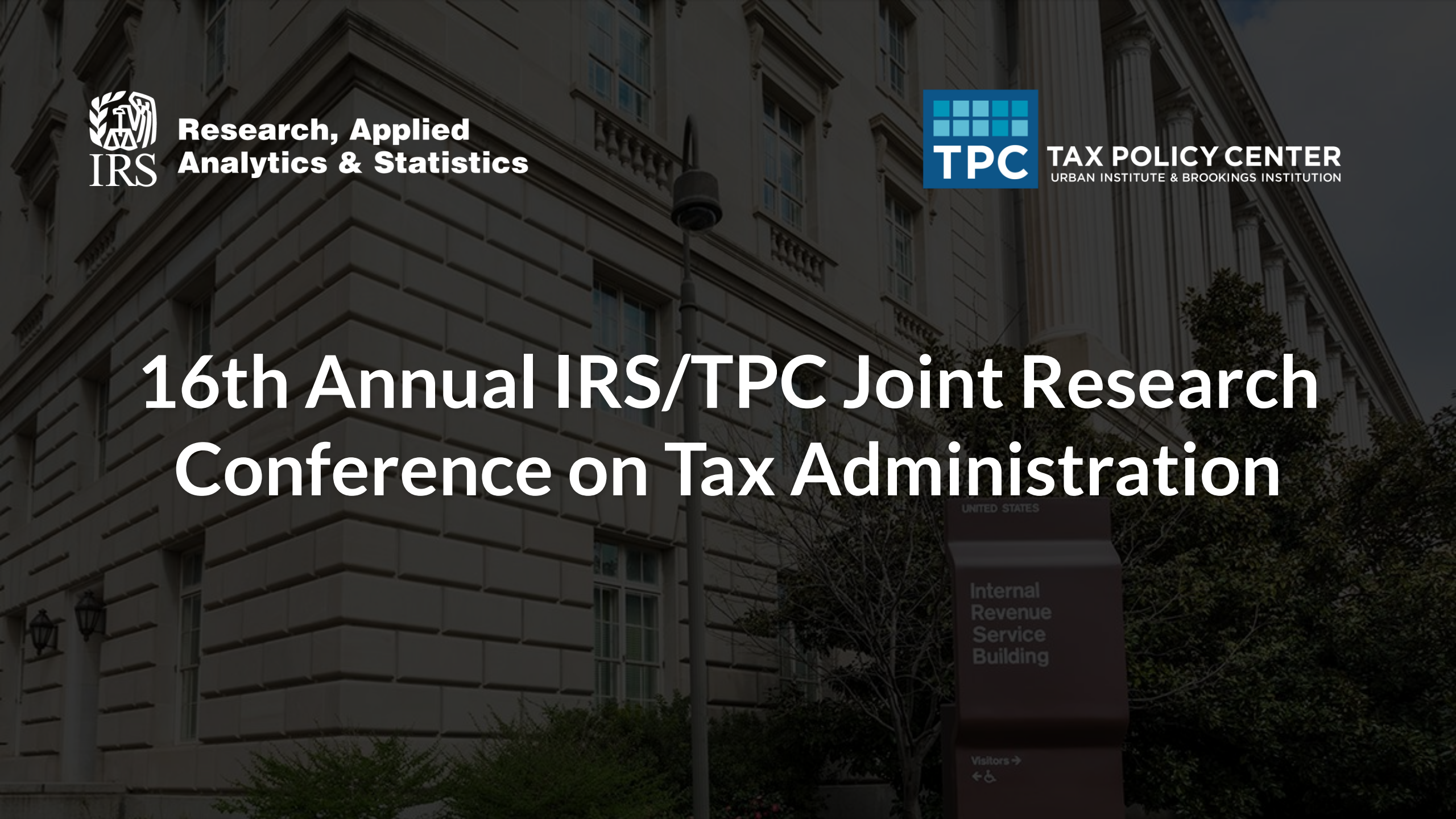


Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

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

Housekeeping

- The event is being recorded, and the recording will be posted online afterwards.
- The slides and related materials are available online.
- Virtual audience can type questions into the Q&A form at any time
- In-person attendees can use the QR code to submit questions.

IRS Research, Applied Analytics & Statistics

TPC TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

Session 1

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

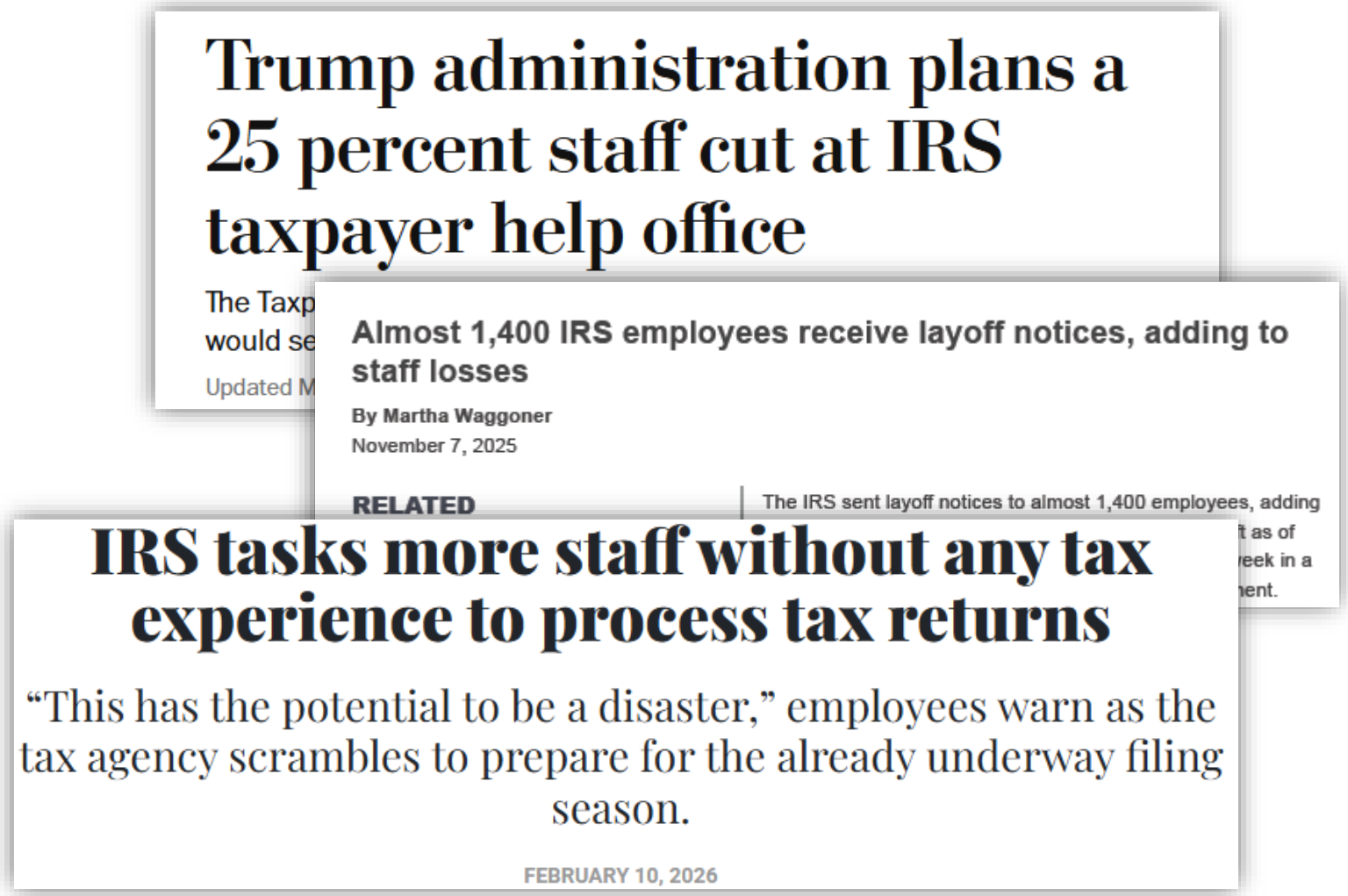
How do IRS staffing reductions impact individual income tax compliance?

June 2026

Vishal Baloria, University of Connecticut

IRS Staffing Cuts – Current Developments

"As the IRS prepares for the next filing season, it has lost 25% of its workforce in the midst of preparing technology and guidance for H.R.1 [One Big Beautiful Bill Act]. There are some provisions in there that I expect are going to raise ***taxpayer demand, tax professional demand, for IRS help and it's going to be very difficult****" Douglas O'Donnell, Former IRS Commissioner*



Trump administration plans a 25 percent staff cut at IRS taxpayer help office

The Taxp
would se
Updated M

Almost 1,400 IRS employees receive layoff notices, adding to staff losses

By Martha Waggoner
November 7, 2025

RELATED

The IRS sent layoff notices to almost 1,400 employees, adding
t as of
eek in a
ent.

IRS tasks more staff without any tax experience to process tax returns

"This has the potential to be a disaster," employees warn as the tax agency scrambles to prepare for the already underway filing season.

FEBRUARY 10, 2026

Research Questions

- What is the effect and economic magnitude of IRS staffing cuts on individual taxpayer compliance for states that are more versus less affected?
- Do the economic channels of IRS staffing cuts affecting taxpayer compliance include reduced enforcement and reduced taxpayer services?

Tension?

- Many proponents of IRS staffing cuts
- Technological shifts (machine learning/ internet/Turbotax)
- Documenting magnitude

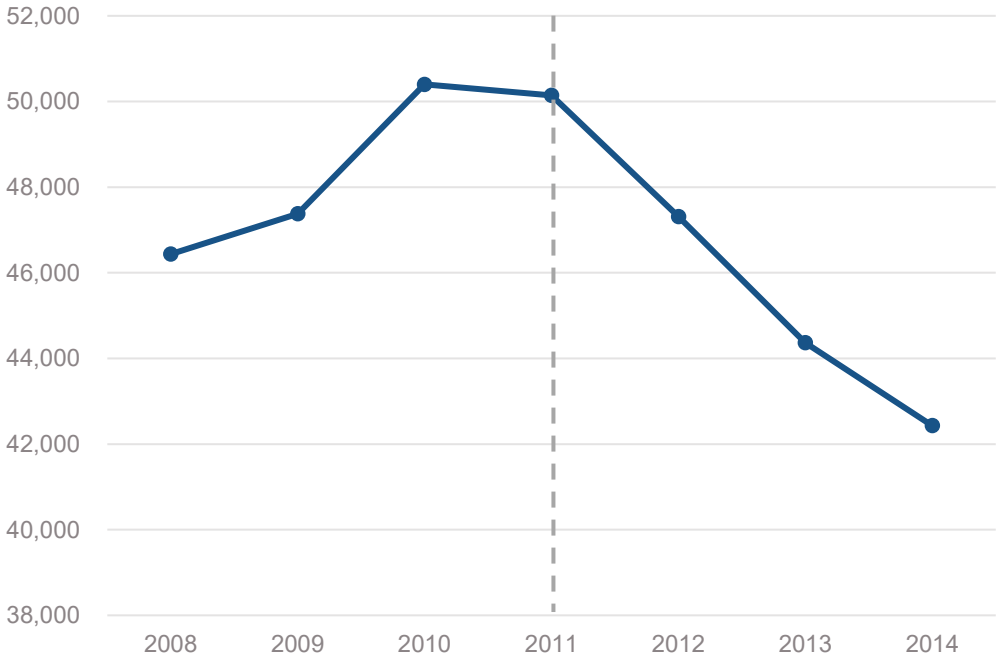
Contribution

- Provide evidence about impact of government agency labor resources on taxpayer behavior
  - Empirical question – reductions advocated by policy makers and could spur an efficient reallocation of resources
  - Study filing compliance; prior research generally examines reporting compliance in narrow settings
  - IRS staffing reduction literature focuses on corporate tax exposure (i.e., well-resourced sophisticated managers)
  - Individual taxation literature
- We shed light on an additional channel, *taxpayer assistance services*, through which IRS staff can impact tax compliance.
- We provide evidence about economic magnitude and moderating factors

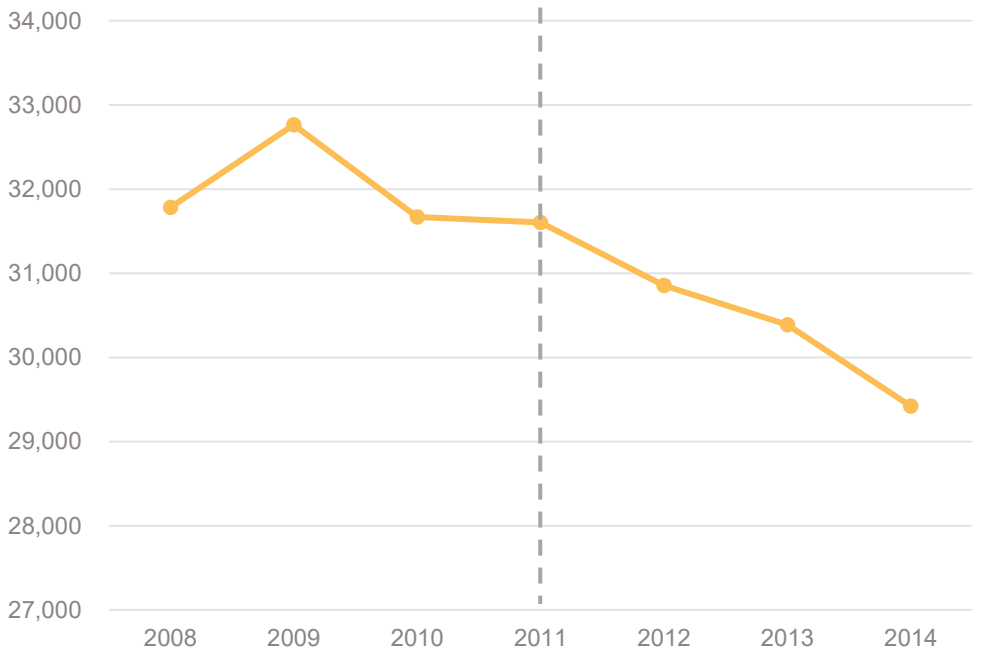
Setting: 2011 IRS Staffing Reduction Event

- Prior to 2025, the most recent substantial IRS workforce reduction followed a December 2010 hiring freeze (~11% overall reduction)

