+5.1 pp, immediately following ARPA's anticipated effective date

Caveat: *Aggregate trend alone could reflect broader digital-payment adoption. We need within-firm variation and heterogeneity tests to identify the strategic-routing channel.*

Table 1. Summary Statistics

Variable	Mean	Std. Dev.	P25	Median	P75
Zelle Share	0.512	0.456	0.000	0.584	1.000
Tax_Rate	0.079	0.144	0.009	0.030	0.083
Low_Tax	0.500	0.500	0	0	1
Inmate_Tie	0.052	0.223	0	0	0
Crypto_Trader	0.220	0.415	0	0	0
Robinhood_User	0.274	0.446	0	0	1
Gambling	0.120	0.325	0	0	0
Overdraft	0.514	0.500	0	1	1
Debt_Payer	0.479	0.500	0	0	1
Cash_Reliant	0.518	0.500	0	1	1
Payroll_Service	0.136	0.343	0	0	0

N = 1,456,349 firm-months across 26,445 unique small businesses, January 2020 – August 2024. Pre-period (≤ 2021) characteristics used in heterogeneity tests. All variables defined in Appendix A.

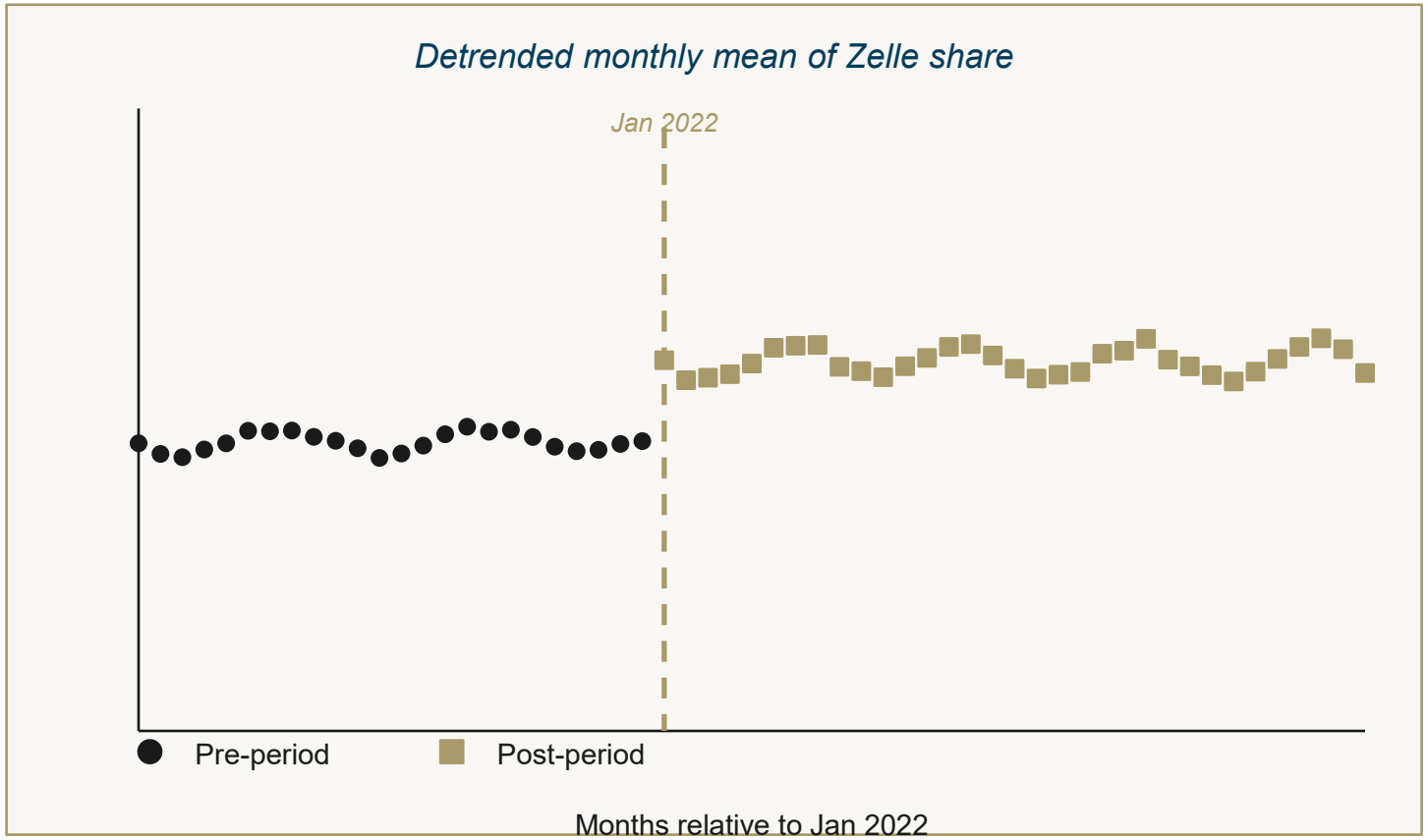
Main Result: Effect of ARPA on Zelle Share