IRS Enforcement Full-Time Equivalent Positions

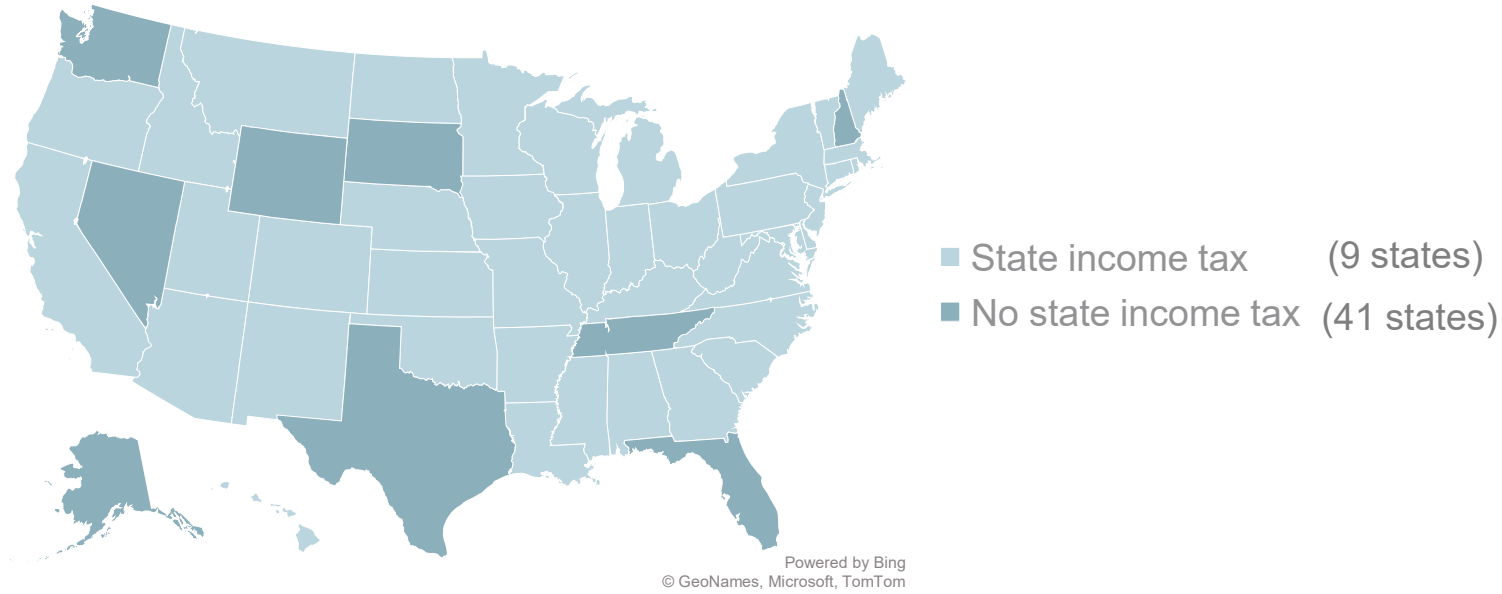


IRS Taxpayer Services Full-Time Equivalent Positions



Identification Strategy: State Individual Income Taxes

- Existence of state individual income tax provides source of plausibly exogenous variation, as state choice to impose income tax long predates our sample period
- Descriptively, income tax states experience greater decreases in IRS staff (23%) vs. non-income tax states (15%)



Hypothesis

- Individuals' decisions about federal tax compliance are linked to state tax compliance (Alm, Erard, and Feinstein 1996)
  - Most states use federal AGI or taxable income as a starting point
- Taxpayers in income tax states face a higher marginal tax rate, increasing incentives for non-compliance
- Taxpayers in income tax states face a higher compliance burden of filing two tax returns
  - Psychology literature finds that lower self-efficacy (i.e., belief you can succeed at a specific task) and task aversiveness (i.e., unpleasant tasks) increase the likelihood of a task will not be completed (e.g., Bandura 1977; Steel 2007)
  - Descriptively, we find that Google searches for IRS assistance are higher in income tax states compared to non-income tax states

Hypothesis: Relative to taxpayers in states less affected by IRS staff cuts, taxpayers in states more affected by IRS staff cuts decrease federal tax compliance after a 2010 IRS hiring freeze.

Research Design

- Difference-in-differences analysis around 2011 event year
  - Examine income tax returns filed for counties in states with an individual income tax versus counties in states without
- County-level analysis using IRS Statistics of Income (SOI) data:

$$Ln\_Filers_{cst} = \beta_0 + \boldsymbol{\beta_1 State\_IIT_{cs}} \times \boldsymbol{Post2011_t} + \beta_{2-6} Controls + County\ FE + Year\ FE + \varepsilon_{cst}$$

  - *Ln_Filers:* Natural log of the number of individual income tax returns filed in county *c* in year *t*
  - *State_IIT*: Indicator variable equal to 1 if county is in a state with an individual income tax
  - Control variables for county-level economic and demographic characteristics

Table 2 DiD Analysis: IRS Staff Reduction Event

- States more affected by IRS staffing cuts have less tax filings
- Dynamic analysis supports parallel trends assumption
- Economic magnitude: 3.2% decrease in tax filings in states more affected
- 2.5% relates to enforcement and 0.7% relates to taxpayer assistance services based on employee classification in cuts

	(1)	(2)
	ln_Filers	*ln_Filers*
State_IIT × Post2011	-0.035**	-0.032**
	(-2.45)	(-2.60)
ln_Employment		0.198***
		(3.11)
ln_Establishments		0.152***
		(2.98)
ln_Wages		0.056
		(1.65)
ln_Nonwage_income		0.054***
		(8.55)
ln_Avg_AGI		-0.248***
		(-5.89)
Popgrowth		0.096
		(0.84)
Observations	18,151	18,151
Within R-squared	0.022	0.198
Fixed Effects	County & Year	County & Year
Cluster	State	State

Table 3 Enforcement and Taxpayer Services

- Use BLS data
- Counties with high per capita tax examiners still have enforcement
- Counties with high per capita paid tax preparers can fill in the gap

	(1)	(2)
	ln_Filers	*ln_Filers*
State_IIT × Post2011	-0.054***	-0.070***
	(-3.22)	(-6.13)
State_IIT × Post2011 × High_examiners	0.046*	
	(1.91)	
State_IIT × Post2011 × High_preparers		0.071***
		(2.90)
High_examiners	1.095	
	(1.26)	
High_preparers		0.275
		(1.08)
Control Variables?	Yes	Yes
Observations	10,713	9,718
Within R-squared	0.318	0.279
Fixed Effects	County & Year	County & Year
Cluster	State	State

Table 4 Taxpayer Income Level

- Based on % of returns in income category for county
- Higher income
  - Enforcement Channel
- Middle income
  - Assistance Channel
- Lower income
  - Assistance demand
  - But, saturated with EITC tax shops

	High-income	Middle-income	Low-income
	> $100,000	$25,000-$100,000	<$25,000
	ln_Filers	ln_Filers	ln_Filers
State_IIT × Post2011	-0.045**	-0.036***	-0.001
	(-2.41)	(-4.42)	(-0.06)
ln_Employment	0.496***	0.102	0.260***
	(7.04)	(1.13)	(3.84)
ln_Establishments	-0.108	0.154**	0.199**
	(-1.19)	(2.30)	(2.63)
ln_Wages	-0.109**	0.073	0.021
	(-2.15)	(1.39)	(0.39)
ln_Nonwage_income	0.040***	0.045***	0.068***
	(3.91)	(5.25)	(4.38)
ln_Avg_AGI	-0.370***	-0.138***	-0.303***
	(-4.51)	(-4.51)	(-6.44)
Popgrowth	0.142	0.068	-0.101
	(0.71)	(0.33)	(-0.91)
Observations	3,765	6,782	7,604
Within R-squared	0.302	0.139	0.250
Fixed Effects	County & Year	County & Year	County & Year
Cluster	State	State	State

Table 5 Reported Nonwage Income as DV

- Extensive margin
- Taxpayers have discretion in reporting nonwage income

	(1)	(2)
	ln_Reported_nonwage_income	*ln_Reported_nonwage_income*
State_IIT × Post2011	-0.109**	-0.078**
	(-2.29)	(-2.53)
ln_Employment		-0.136
		(-1.03)
ln_Establishments		0.321***
		(3.54)
ln_Wages		0.509***
		(6.41)
ln_Nonwage_income		0.155***
		(7.46)
Popgrowth		0.511**
		(2.22)
Observations	18,151	18,151
Within R-squared	0.016	0.084
Fixed Effects	County & Year	County & Year
Cluster	State	State

Table 6 Alternative Treatment Variables

- Overall factors:
  - State with high percentage of IRS staff cuts (NTEU data)
  - States with high levels of sequestration cuts
- Enforcement factors:
  - Counties with high audit rates
  - States with higher income tax rate
- Taxpayer assistance services factors:
  - Counties where nearest TAC closed
  - States with relatively higher IRS.GOV web searches (Google Trends)

We find consistent results across all these variables of lower filings after IRS staff cuts in geographic areas more affected.

Table 6 Alternative Treatment Variables

	Overall Factors		Enforcement Factors		Taxpayer Assistance Services Factors	
	(1)	(2)	(3)	(4)	(5)	(6)
	ln_Filers	ln_Filers	ln_Filers	ln_Filers	ln_Filers	ln_Filers
High_staff_cuts × Post2011	-0.022***					
	(-8.80)					
High_sequestration_cuts × Post2011		-0.010***				
		(-4.44)				
High_county_audit_rate × Post2011			-0.036***			
			(-12.57)			
High_state_tax_rate × Post2011				-0.011***		
				(-3.83)		
Nearest_TAC_closed × Post2011					-0.026***	
					(-3.63)	
High_irs.gov_google_search × Post2011						-0.017***
						(-4.66)

Additional Analyses

- Alternative event years
  - 1995 (IRS experienced ~5% staff cut)
  - 2020 (pandemic limited in-person taxpayer assistance): Reduction in filings stronger in income-tax states
  - 2007 falsification test (SEC but not IRS staff cuts): insignificant results
- 2006-2022 sample period
  - Positive relationship between IRS staffing levels (overall, and disaggregated into taxpayer services, enforcement, and operations support) and filings is stronger in income-tax states
- Parallel trends analysis
  - We do not observe any violation of the parallel trends assumption

Conclusion

- Timely evidence on the effects of IRS staffing levels on individual taxpayer compliance, through both enforcement and taxpayer service channels
- We find that 11% IRS workforce reductions starting in 2011 result in 3% lower individual tax filings in states with income taxes
  - Effects are weaker in states with more tax examiners and collectors, consistent with the critical role of enforcement
  - Greater access to tax preparers (an alternative form of assistance) has a mitigating effect
  - Stronger effects in counties with a greater reliance on IRS Taxpayer Assistance Centers
- Contributions
  - Provide broad-based archival evidence on the roles of both enforcement and taxpayer assistance, consistent with the expanded taxpayer service paradigm of Alm et al. (2010)
  - Contribute to the literature on the impact of government agency resources on taxpayer behavior


thank you

Table 1, Panel A: Sample Selection

Criteria	County-Years
SOI county-year observations with populated number of returns filed and AGI for filing years 2008-2014 (excluding 2011)	18,682
Less: observations missing control variables	(530)
Less: singleton observations excluded from fixed effects models	(1)
Total County-Years	**18,151**

Table 1, Panel B: Descriptive Statistics

Variable	N	Mean	Std Dev	25th Pctl	50th Pctl	75th Pctl
State_IIT	18,151	0.81	0.39	1.00	1.00	1.00
Post2011	18,151	0.50	0.50	0.00	1.00	1.00
Filers	18,151	38,315	85,350	4,510	10,640	28,340
Reported_nonwage_income (1000's)	18,151	573,858	1,427,831	52,667	129,042	373,690
Employment	18,151	35,549	87,339	3,103	8,081	23,232
Establishments	18,151	2,313	5,352	280	628	1,632
Wages (1000's)	18,151	1,544,351	4,373,727	93,498	260,223	832,489
Nonwage_income (1000's)	18,151	1,403,393	3,026,053	199,715	431,883	1,071,603
Avg_AGI (1000's)	18,151	46.46	11.69	38.78	44.38	51.46
Popgrowth	18,151	0.003	0.013	-0.005	0.002	0.010

Table 7 Alternative Event Years

	1995 IRS Staffing Reduction		2020 COVID-19 Shock	2007 Falsification Test
	(1)	(2)	(3)	(4)
	ln_Filers	*ln_Filers*	*ln_Filers*	*ln_Filers*
State_IIT × Post1995	-0.019***	-0.043***		
	(-3.02)	(-5.99)		
State_IIT × Post1995 × High_internet		0.040***		
		(3.17)		
State_IIT × Post2020			-0.013**	
			(-2.22)	
State_IIT × Post2007				-0.008
				(-1.27)
Observations	18,230	800	17,827	17,847
Within R-squared	0.212	0.669	0.220	0.136
Fixed Effects	County & Year	County & Year	County & Year	County & Year
Cluster	State	State	State	State
Control Variables	Yes	Yes	Yes	Yes

Table 8 2006-2022 Sample Period

	(1)	(2)	(3)	(4)
	ln_Filers	*ln_Filers*	*ln_Filers*	*ln_Filers*
State_IIT × ln_IRS_Staffing	0.135***			
	(2.94)			
State_IIT × ln_TP_Services		0.167**		
		(2.34)		
State_IIT × ln_Enforcement			0.086***	
			(3.04)	
State_IIT × ln_Support_services				0.136**
				(2.35)
Observations	50,583	50,583	50,583	50,583
Within R-squared	0.357	0.354	0.357	0.353
Fixed Effects	County & Year	County & Year	County & Year	County & Year
Cluster	State	State	State	State
Control Variables	Yes	Yes	Yes	Yes

Fig. 3: Dynamic Analysis Around 2011 IRS Staffing Event

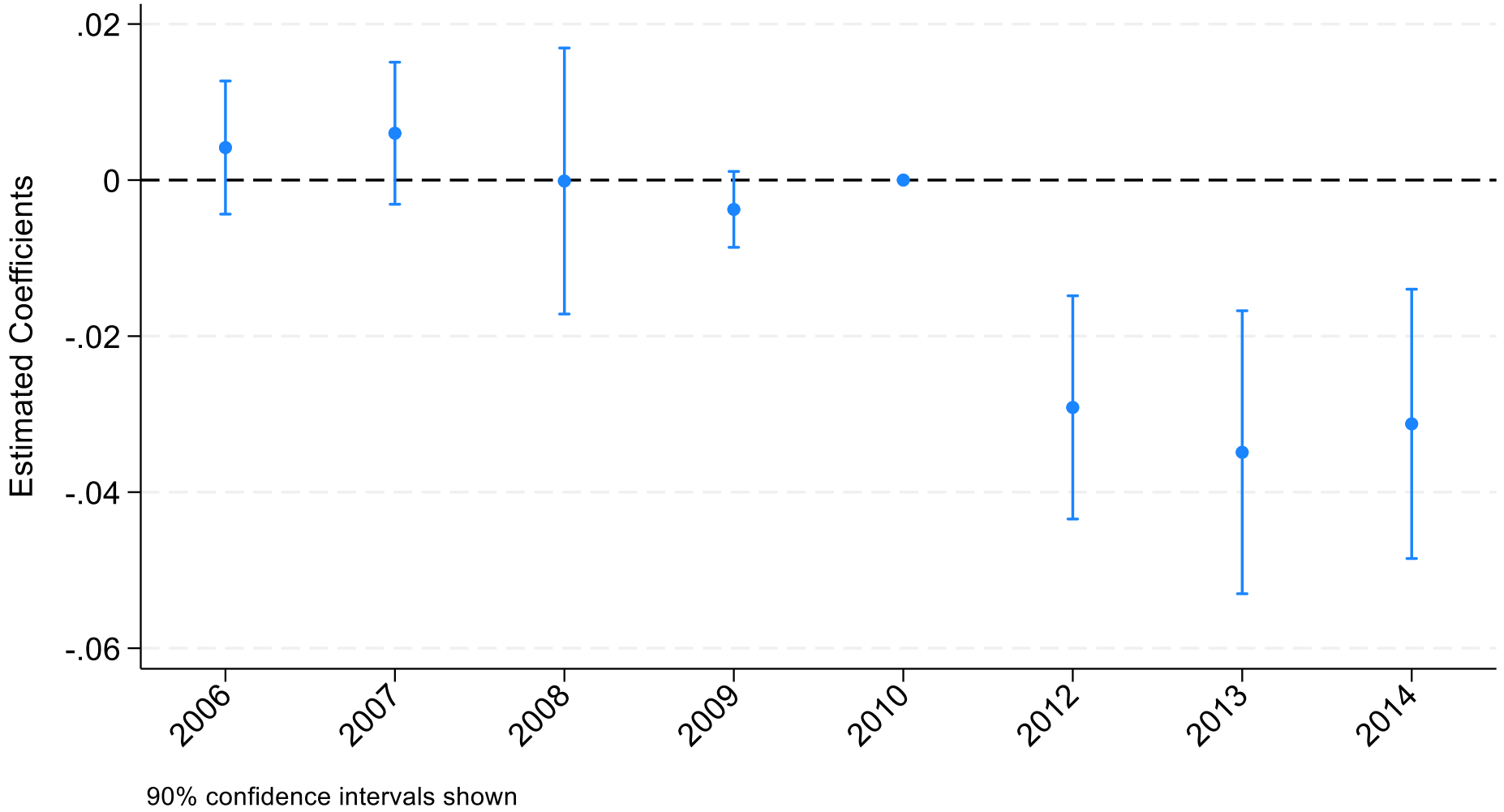
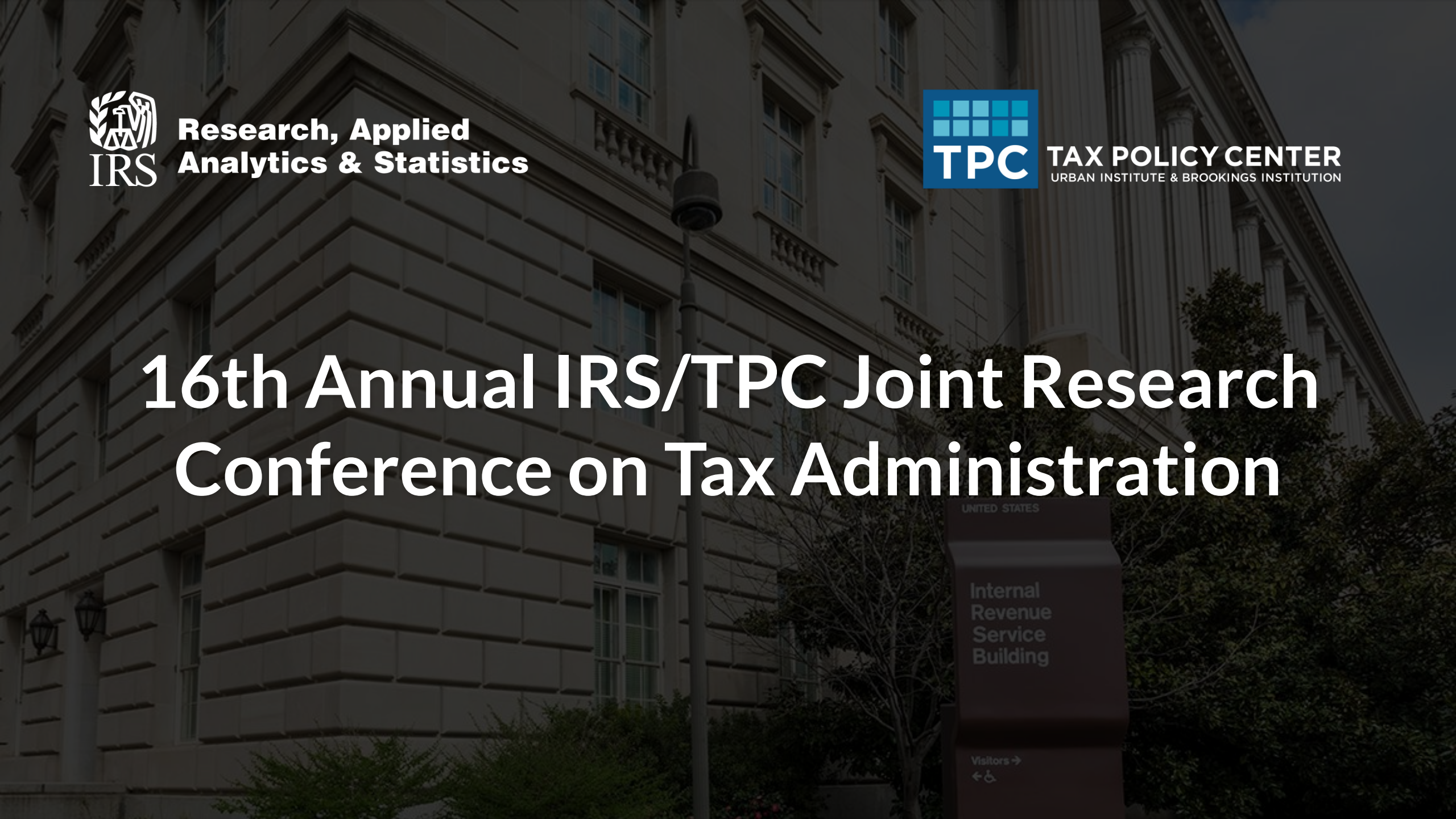


Table 9 Robustness Test: Trend Controls

	(1)	(2)
	ln_Filers	*ln_Filers*
State_IIT × Post2011	-0.029***	-0.030***
	(-4.74)	(-4.75)
State_IIT × trend	-0.001	0.000
	(-0.20)	(0.03)
State_IIT × trend_sq		-0.000
		(-0.15)
Observations	18,151	18,151
Within R-squared	0.198	0.198
Fixed Effects	County & Year	County & Year
Cluster	State	State
Control Variables	Yes	Yes



Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

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

From Self-Reporting to System-Reporting

Measuring Compliance Gains from Event-Based Tax Administration

The tax gap is also an information-architecture problem.

IRS Tax Policy Research Conference

Ali Ekmen — Former Chief Tax Inspector | MSc Taxation, University of Oxford

Takeaway: This presentation asks how the design of reporting changes the practical scope for misreporting.

PROBLEM

Opening question

Why is misreporting so low for some income categories, but so high for others?

Same tax system

Different income categories face very different reporting environments.

Different visibility

Some income is already visible before filing; some remains mainly in taxpayer records.

Different outcomes

Measured misreporting differs sharply across these environments.

Takeaway: The question is not only who the taxpayer is, but what the reporting system already knows.

Core argument

Compliance depends not only on taxpayer motivation, audits and penalties, but also on how tax-relevant information enters the system.

The paper does not reject behavioural theories of compliance.

It adds an institutional question: when, how and through whom does taxable information become visible?

Reporting architecture affects the opportunity for omission, error and reclassification.

Takeaway: Tax compliance is shaped by information flows as well as taxpayer choices.

PROBLEM

The conventional view is important — but incomplete

Behaviour

honesty, tax morale, trust

Deterrence

audit probability and penalties

Administration

risk selection and enforcement

These factors matter, but they do not fully explain why misreporting differs so sharply across income types.

Takeaway: The missing dimension is the reporting environment in which the taxpayer makes the reporting decision.

What this paper adds

From taxpayer motivation

Why does the taxpayer comply or not comply?

Focus: intention, incentives, morality and deterrence.

To reporting architecture

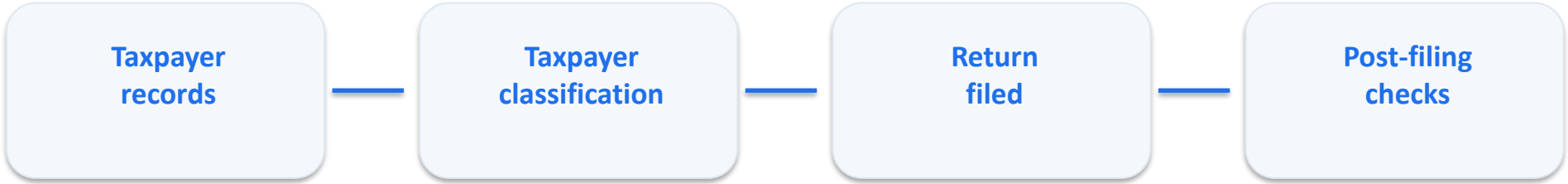
How much room does the system leave for omission, understatement or misclassification?

Focus: visibility, verification and timing.

Takeaway: The contribution is a shift from “taxpayer as decision-maker” to “taxpayer inside an information architecture”.

CONCEPT

Self-reporting: the return begins with taxpayer reconstruction



The taxpayer identifies income and expenses.

The taxpayer applies legal classifications.

The administration normally verifies after filing.

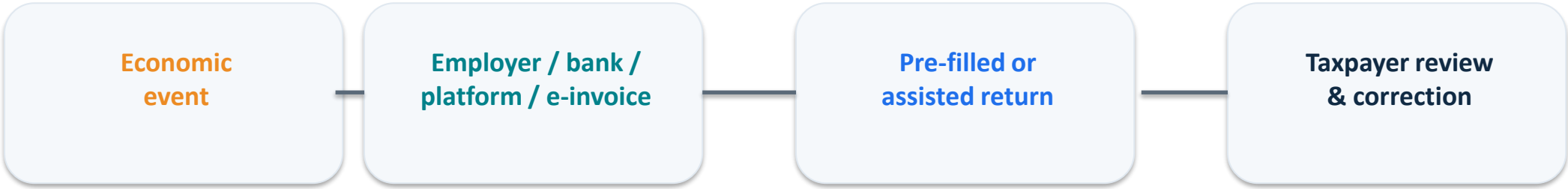
Risk

A digital return can still be self-reporting in substance if the information is reconstructed only at period end.

Takeaway: Self-reporting is not defined by paper forms; it is defined by who constructs the information base.

CONCEPT

System-reporting: the return begins with visible information



Information enters the tax system before or during filing.

The taxpayer remains responsible for the final tax position.

The role changes: from sole reporter to reviewer, corrector and completer.

Takeaway: System-reporting supports self-reporting; it does not abolish taxpayer responsibility.

CONCEPT

The key distinction

The question is not whether the tax return exists.
The question is where the tax return begins.

Blank self-declaration

The taxpayer creates the main information base from their own records.

System-visible facts

The return starts from information already reported, transmitted or verifiable.

Takeaway: This is the conceptual bridge from compliance theory to reporting design.

EVIDENCE

Why IRS tax gap data?

The IRS separates the gross tax gap into underreporting, non-filing and underpayment.

This distinction matters because each component reflects a different compliance problem.

Underreporting is closest to the paper's reporting-architecture question.

Underreporting

Incomplete or inaccurate reported liability

Non-filing

Required return not filed

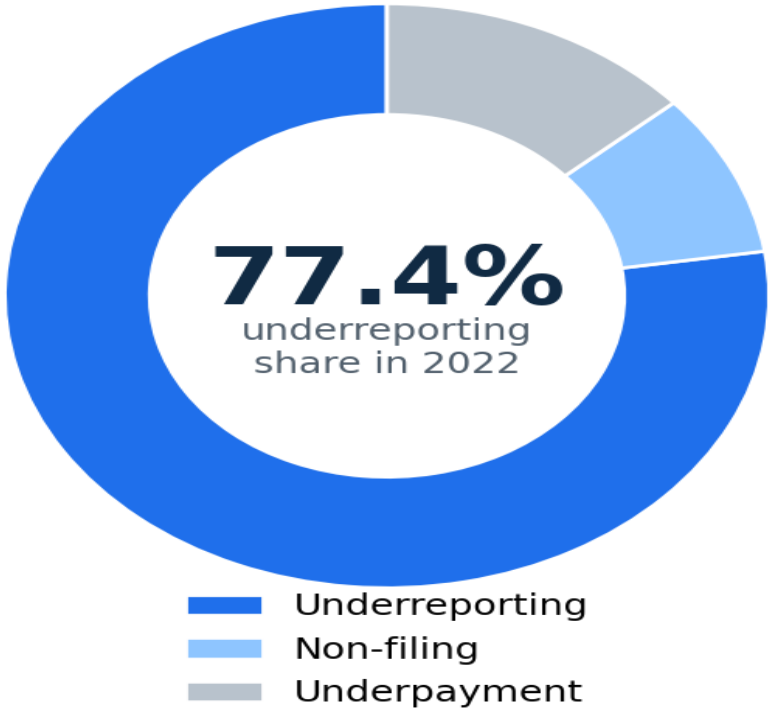
Underpayment

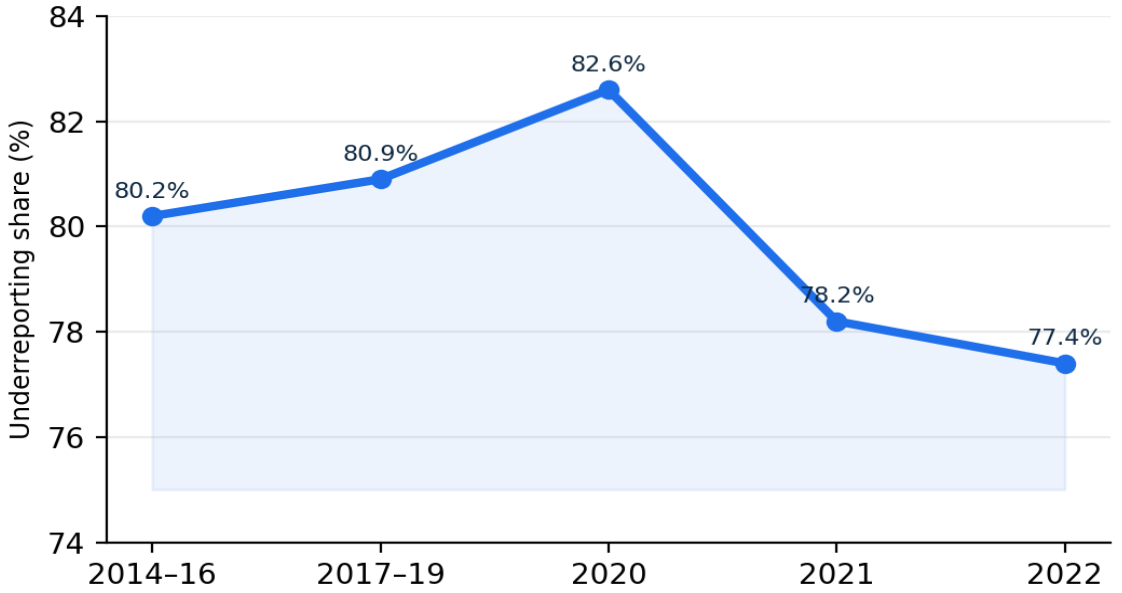
Reported or assessed tax not paid

Takeaway: The empirical focus is not payment enforcement alone; it is the accuracy of information in filed returns.