$$Zelle\ Share_{it} = \alpha + \beta_1\ Post_t + \gamma\ X_t + \mu_i + \varepsilon_{it}$$

	(1) Baseline	(2) Linear trend	(3) Placebo
Post	0.088***	0.005***	
	(66.941)	(3.469)	
Trend		0.003***	0.004***
		(38.270)	(20.223)
Trend × Post		−0.001***	
		(5.360)	
Placebo Post			−0.002
			(0.652)
Trend × Placebo Post			−0.000
			(0.481)
Firm fixed effects	Yes	Yes	Yes
Observations	1,456,349	1,456,349	636,181
Adj. R^2	0.667	0.670	0.737

Baseline: Zelle share rises +8.8 pp post-2022 (20% relative increase). **With linear trend:** level shift attenuates to +0.5 pp but remains significant. **Placebo:** no comparable jump.

Structural Break in Zelle Share



Key takeaways

1. After partialling out the linear pre-trend, the detrended Zelle share displays a **discrete level shift** at Jan 2022 — the originally scheduled ARPA effective date.
2. The estimated jump (β_1 on Post) = **0.005** (t = 3.47, p < 0.001) — small in absolute terms, but identified after removing a smooth time trend.
3. The pre-period slope is approximately flat, and the post-period series stabilizes at a visibly higher level — consistent with a one-time structural break rather than a smooth growth path.

Table 4. Heterogeneity: Weak Tax Compliance

	(1) Low tax	(2) Inmate-related
Post × Low_Tax	0.006**	
	(2.027)	
Post × Inmate_Tie		0.028***
		(4.400)
Small Business FE	Yes	Yes
Year-Month FE	Yes	Yes
Clustering	Firm	Firm
Observations	1,183,455	1,456,349
Adjusted R^2	0.669	0.670

*Dependent variable: Zelle Share. Low_Tax = below-median pre-period estimated tax margin. Inmate_Tie = presence of inmate-related transactions in the pre-period. Robust t-statistics clustered at the firm level in parentheses. ***, **, * = 1%, 5%, 10%.*

Heterogeneity: Weak Tax Compliance

Are firms with ex-ante markers of weaker compliance more responsive to the reporting expansion?

Interaction term	Coefficient	In pp
Post × Low_Tax (t = 2.03)	**0.006****	+0.6 pp
Post × Inmate_Tie (t = 4.40)	**0.028*****	+2.8 pp

Interpretation

Low_Tax: firms with below-median pre-period effective tax rates display a 0.6 pp larger increase in Zelle share — consistent with greater ex-ante incentive to underreport.

Inmate_Tie: retailers with inmate-related transactions (proxy for social ties to informal economic activity) display a 2.8 pp larger increase — the largest heterogeneity coefficient in this panel.

Motivation: *Allingham–Sandmo–Yitzhaki framework; tax-evasion social networks (Alstadsæter et al. 2019; Garin et al. 2025).*

Table 5. Heterogeneity: Risk Aversion

	(1) Robinhood	(2) Crypto	(3) Gambling
Post × Robinhood_User	0.018***		
	(4.590)		
Post × Crypto_Trader		0.013***	
		(4.033)	
Post × Gambling			0.010**
			(2.363)
Small Business FE	Yes	Yes	Yes
Year-Month FE	Yes	Yes	Yes
Clustering	Firm	Firm	Firm
Observations	1,456,349	1,456,349	1,456,349
Adjusted R^2	0.670	0.670	0.670

*Dependent variable: Zelle Share. Each indicator captures pre-period use of the respective platform/activity. Robust t-statistics clustered at the firm level in parentheses. ***, **, * = 1%, 5%, 10%.*

Heterogeneity: Risk Aversion

Are firms with lower risk aversion more willing to operate at the edge of reporting compliance?