EVIDENCE

Underreporting dominates the US gross tax gap





2022 gross tax gap

Underreporting: $539bn | Non-filing: $63bn | Underpayment: $94bn

Takeaway: If underreporting is the dominant component, the design of information flows at the reporting stage is central.

Why underreporting matters for reporting architecture

Underreporting means the taxpayer enters the filing system, but the reported information is incomplete, inaccurate or understated.

The issue is not simply whether a return is filed.

The issue is whether the information in the return can be independently observed or verified.

This makes visibility a core compliance variable, not a technical detail.

Takeaway: The reporting stage is where system design can prevent errors and omissions from entering the return.

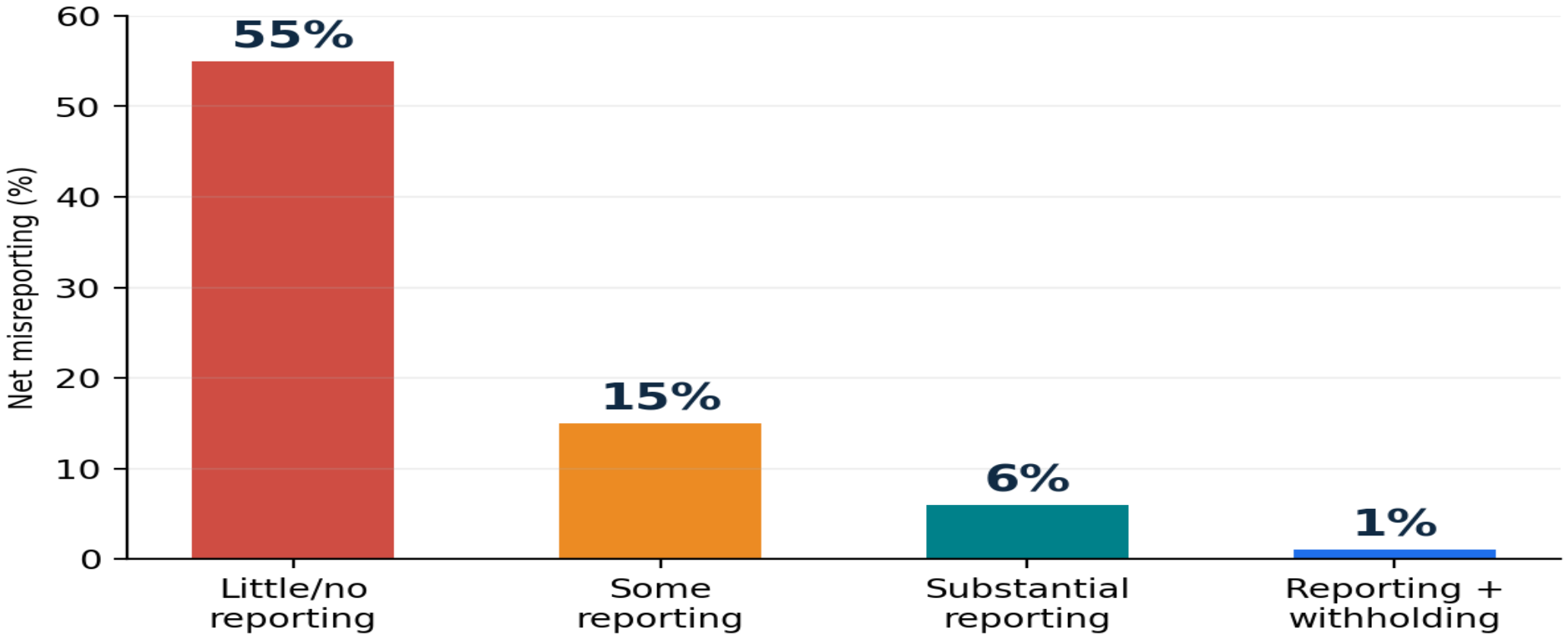
IRS visibility categories: the central evidence

IRS visibility category	Reporting intensity	Net misreporting
Substantial information reporting + withholding	4	1%
Substantial information reporting, no withholding	3	6%
Some information reporting	2	15%
Little or no information reporting	1	55%

Takeaway: The same tax system produces very different misreporting outcomes depending on information visibility.

EVIDENCE

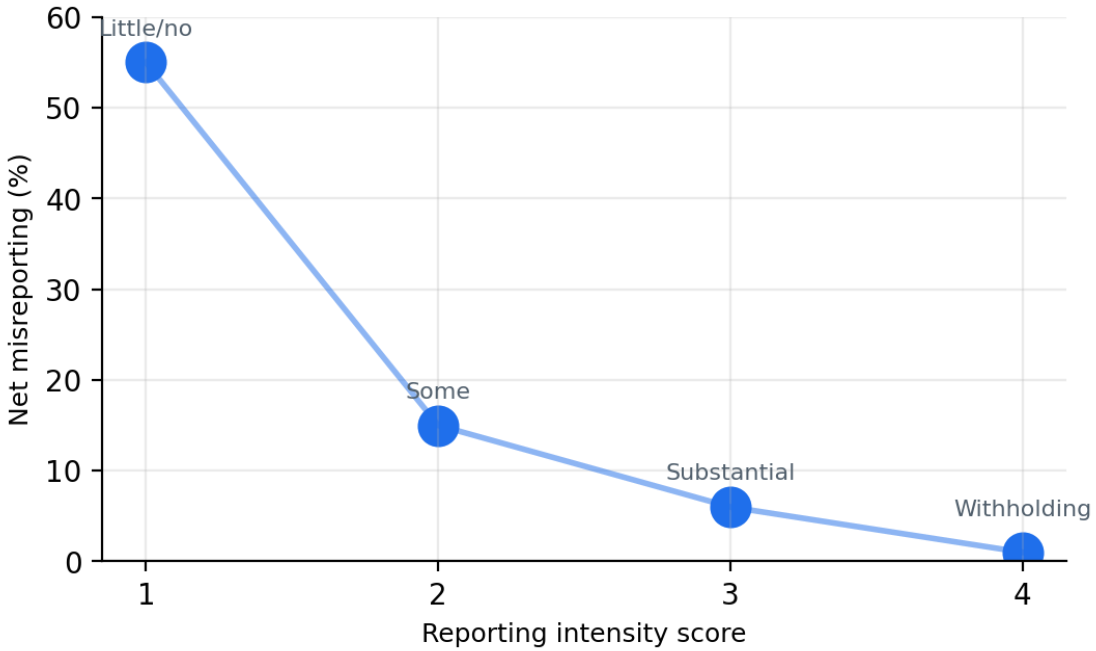
The visibility gradient



Takeaway: As information visibility falls, net misreporting rises sharply — from 1% to 55%.

EVIDENCE

Reporting intensity and net misreporting



Linear OLS: -17.10

Higher reporting intensity is associated with lower misreporting.

Log OLS: -1.29

The negative pattern remains after log transformation.

Spearman: -1.00

A perfect inverse ranking across the IRS categories.

Read these as descriptive summaries, not causal estimates.

Takeaway: The statistics formalise the pattern already visible in the IRS table and chart.

How to interpret the IRS evidence

What the evidence does show

Measured misreporting is lowest where income is already visible through third-party reporting and withholding, and highest where the administration depends mainly on unilateral self-declaration.

What it does not show

It does not prove that reporting intensity alone causes the entire difference. Income types also differ by cash intensity, deductions, legal complexity and business structure.

The claim is limited, but important: reporting environment is closely related to compliance outcomes.

Takeaway: This careful interpretation strengthens the paper because it avoids overstating causality.

Comparative consistency check

Canada

A large share of the federal tax gap arises at the reporting stage rather than only at the payment stage.

United Kingdom

HMRC behavioural categories show that reporting accuracy, error and care are central tax gap issues.

Denmark

Kleven et al. show very low evasion for third-party reported income and much higher evasion for self-reported income.

These sources are not directly comparable.

They are used as a consistency check, not as a cross-country regression.

Takeaway: The IRS pattern is not merely a US-specific curiosity; it reflects a broader tax administration issue.

From visibility evidence to reporting design

If reliable information is already visible, it should not be used only after filing.

Traditional use

Post-filing matching, audit selection and enforcement.

System-reporting use

Pre-filing visibility, assisted filing and taxpayer review.

Takeaway: The policy move is from retrospective control to earlier, structured visibility.

Concrete forms of system-reporting

Withholding
Tax connected directly to payment

Third-party reporting
Independent verification record

E-invoicing
Transaction traceability and matching

Pre-filled returns
Reduced omission and reporting error

Platform reporting
Visibility of platform-generated income

These mechanisms are not identical.

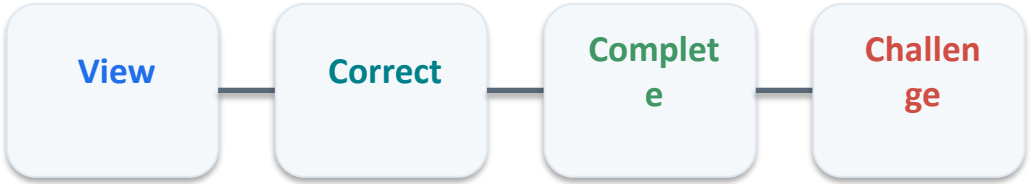
Their common function is to reduce reliance on a blank return prepared entirely from taxpayer records.

Takeaway: System-reporting is a family of mechanisms, not a single technology.

Limits and safeguards

System-reporting does not mean full automation

Visible income does not automatically determine taxable profit. Taxpayer input remains necessary for deductible expenses, mixed-use assets, timing, losses and legal classification.



Taxpayers must know who reported the information, why it appears in the return, and how it can be corrected.

Transparency and correction channels are legitimacy safeguards, not administrative extras.

Takeaway: The taxpayer's role changes, but procedural rights become more important, not less.

CONCLUSION

Final message

The tax return should not begin from a blank page.

**It should begin from system-visible facts,
subject to taxpayer correction, legal classification
and procedural safeguards.**

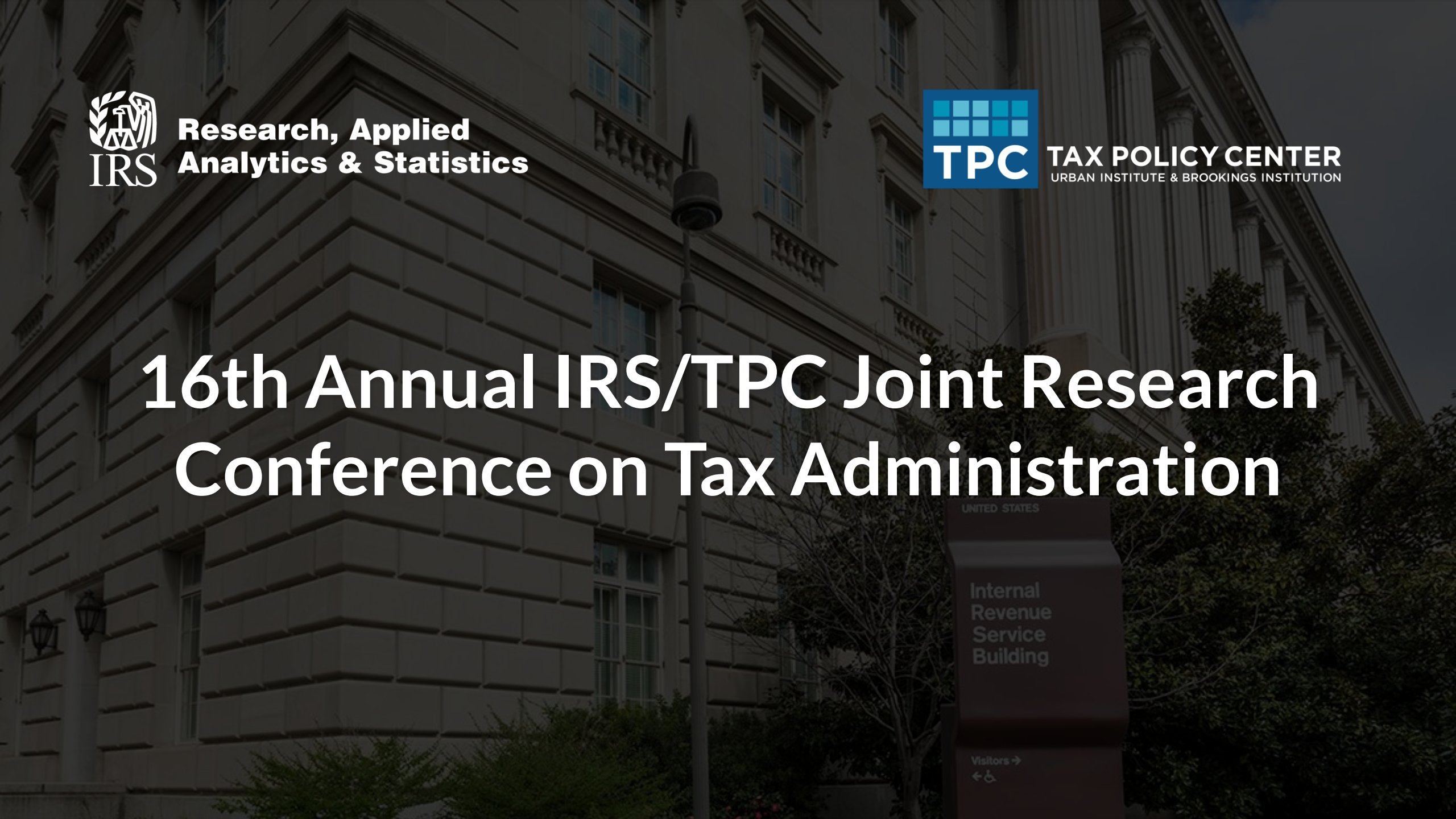
Not abolition of self-reporting

The return and taxpayer responsibility remain.

Change of starting point

From blank disclosure to assisted, verifiable reporting.

Takeaway: The design question is where reliable information first arises — and how the return can begin from that point.



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 2

Leave One Out Estimator for Auditor Detection of Underreported Tax

Ishani Roy and Alex Turk
RAAS- Knowledge Development and Application
Internal Revenue Service

2026 IRS-TPC Research Conference
June 25, 2026

Leave One Out Estimator for Auditor Detection of Underreported Tax

Problem: Auditors may be imperfect at detecting the “true” underreported tax

- Detection Controlled Estimation – Feinstein (1991), Erard and Feinstein (2011)
- Bayesian Shrinkage - Vossler *et. al.* (2025)

Objective: Develop a method for accounting for auditor effects in detecting reporting compliance that are

- Integrated into the reporting compliance model estimation
- Agnostic to assumptions about auditor human capital
- Agnostic to model specification
- Makes no assumptions about detectability based on visibility
- Makes no assumptions about the numbers of audits that an auditor must do
- Easily extends to models that incorporate non-NRP audit data
- Consistent with auditor having issue specific expertise

Overview of Presentation

- Discussion of the proposed estimation methods
  - Work in progress – simple models to demonstrate the methods
- Demonstrate the methods with synthetic data with engineered auditor effects
  - Synthetic data derived from linear model
  - Synthetic data derived from limited dependent variable model (i.e. Tobit)
- Demonstrate the methods with historical National Research Program data
  - Simple linear models
  - Separately for subpopulation of Individudal/Form 1040 tax returns.
- Discussion of next steps and other applications.

Leave One Out Estimator for Auditor Detection of Underreported Tax

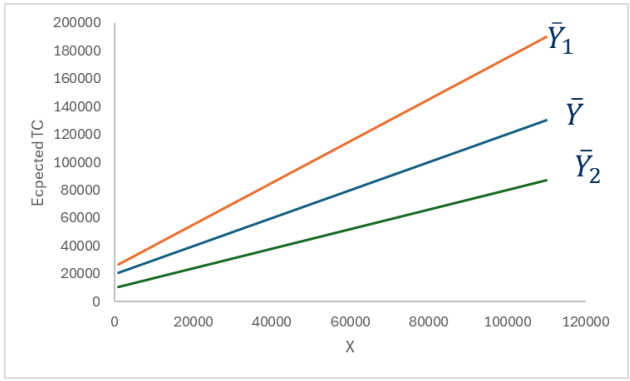
Illustrative Example: Suppose we have audit data and two different types of auditors; **Group 1** are good ('top') auditors and **Group 2** are average auditors and groups 1 and 2 are otherwise homogeneous. Let the Group 1 auditors handle α proportion of cases and Group 2 auditors handle remaining $(1-\alpha)$ proportion.

- $\bar{Y}_1$ be the expected audit outcome if group 1 does the audit
- $\bar{Y}_2$ be the expected audit outcome if group 2 does the audit
- Estimation of Group 1 outcome $\bar{Y}_1$ directly is not possible due to smaller sample size.

$$\bar{Y} = (\alpha\bar{Y}_1 + (1-\alpha)\bar{Y}_2)$$

Hence, we can rewrite as

$$\bar{Y}_1 = \alpha^{-1}(\bar{Y} - (1-\alpha)\bar{Y}_2)$$



We can estimate the audit outcomes of the Group 1 auditors if we can estimate that of Group 2 auditors.

Overall Methodology

- There are four parts:
  - Identify an overall model, ignoring the auditor
  - Identify auditor group labels (i.e., ‘top’ auditors vs. average auditors)
  - Estimate the model using only group 2 auditors (i.e., all the top auditors removed).
  - Estimation of the average predicted adjustment due to group1 auditors as a weighted difference between the overall model and the group 2 auditor model.

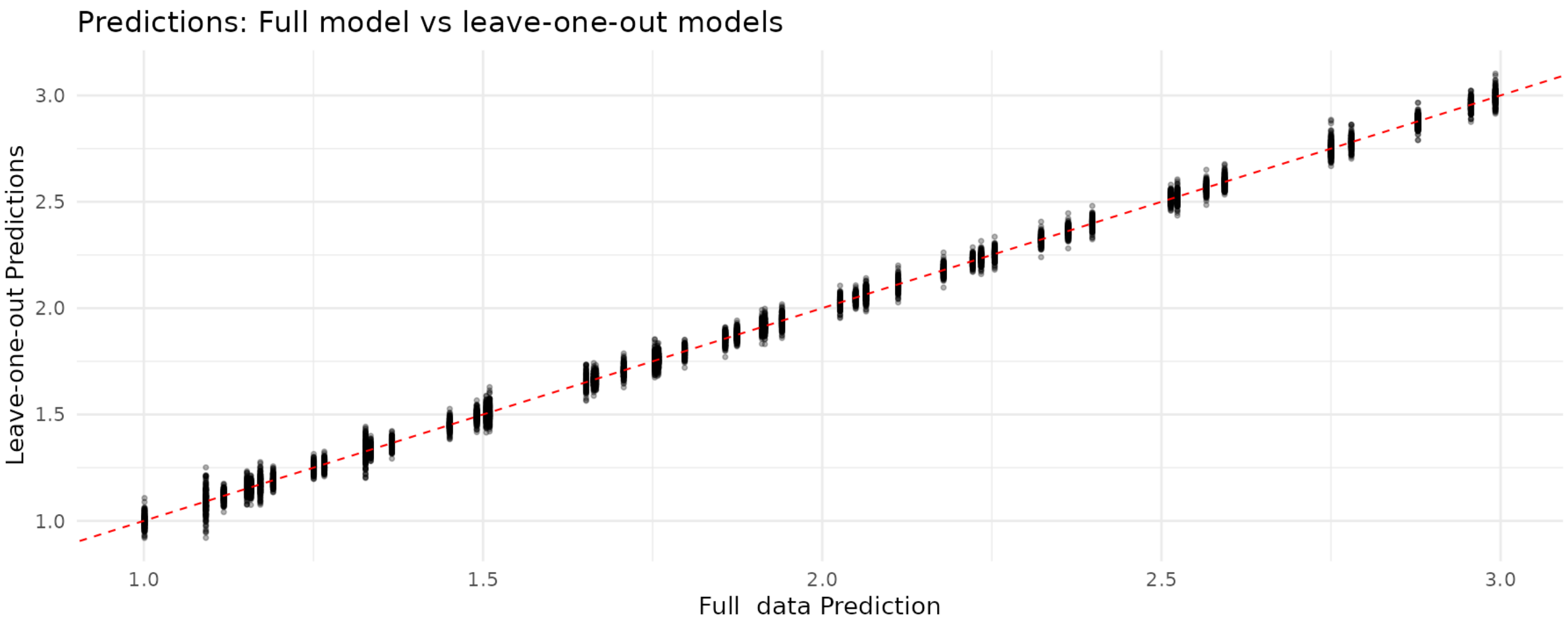
Identification of Group 1 'top' auditors with highest detection rate:

Assume there are I cases, with J different auditors

- Estimate an overall model with all cases
- Re-estimate the model J times, each time removing a different auditor's cases
- This produces a distribution of J different predictions for each case
- For a given case, the 'top' auditors are identified as those, when removed from the training sample, yield models that give smallest predictions for that case compared to the overall model.
- "Top" is defined that they have **K** predictions in the top **p%** of the **test data** set e.g. at least 10 times they are in the top 1% (largest negative) of the prediction distribution
- In other words, top means the larger negative impact on the prediction when they are held out.

Leave One Out Estimator for Auditor Detection of Underreported Tax

A visual depiction of how to identify the top auditors.



Incorporating the auditor detection rates into the model:

Three (at least three) options

- Estimate the model with only the top auditors
  - May not be feasible given with small samples
- Impose assumption on how auditor effect enter the model
  - Assumptions regarding which features have differential effects
- Allow all features to have differential effects
  - This is the approach we take in this paper

Leave One Out Estimator for Auditor Detection of Underreported Tax

Estimated average adjustment by 'Top' auditors and Multiplier for tax gap estimation

- Re-estimate the model removing all the "top" auditors to get $\bar{Y}_2$
- Assuming a uniform detection rate among the respective group of auditors, the average adjustment from the overall model can be written as a weighted aggregate from Group 1 model and Group 2 model

$$\bar{Y} = (\alpha\bar{Y}_1 + (1-\alpha)\bar{Y}_2).$$

- Inverting the relation, average adjustment due to Group 1 (top) auditors is

$$\bar{Y}_1 = \alpha^{-1}(\bar{Y} - (1-\alpha)\bar{Y}_2)$$

- The implied multiplier M is $\bar{Y}_1/\bar{Y}$ (in the dollar scale) and **exp($\bar{Y}_1 - \bar{Y}$)** in the log-dollar scale

SYTHETIC MODEL FRAMEWORK

True Model

- **Latent Change (log-scale):** We model the underlying positive change using a log transformation to stabilize variance and ensure positivity:

$$A_i = \log(\textit{positive true understatment})$$

- **Linear Model (log dollars):**

$$A_i = x_i'\beta + \varepsilon_i,\ \ \varepsilon_i \sim N(0, \sigma^2)$$

- **Observation Model**: The observed outcome reflects both the true latent change and how much of the change is detected:

$$Y_i = A_i + \log(D_i)$$

- $D_i \in [0, 1]$ represents the fraction of the change that is captured, introducing attenuation in the observed signal.

Synthetic Response Generation Overview

- Generate data using synthetic variables and known coefficients β and noise variance
- Auditors assigned to a random number of cases by assigning a random auditor ID to each case. The total number of cases is 170,000 and the number of auditors in the simulation experiment is 1500 with the number of assigned cases ranging from 60 to 240. Four synthetic features:
- Four synthetic features: X_1 and X_4 are continuous features randomly sampled from normal distributions and X_2 and X_3 are binary features sampled from Bernoulli distributions

$$A_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4 + \epsilon_i$$

- Generate synthetic response Y_i (in log-dollar) after incorporating imperfect detection

$$Y_i = A_i + log(D_i)$$

Leave One Out Estimator for Auditor Detection of Underreported Tax

- **Detection Fraction Model**

$$\boldsymbol{D_i = \Phi(W_i + \theta_2 \overline{A_i})}$$

Interpretation: D_i :fraction of true change that is observed

- Φ: standard normal CDF; Probit link ensures $D_i \in [0, 1]$
- W_i are examiner specific random effect; $\boldsymbol{W_i \sim N(-1 + \mu, \tau^2)}$
- θ_2 is very small or set to 0 initially
- $\overline{A_i}$:normalized case-specific effect $\overline{A_i} = \frac{(\max_j A_j - A_i)}{\max_j A_j}$

The parameter μ in the detection fraction model allows for ability to discriminate between good auditors and the rest. For good auditors μ is set to large positive so that detection fraction is approximately one (perfect detection)

Leave One Out Estimator for Auditor Detection of Underreported Tax

Table1: Linear Model Synthetic Data Summary Statistics

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Y (Mu=0)	-2.9403	0.4445	1.97	1.9694	3.4887	6.7203
Y (Mu=3)	-2.6635	0.7795	2.3339	2.3321	3.8537	10.987
X1	-13	32	40	40.02	48	94
X4	3	13	24	24.01	35	45
X2	84978 (Y)			85022(N)		
X3	113570(Y)			56430(N)		

Source: Author generated synthetic data

Illustration of Auditor Effect Using the Parametric Model:

- In the parametric model the auditor effect can be captured through the parameters
- Suppose Z is a dummy indicating a Group 2 auditor and $X_z = [ZX_1 : \cdots : ZX_p]$, Consider the expanded model with interactions

$$Y = X\beta + X_z\beta_z + \epsilon$$

- The main effect in the interaction model β is the same as the coefficient in the model with only Group 1 observations; the interaction parameter β_z is the difference in the coefficients between Group 2 observations and Group 1 observations.

$$\begin{pmatrix} \hat{\beta} \\ \hat{\beta}_z \end{pmatrix} = \begin{pmatrix} \hat{\beta}_{(1)} \\ \hat{\beta}_{(2)} - \hat{\beta}_{(1)} \end{pmatrix}$$

Leave One Out Estimator for Auditor Detection of Underreported Tax

Table 2: Linear Model Parameter Estimates Under Varying Auditor Selection Rules with no Auditor Effects*

Mu = 0							
Coeff	TRUE	ORACLE	K=1	K=10	K=50	K=100	NAÏVE
B0	-0.16	-2.039	-1.968	-1.967	-1.963	-1.97	-2.013
B1	0.01	0.011	0.01	0.01	0.01	0.01	0.01
B2	-0.005	0.006	0.002	0.003	0.003	0.002	-0.003
B3	0.14	0.147	0.135	0.135	0.135	0.137	0.14
B4	0.145	0.145	0.145	0.145	0.145	0.145	0.145
B0z	0	0.033	-0.055	-0.056	-0.061	-0.052	0
B1z	0	-0.001	0	0	0	0	0
B2z	0	-0.012	-0.007	-0.007	-0.007	-0.006	0
B3z	0	-0.009	0.006	0.006	0.006	0.004	0
B4z	0	0	0	0	0	0	0

Source: Analysis of author generated synthetic data

*The first column gives the true coefficients, and the second column gives those where auditor group in know. The final column gives the estimates under the naïve model that ignores the auditor effect. The auditor effect parameter is zero, i.e. no difference between Group 1 and Group 2 auditors and all auditor under-detect.

Table 3: Linear Model estimates of Parameters Under Varying Auditor Selection Rules with Auditor Effects*

Mu = 3							
Coeff	TRUE	ORACLE	K=1	K=10	K=50	K=100	NAÏVE
B0	-0.16	-0.213	-0.181	-0.173	-0.177	-0.177	-1.639
B1	0.01	0.011	0.01	0.01	0.01	0.01	0.01
B2	-0.005	0.006	0.015	0.013	0.014	0.014	-0.007
B3	0.14	0.149	0.136	0.136	0.14	0.143	0.145
B4	0.145	0.145	0.145	0.145	0.145	0.145	0.145
B0z	0	-1.793	-1.548	-1.546	-1.538	-1.536	0
B1z	0	-0.001	0	0	0	0	0
B2z	0	-0.011	-0.02	-0.019	-0.02	-0.02	0
B3z	0	-0.011	0.01	0.009	0.003	0.001	0
B4z	0	0	0	0	0	0	0

Source: Analysis of author generated synthetic data

. *The first column gives the true coefficients, and the second column gives those where auditor group in know. The final column gives the estimates under the naïve model that ignores the auditor effect. The auditor effect parameter is 3, i.e., 20% of all auditors are Group 1 auditors.

Leave One Out Estimator for Auditor Detection of Underreported Tax

Table 4: Estimated Multipliers by Level of Auditor Effect (Mu) and the Number of Times the Auditor is Observed as a "Group 1" Auditor (k)*

Mu	Multiplier				
	TRUE	k = 1	k=10	k=50	k=100
0	1.002	1.026	1.027	1.028	1.029
1	2.514	2.533	2.527	2.530	2.530
2	3.819	3.829	3.83	3.832	3.832
3	4.310	4.321	4.322	4.319	4.323
5	4.392	4.402	4.404	4.401	4.405

Source: Analysis of author generated synthetic data

* The percentage of top auditor bracket, p% is fixed at 1%.