Post × Robinhood_User	***Post × Crypto_Trader***	***Post × Gambling***
+1.8 pp	**+1.3 pp**	**+1.0 pp**
0.018***	0.013***	0.010**
(t = 4.59)	(t = 4.03)	(t = 2.36)

All three proxies point in the same direction with significance at the 1–5% level. *Consistent across conceptually distinct markers of risk-tolerant behavior (investment platforms, cryptocurrency, sports betting), reinforcing that the migration reflects intentional choice rather than passive technology diffusion.*

Table 6. Heterogeneity: Financial Constraints

	(1) Debt obligations	(2) Overdraft charges
Post × Debt_Payer	0.007**	
	(2.426)	
Post × Overdraft		0.016***
		(6.215)
Small Business FE	Yes	Yes
Year-Month FE	Yes	Yes
Clustering	Firm	Firm
Observations	1,456,349	1,456,349
Adjusted R^2	0.671	0.670

*Dependent variable: Zelle Share. Debt_Payer = pre-period debt-servicing activity (loans, credit-card repayments). Overdraft = pre-period overdraft / insufficient-funds charges. Robust t-statistics clustered at the firm level in parentheses. ***, **, * = 1%, 5%, 10%.*

Heterogeneity: Financial Constraints

Do financially constrained firms have stronger incentives to reduce visible tax obligations?

Post × Debt_Payer

+0.7 pp

0.007**

(t = 2.43)

Firms with regular debt-servicing transactions (loans, mortgages, credit cards)

Post × Overdraft

+1.6 pp

0.016***

(t = 6.21)

Firms with overdraft or insufficient-funds charges

Mechanism. *Financially constrained firms place a higher marginal value on unremitted tax dollars — they have **more to gain** from reducing visible income, all else equal. The overdraft effect is the largest t-statistic among all heterogeneity interactions in the paper.*

Table 7. Heterogeneity: Information Environment

	(1) High cash deposits	(2) Payroll software
Post × Cash_Reliant	0.010***	
	(3.976)	
Post × Payroll_Service		−0.017*
		(1.771)
Small Business FE	Yes	Yes
Year-Month FE	Yes	Yes
Clustering	Firm	Firm
Observations	1,456,349	1,456,349
Adjusted R^2	0.670	0.670

*Dependent variable: Zelle Share. Cash_Reliant = pre-period cash/check deposit activity. Payroll_Service = pre-period payments to payroll processors (e.g., ADP, Gusto). Note the opposing signs. Robust t-statistics clustered at the firm level in parentheses. ***, **, * = 1%, 5%, 10%.*

Heterogeneity: Information Environment

Does the migration scale with how visible the firm's broader operations already are?

Low-visibility firms

Post × Cash_Reliant

+1.0 pp

0.010*** (t = 3.98)

Cash-intensive retailers (already low-visibility) shift more toward Zelle.

High-visibility firms

Post × Payroll_Service

−1.7 pp

−0.017* (t = 1.77)

Firms using ADP/Gusto (already high-visibility) shift less toward Zelle.

The sign flip is the falsification. A pure technology-adoption story predicts that firms with developed back-office infrastructure adopt new payment apps **earlier**, not later. The opposite sign is hard to reconcile with adoption — but is precisely what a **strategic-visibility** story predicts.

Table 8. Consequences on Tax Margins

	(1) Annual estimated tax margin (Tax_Rate)
Post × High_Zelle	−0.011***
	(4.529)
Small Business FE	Yes
Year FE	Yes
Clustering	Firm
Observations	79,775
Adjusted R^2	0.348

*Dependent variable: Tax_Rate (annual estimated tax margin from IRS / Treasury transactions). High_Zelle = 1 for firms with above-median Zelle-share growth in the post period. Firm-year level, excluding 2024. Robust t-statistics clustered at the firm level in parentheses. ***, **, * = 1%, 5%, 10%.*

Consequences: Tax Margins

Question. Do firms that migrate most aggressively toward Zelle subsequently exhibit lower estimated tax margins?

Post × High_Zelle

−1.1 pp

decline in annual estimated tax margin

≈ 14%

of the 7.9% sample-mean tax margin

Table 8 specification

Dependent variable	**Tax_Rate (annual)**
Coefficient (t-stat)	**−0.011*** (4.529)**
Sample	**Firm-year**
Observations	**79,775**
Fixed effects	**Firm + Year**
Clustering	**Firm**
Adj. R^2	**0.348**

Associational, not strictly causal: *High_Zelle and tax margins may be jointly determined. Nonetheless, the direction and magnitude are coherent with the heterogeneity evidence in Tables 4–7.*

Conclusions

1

Strategic substitution is real and measurable.

Following the anticipated ARPA expansion, small businesses redirected receipts toward Zelle.

2

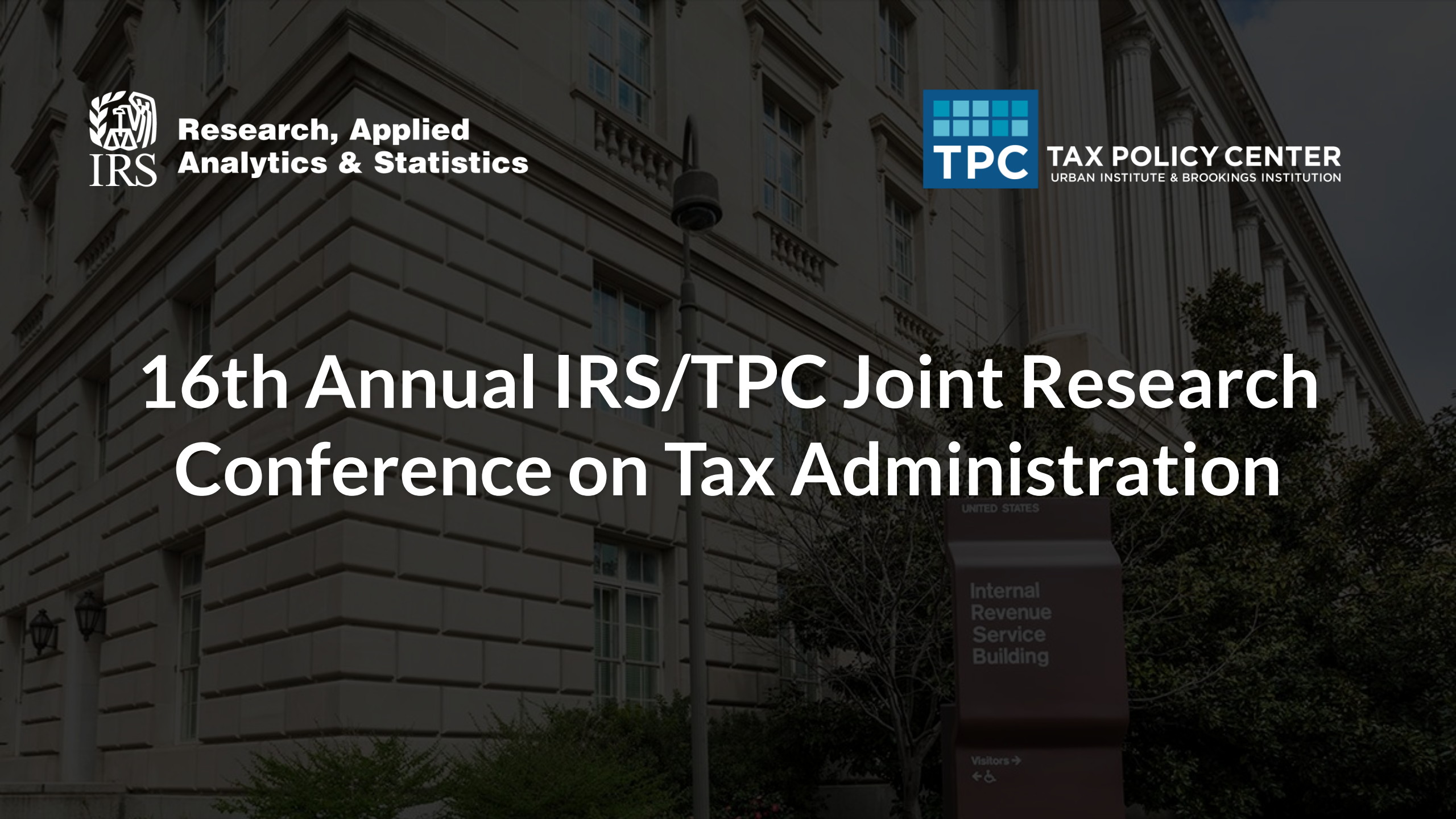
The response is concentrated, as theory predicts.