Leave One Out Estimator for Auditor Detection of Underreported Tax

A Modified True Model Accounting For Zero Detection:

- **Latent Change (log-scale):**

We model the latent change (log-scale):

$$A_i^* = x_i'\beta + \varepsilon_i,\ \ \varepsilon_i \sim N(0, \sigma^2)$$

- **TOBIT Model:**

$$A_i = \begin{cases} 0 & \text{if } A_i^* \le 0 \\ A_i^* & \text{otherwise} \end{cases}$$

The observations are censored if the latent change is nonpositive.

Observation Model:

The observed outcome reflects censoring of the true latent change when affected by an auditor effect:

$$Y_i = \begin{cases} 0 & \text{if } A_i^* + \log(D_i) \le 0 \\ A_i^* + \log(D_i) & \text{otherwise} \end{cases}$$

Synthetic Response Generation Overview

- **Goal:**
  Generate observed response Y_i (in dollar scale) from latent change A_i ,incorporating imperfect detection and measurement attenuation.

Detection fraction:

- Detection fraction D_i:
  Continuous factor in $[0, 1]$representing the proportion of the true latent change that is captured when detection occurs.
- **Final synthetic response (in log-dollars):**

$$Y_i = \begin{cases} 0 & \text{if } A_i^* + \log(D_i) \leq 0 \\ A_i^* + \log(D_i) & \text{otherwise} \end{cases}$$

Leave One Out Estimator for Auditor Detection of Underreported Tax

Table 5: Estimated Multipliers by Level of Auditor Effect (Mu) for different p and k

Mu = 0, True Multiplier = 0.9834951			
p\k	5	10	20
0.01	1.094177	1.115504	1.11796
0.02	1.075357	1.08809	1.108023
0.03	1.073005	1.077368	1.098079

Mu = 3, True Multiplier = 1.687599			
p\k	5	10	20
0.01	1.43726	1.529661	1.829168
0.02	1.313453	1.379641	1.417914
0.03	1.248803	1.295772	1.3295

Source: Analysis of author generated synthetic Tobit data

NRP DATA ANALYSIS

Data and Model

- We used TY 2010-TY2015 NRP data. Data from TY2010-TY2014 were used for training and TY20015 were used as test data.
- Analysis was done by activity code.
- For illustration, results are presented for EITC returns (AC 270)
  - Linear model of Log of positive tax adjustment (Y=log(max(aud_res_sum,1)))
  - 20,699 observations in the training sample and 2,317 observations in the test sample.
  - There are 3,318 distinct auditors in the training sample.
- Multiple linear regression is used as the predictive model. However, more complex predictive model can be easily incorporated in the methodology.

Table 6: Estimation of Multiplier and Associated Parameters for Form 1040 EITC return under Varying Auditor Selection Rules*

Returns with Earn Income Tax Credit > $0, Total Gross Receipts below $25K (Activity Code 270)					
Number of examiner = 3,318	Response variable Y=log(max(Tax Change ,1))				
	Top Group (p%,k)				
	1%, k=1	1%, k-10	1%, k = 50	1%,k=100	5% k= 1
Proportion of examiners in Top Group	0.2	0.12	0.07	0.05	0.53
Proportion of cases in Top Group (α)	0.46	0.34	0.24	0.2	0.81
Mean Predicted Adjustment from Full Model (Ybar)	3.63	3.63	3.63	3.63	3.63
Predicted Mean Adjustment removing Top Group (Ybar2)	3.20	3.24	3.31	3.35	2.49
Estimated Mean Adjustment for Top Group (Ybar1)	4.14	4.39	4.67	4.77	3.89
Multiplier: exp (Ybar1 - Ybar)	1.66	2.13	2.82	3.13	1.3

Source: Authors analysis of National Research Program closed audits, Tax Years 2010-2015

**The estimates are illustrative only. The underlying model is overly simplified and does not account for the censoring of tax change*

Leave One Out Estimator for Auditor Detection of Underreported Tax

*Table 7: Estimation of Multiplier for Linear Model of Tax Change for Form 1040 Returns Under Varying Auditor Selection Rules**

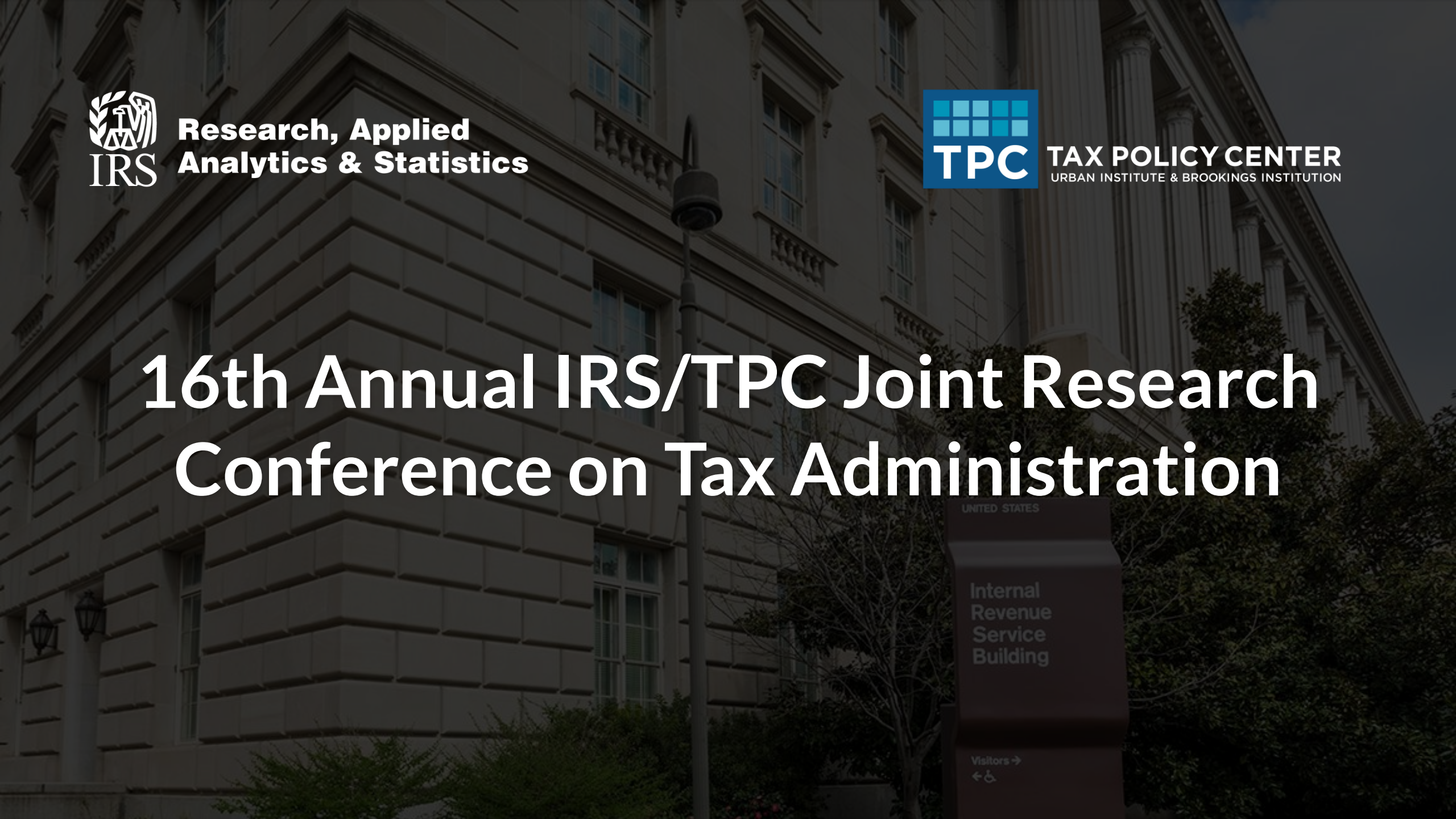
Activity Code		**Multiplier: exp (Ybar1 - Ybar)**				
		Response variable Y=log(max(Tax Change ,1))				
	Definition	1%, k=1	1%, k-10	1%, k = 50	1%,k=100	5% k= 1
270	Earn Income Tax Credit > $0, Total Gross Receipts below $25K	1.66	2.13	2.82	3.13	1.3
271+274	Earn Income Tax Credit >0, Total Gross Receipts > $25K or Activity Code 274	1.45	1.81	2.12	2.08	1.20
272	TPI < $200K, no schedule C,E,F or 2106	1.99	2.27	3.87	4.81	1.47
273	TPI > $200K, no schedule C,E,F or 2106	2.69	4.04	5.81	5.99	1.58
274	Sch C; Total Gross Receipts < $25K; TPI < $200K	1.57	1.96	2.59	2.74	1.22
275	Sch C; Total Gross Receipts between 25K and $100K; TPI<$200K	1.43	2.36	2.40	3.46	1.22
276	Sch C; Total Gross Receipts between $100K and $200K; TPI<$200K	1.18	1.83	3.06	2.63	1.10
277	Sch C; Total Gross Receipts over $200K; TPI<$200K	1.24	1.83	3.06	5.19	1.12
278	Sch F Not Classified Elsewhere; TPI<$200K	1.82	2.51	3.71	4.99	1.35
279	No Sch C/F; TPI between $200K and $1M	2.76	3.62	5.16	7.31	1.77
280	Sch C/F; TPI between $200K and $1M	1.99	3.43	6.14	9.33	1.47
281	TPI over $1M	2.91	4.18	4.85	5.66	2.10

Source: Authors analysis of National Research Program closed audits, Tax Years 2006-2015

*The estimates are illustrative only. The underlying model is overly simplified and does not account for the censoring of tax change

Summary

- Data driven method identifying auditors with higher detection and for estimating a detection corrected underreporting
- The results conform to some of the historical values of multipliers
- The method does not assume a tax compliance model; illustrations use paremtric regressions, but other complex models can be substituted
- Operational audits can be also used in the same manner as the NRP data.
  - Either integrated into the tax underreporting model or
  - As auxiliary data set to identify “good” auditors
- More analysis needed to refine the methods for determine the group of top auditors
- Approach could be applied in estimate differential treatment effects e.g. Judges Model



Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

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

Refresh Rate Matters

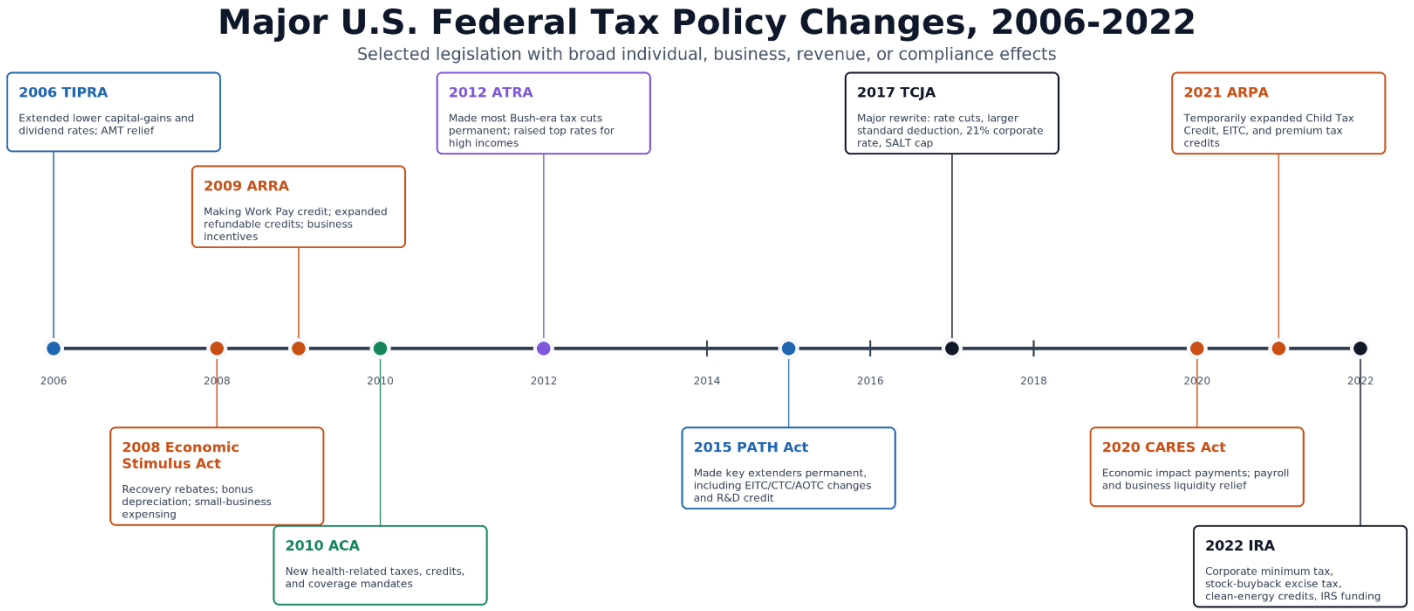
Monthly Retraining Improves Audit-Selection Under Concept Drift

Brandon Anderson[1], Austin Miller[1], Nan Xu[1], Alex Turk[1]

[1]Internal Revenue Service – RAAS

2026 IRS-Tax Policy Center Research Conference on Tax Administration

A Dynamic Environment



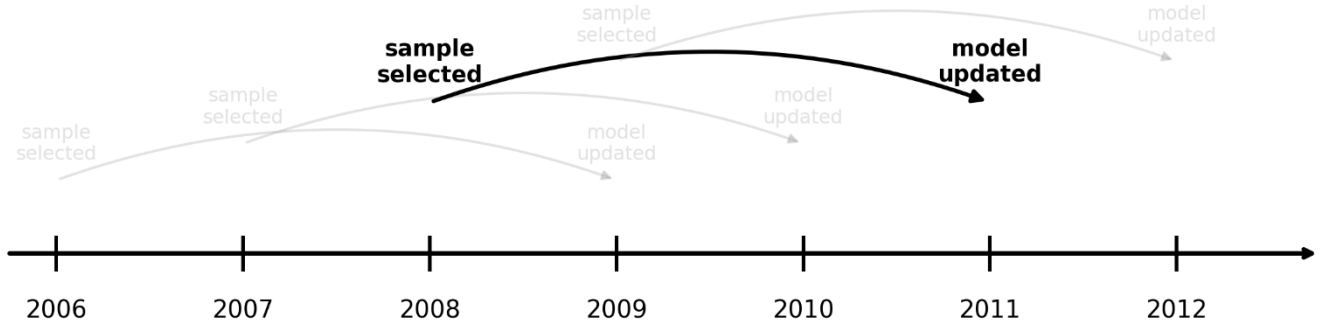
Concept drift sources

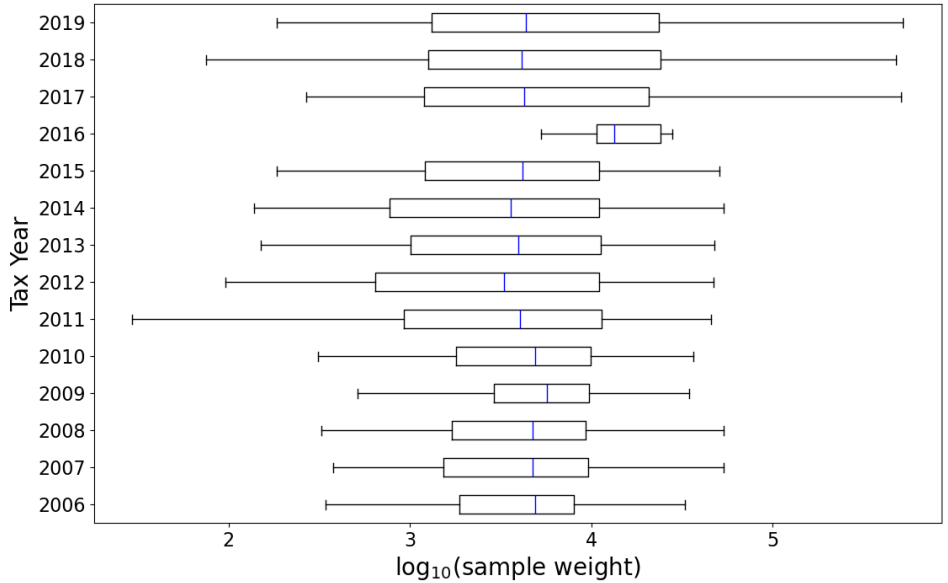
- Policy – features, thresholds, and meanings all change
- Behavior – tendencies (think TCJA + standard deduction), schemes, understanding, etc.