Firms with weaker tax compliance, lower risk aversion, tighter financial constraints, and lower-visibility operations respond more.

3

Incomplete reporting regimes leak.

Firms with the largest Zelle migration subsequently exhibit a 1.1 pp (≈14%) reduction in estimated tax margins, suggesting the compliance gains of expanded reporting were partially dissipated by substitution to uncovered channels.



Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

16th Annual IRS/TPC Joint Research Conference on Tax Administration

DISCUSSION: MAKING THIRD-PARTY DATA WORK HARDER

Jamie McGuire, Joint Committee on Taxation

June 25, 2026

This presentation embodies work undertaken for the staff of the Joint Committee on Taxation, but as members of both parties and both houses of Congress comprise the Joint Committee on Taxation, this presentation should not be construed to represent the position of any member of the Committee.

Main Takeaways

- Machine learning recovers missing admin detail. (Geiger et al)
- Wait, people underreport wages?! (Boning et al)
- Businesses practice platform arbitrage. (Basu et al)

Geiger et al

- Problem: SSA consolidates some box 12 data before sending W-2s to IRS.

- In an ideal world, change the data-sharing agreement with SSA.

- Second-best solution:
  - SOI gets a sample of taxpayer-submitted W-2 data that includes box 12 codes D-H.
  - This paper shows how box 12 codes D-H could be imputed to the population.

Geiger et al (continued)

- Why were CART and GBM chosen?
  - Maybe direct the reader to an explainer of these methods.
- Understanding the output:
  - Show what the models do with a simple example.
  - How accurately does it produce distributions of values?
  - Interaction with SOI sample weights?
- Could employer information be leveraged?
- Other tax-related applications?
  - Use 10-K to predict private firm data?

Boning et al

- Wages, salaries, and tips in TY 2014-2016 tax gap:
  - $7 billion per year in unpaid tax due to underreporting.
  - 1% of total tax gap, 2% of individual underreporting tax gap.
  - 1% net misreporting percentage.

- Pretty much perfect compliance except for tips?

- No, people omit W-2s!
  - 4% of returns for 2012 → 7% of returns for 2022 and 2023.
  - Also omit withholding from those W-2s.

Boning et al (continued)