Data Flow

National Research Program

- Appx. 8000-15,000 stratified random audits per year
- Each sample takes 3 years to complete (at 99%)
- NRP-based models are updated only with complete samples
- Cases currently close at an approximately even rate





Sequential Decision-Making

agent – a self-contained ‘mini-IRS’

- Selects audits from population
- Maintains its own risk model
- Tracks outcomes

policy – rules that dictate agent behavior

- How to turn risk into audit selections
- How often to update models
- What data to update models with

environment – simulated state of operations

- What population is available
- When do cases open and close

Historical Playback

Pseudocode

initialize simulation environment and agents

for each scheduled time step:
    update environment

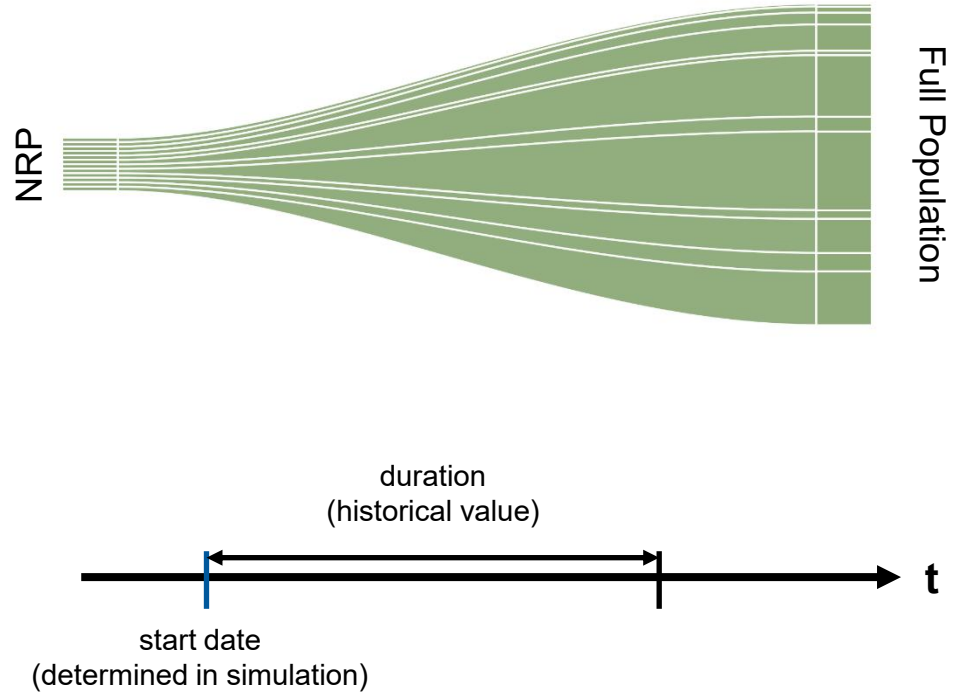
    for each agent:
        observe current population and prior outcomes
        *update model using available outcomes
        *choose new cases under budget constraints

*policy-dependent

Historical Playback

Simulation Environment

1. Inflate NRP to give full yearly populations
2. Cases can be opened at any point within their original processing year and close after their original duration
3. Limit number of simultaneous open cases
4. Idealize workforce



Policy Space

Contextual Decisions

- How much does updating the model matter at all?
- Some concern over (case-length) fixation. Is exploration an effective counter?

Table 2: Policy dimensions defining Π.

Dimension	Values
Retraining cadence	monthly, annual
Sample delay	immediate, 3-year
Exploration	$\varepsilon \in \{0,\ 0.10\}$
Updating regime	ongoing, one-shot

Overall Results

Metrics:

- Total Yield
- No-Change Rate
- ROI/case-hour
- ROI/case
- Marginal versions

Overall				Top 10%		
Yield	NCR	ROI_t	ROI_c	NCR@10%	ROI_t@10%	ROI_c@10%

Overall Results

Metrics:

- Total Yield
- No-Change Rate
- ROI/case-hour
- ROI/case
- Marginal versions

Policies:

- **High Cadence Updates**
- Sample Delay
- Epsilon Exploration
- Model Updates

Table 3: Policy Scorecard. Best performers under each metric are bold-faced and the row for the baseline model is highlighted. Rows are sorted by ROI_{case}.

Policy				Overall				Top 10%		
HC	Delay	ε	Upd	Yield	NCR	ROI_t	ROI_c	NCR@10%	ROI_t@10%	ROI_c@10%
1	0	0.1	1	**$186.2B**	**0.32**	**$301**	**$20,110**	**0.33**	$626	$26,588
1	0	0.0	1	$158.6B	0.33	$244	$17,962	0.38	**$1,103**	**$35,089**
0	0	0.1	1	$146.1B	0.38	$242	$16,128	0.40	$297	$20,078
0	0	0.0	1	$141.3B	0.37	$227	$15,522	0.39	$339	$24,768
0	0	0.0	0	$132.4B	0.37	$227	$14,628	0.40	$253	$16,154
0	0	0.1	0	$127.5B	0.40	$228	$13,704	0.41	$243	$14,352
0	1	0.1	1	$131.6B	0.43	$238	$13,589	0.44	$279	$16,557 ← benchmark
1	1	0.1	1	$127.7B	0.44	$229	$13,161	0.46	$288	$16,708
0	1	0.0	1	$126.5B	0.45	$235	$12,875	0.45	$310	$18,763
0	0	1.0	0	$18.8B	0.58	$68	$1,504	0.58	$70	$1,496

"Generally superior except when coupled with delay."

Overall Results

Metrics:

- Total Yield
- No-Change Rate
- ROI/case-hour
- ROI/case
- Marginal versions

Policies:

- High Cadence Updates
- **Sample Delay**
- Epsilon Exploration
- Model Updates

Table 3: Policy Scorecard. Best performers under each metric are bold-faced and the row for the baseline model is highlighted. Rows are sorted by ROI_{case}.

Policy				Overall				Top 10%		
HC	Delay	ε	Upd	Yield	NCR	ROI_t	ROI_c	NCR@10%	ROI_t@10%	ROI_c@10%
1	0	0.1	1	**$186.2B**	**0.32**	**$301**	**$20,110**	**0.33**	$626	$26,588
1	0	0.0	1	$158.6B	0.33	$244	$17,962	0.38	**$1,103**	**$35,089**
0	0	0.1	1	$146.1B	0.38	$242	$16,128	0.40	$297	$20,078
0	0	0.0	1	$141.3B	0.37	$227	$15,522	0.39	$339	$24,768
0	0	0.0	0	$132.4B	0.37	$227	$14,628	0.40	$253	$16,154
0	0	0.1	0	$127.5B	0.40	$228	$13,704	0.41	$243	$14,352
0	1	0.1	1	$131.6B	0.43	$238	$13,589	0.44	$279	$16,557
1	1	0.1	1	$127.7B	0.44	$229	$13,161	0.46	$288	$16,708
0	1	0.0	1	$126.5B	0.45	$235	$12,875	0.45	$310	$18,763
0	0	1.0	0	$18.8B	0.58	$68	$1,504	0.58	$70	$1,496

"The worst performers aside from random selection."

Overall Results

Metrics:

- Total Yield
- No-Change Rate
- ROI/case-hour
- ROI/case
- Marginal versions

Policies:

- High Cadence Updates
- Sample Delay
- **Epsilon Exploration**
- Model Updates

Table 3: Policy Scorecard. Best performers under each metric are bold-faced and the row for the baseline model is highlighted. Rows are sorted by ROI_{case}.

Policy				Overall				Top 10%		
HC	Delay	ε	Upd	Yield	NCR	ROI_t	ROI_c	NCR@10%	ROI_t@10%	ROI_c@10%
1	0	0.1	1	**$186.2B**	**0.32**	**$301**	**$20,110**	**0.33**	$626	$26,588
1	0	0.0	1	$158.6B	0.33	$244	$17,962	0.38	**$1,103**	**$35,089**
0	0	0.1	1	$146.1B	0.38	$242	$16,128	0.40	$297	$20,078
0	0	0.0	1	$141.3B	0.37	$227	$15,522	0.39	$339	$24,768
0	0	0.0	0	$132.4B	0.37	$227	$14,628	0.40	$253	$16,154
0	0	0.1	0	$127.5B	0.40	$228	$13,704	0.41	$243	$14,352
0	1	0.1	1	$131.6B	0.43	$238	$13,589	0.44	$279	$16,557
1	1	0.1	1	$127.7B	0.44	$229	$13,161	0.46	$288	$16,708
0	1	0.0	1	$126.5B	0.45	$235	$12,875	0.45	$310	$18,763
0	0	1.0	0	$18.8B	0.58	$68	$1,504	0.58	$70	$1,496

"Improves every policy except for frozen models."

Overall Results

Metrics:

- Total Yield
- No-Change Rate
- ROI/case-hour
- ROI/case
- Marginal versions

Policies:

- High Cadence Updates
- Sample Delay
- Epsilon Exploration
- **Model Updates**

Table 3: Policy Scorecard. Best performers under each metric are bold-faced and the row for the baseline model is highlighted. Rows are sorted by ROI_{case}.

Policy				Overall				Top 10%		
HC	Delay	ε	Upd	Yield	NCR	ROI_t	ROI_c	NCR@10%	ROI_t@10%	ROI_c@10%
1	0	0.1	1	**$186.2B**	**0.32**	**$301**	**$20,110**	**0.33**	$626	$26,588
1	0	0.0	1	$158.6B	0.33	$244	$17,962	0.38	**$1,103**	**$35,089**
0	0	0.1	1	$146.1B	0.38	$242	$16,128	0.40	$297	$20,078
0	0	0.0	1	$141.3B	0.37	$227	$15,522	0.39	$339	$24,768
0	0	0.0	0	$132.4B	0.37	$227	$14,628	0.40	$253	$16,154
0	0	0.1	0	$127.5B	0.40	$228	$13,704	0.41	$243	$14,352
0	1	0.1	1	$131.6B	0.43	$238	$13,589	0.44	$279	$16,557
1	1	0.1	1	$127.7B	0.44	$229	$13,161	0.46	$288	$16,708
0	1	0.0	1	$126.5B	0.45	$235	$12,875	0.45	$310	$18,763
0	0	1.0	0	$18.8B	0.58	$68	$1,504	0.58	$70	$1,496

"Frozen models surprisingly outperform updates with delay."

Overall Results

One clear winner:

- ✓ High Cadence Updates
- ✓ No Sample Delay
- ✓ Epsilon Exploration

Table 3: Policy Scorecard. Best performers under each metric are bold-faced and the row for the baseline model is highlighted. Rows are sorted by ROI_{case}.

Policy				Overall				Top 10%		
HC	Delay	ε	Upd	Yield	NCR	ROI_t	ROI_c	NCR@10%	ROI_t@10%	ROI_c@10%
1	0	0.1	1	**$186.2B**	**0.32**	**$301**	**$20,110**	**0.33**	$626	$26,588
1	0	0.0	1	$158.6B	0.33	$244	$17,962	0.38	**$1,103**	**$35,089**
0	0	0.1	1	$146.1B	0.38	$242	$16,128	0.40	$297	$20,078
0	0	0.0	1	$141.3B	0.37	$227	$15,522	0.39	$339	$24,768
0	0	0.0	0	$132.4B	0.37	$227	$14,628	0.40	$253	$16,154
0	0	0.1	0	$127.5B	0.40	$228	$13,704	0.41	$243	$14,352
0	1	0.1	1	$131.6B	0.43	$238	$13,589	0.44	$279	$16,557
1	1	0.1	1	$127.7B	0.44	$229	$13,161	0.46	$288	$16,708
0	1	0.0	1	$126.5B	0.45	$235	$12,875	0.45	$310	$18,763
0	0	1.0	0	$18.8B	0.58	$68	$1,504	0.58	$70	$1,496

"Best by every metric except upper margin (epsilon dilution)."