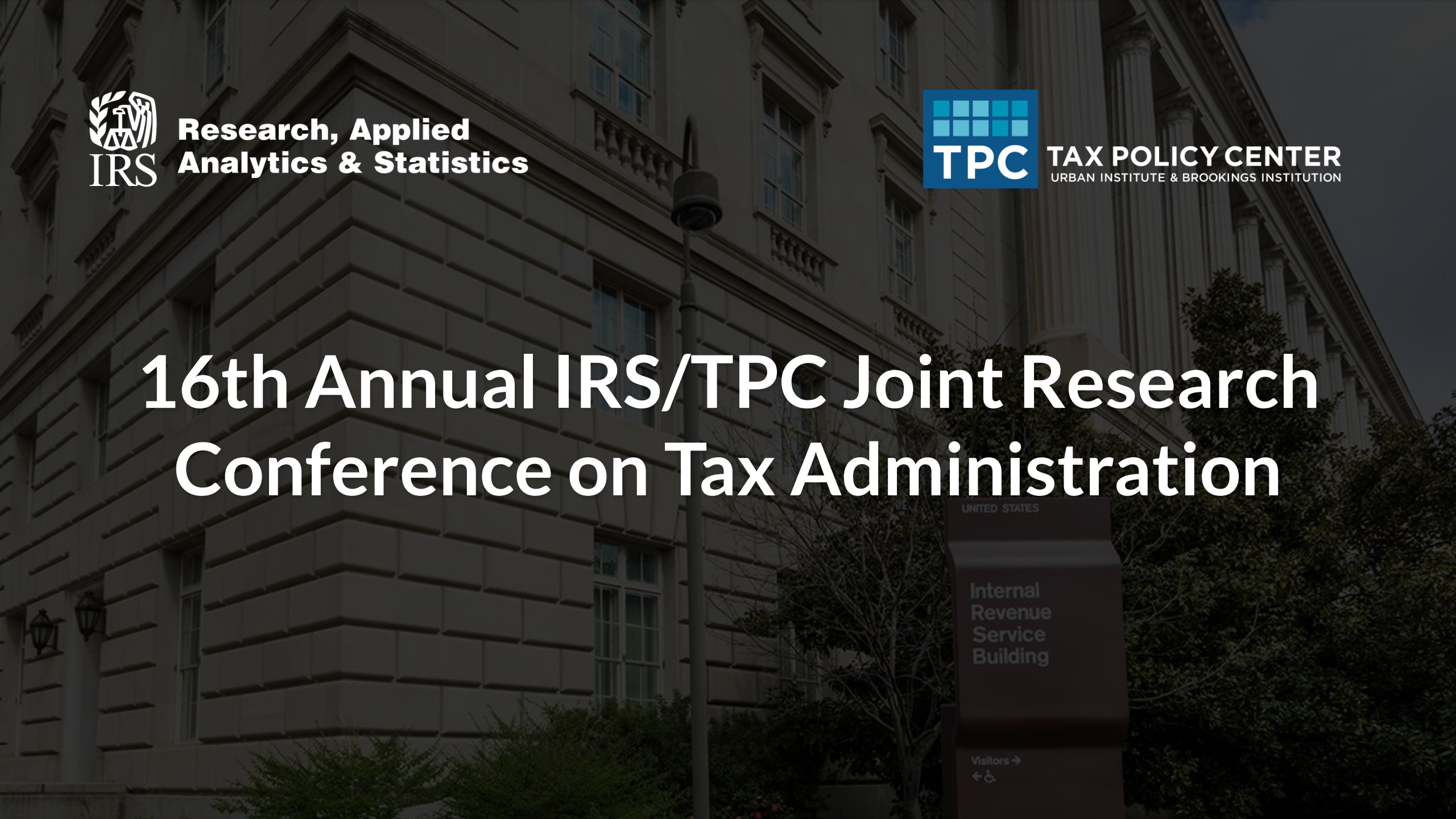
- Role of mistakes vs. intentional noncompliance?
  - Many have reasonable excuse, e.g. small short-term job, moving.
  - Some are worse off due to not getting credit for withholding.
  - 2.2% overreport wages for 2022-2023 (some weird cases here).
  - How many omit W-2s repeatedly over time?
- Does AUR flag omitted W-2s and other mismatches?
  - Enforcement and refunding overpayments should be easy here.
- Why is trend increasing?
  - Are W-2s increasingly sent electronically and forgotten about?

Basu et al

- Policymakers identified reporting gap between labor income (Form 1099-NEC) and income from platforms matching buyers and sellers (Form 1099-K).
  - Gap arose from reg authority to avoid double reporting in 6050W(g).
  - Transportation, delivery, short-term rentals, etc. (Bruckner's testimony)
  - Pure payment platforms were not central to the original discussions.
- ARPA (2021) lowered sec 6050W(e) threshold for TPSOs from >$20k, > 200 transactions to $600 with no transaction threshold.
  - Delays were announced after most transactions had occurred, so businesses acted as though the new threshold would apply.
  - Payments switched to the uncovered platform.

Basu et al (continued)

- Old rule covered > $20k, > 200 transactions. ARPA covered:
  - $600 to $20k
  - > $20k, < 200 transactions
- Sample includes accounts with payment app receipts > $10k/year, >= 50 transactions per month.
  - Provide concrete categories of the affected businesses?
  - ARPA shouldn't affect > $20k, > 200 transactions: falsification test?
- Future research: "Payments between friends" vs. "payments for goods and services"?



Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

16th Annual IRS/TPC Joint Research Conference on Tax Administration