Timeline Breakdown

Cold Start

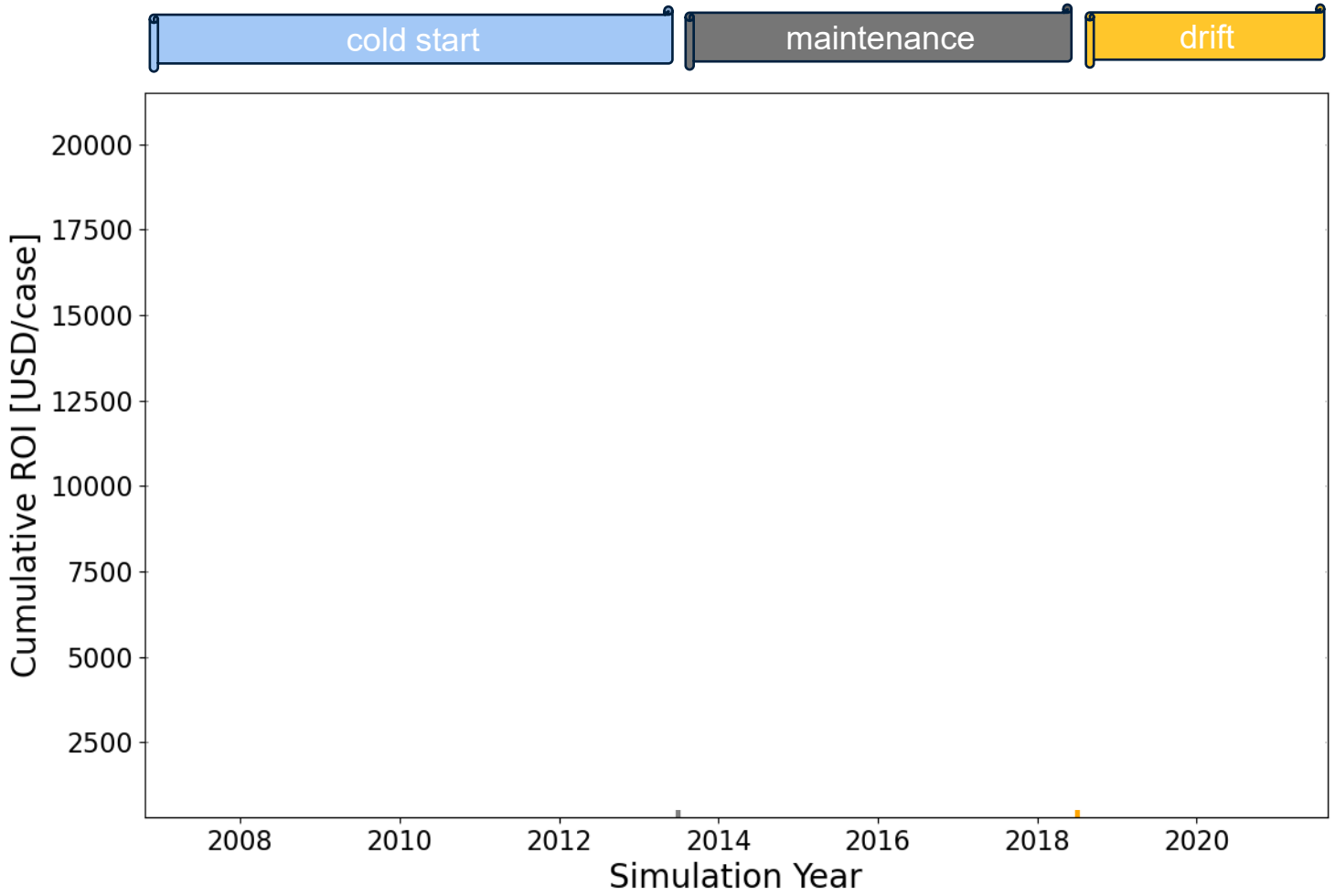
- Low data (drift+)
- Huge opportunity cost

Maintenance

- High data
- Small drift

Drift Event

- High data
- Large drift



Timeline Breakdown

Cold Start

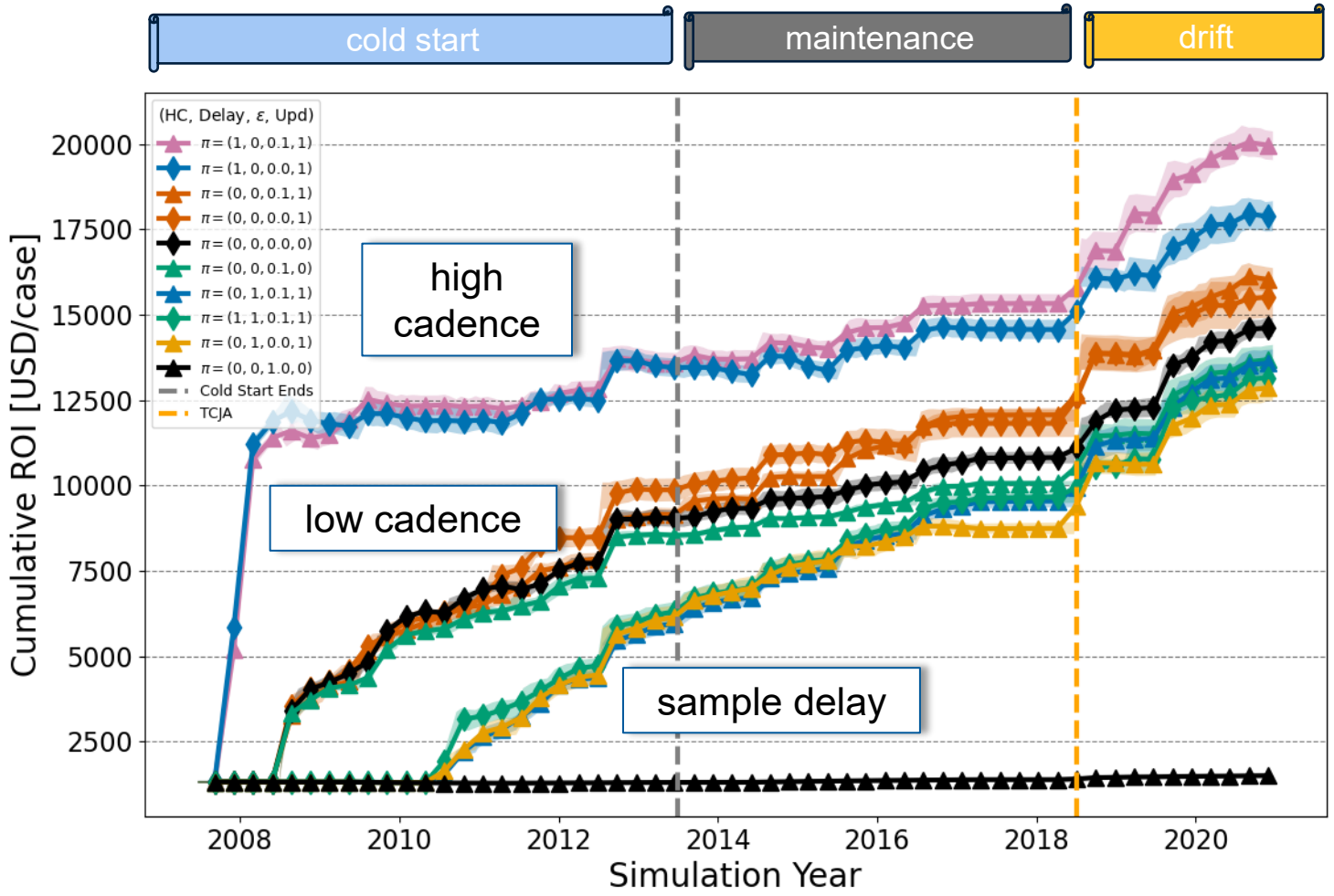
- Low data (drift+)
- Huge opportunity cost

Maintenance

- High data
- Small drift

Drift Event

- High data
- Large drift



Period-Specific Effects

Cold Start

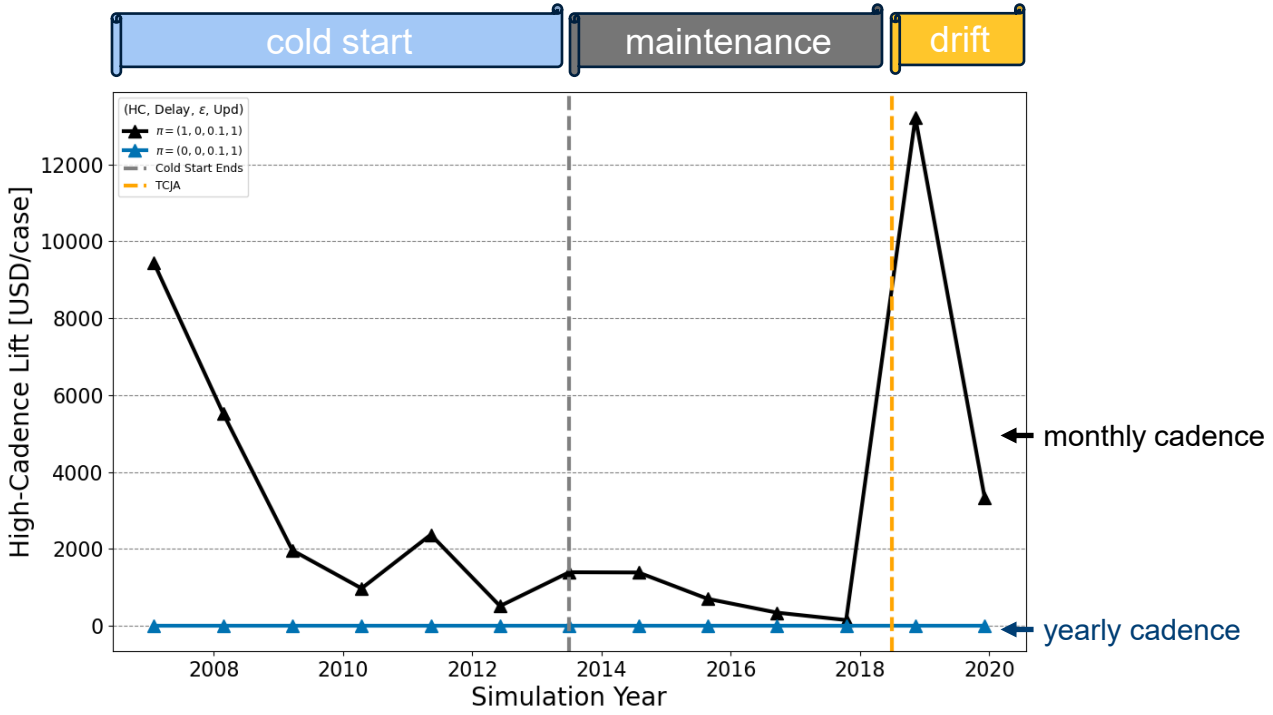
- Big lift from no delay and quick updates
- Epsilon almost irrelevant (untrained model is pseudo-random)

Maintenance

- Policy matters, though to a lesser degree.
- Epsilon more important.

Drift Event

- Policy is critical. All champion policy choices are relevant.



Policy	ROI_{case}			
	Overall	Cold Start	Maintenance	Drift Event
Policy	Δ	Δ	Δ	Δ
Sample Delay	-$6,360	-$7,270	-$1,590	-$6,640
Updating	$5,320	$4,550	$4,350	$7,420
High-Cadence	$3,940	$3,710	$1,020	$6,250
Epsilon	$2,060	$139	$1,950	$6,780

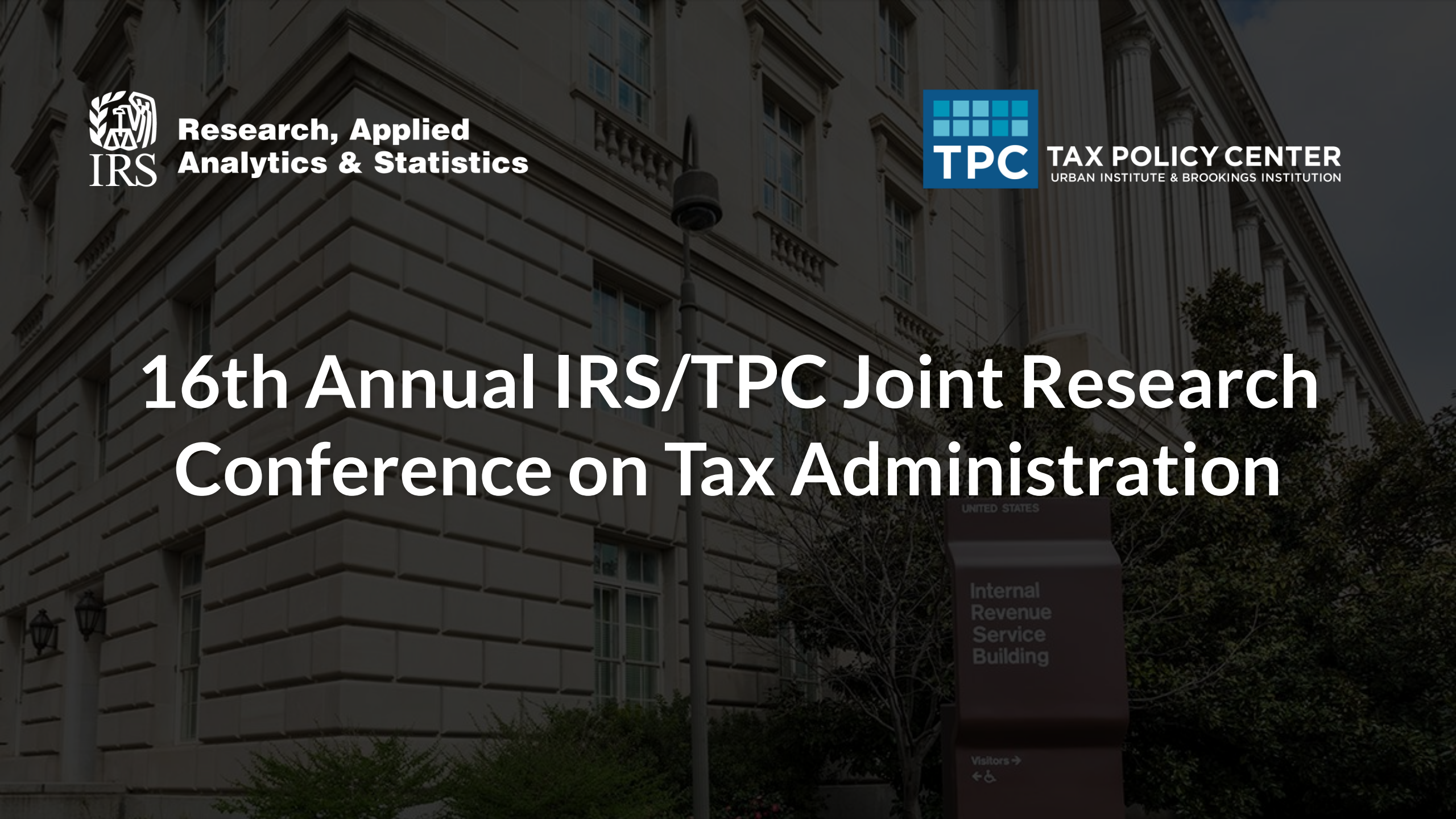
Conclusions

The Takeaway

- The individual tax filing environment is one where **frequent model updates provide a significant improvement** in risk identification, particularly following major policy changes.

Open Questions

- How does the picture change under a fixed workforce?
- What is the optimal update cadence?
- Can incomplete-sample updates feed an unbiased population estimation?
- Can drift detection-based explore give lift?



Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

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

Rethinking Post-Audit Compliance:

Heterogeneous and Persistent Effects in Experimental Data

Andrea Lopez-Luzuriaga · Gabriela Mejía · Carlos Scartascini

Inter-American Development Bank

IRS-TPC Research Conference

June 25, 2026 · Urban Institute, Washington, DC

Do Audits Change Behavior? A Contested Question

- Auditing is the primary enforcement tool of tax administrations
- Understanding post-compliance behavior matters greatly to plan any enforcement strategy
- Classic theory indicates that a higher likelihood of audits raises the expected cost of evasion, making compliance rise

(Allingham and Sandmo, 1972)

- Given that most individuals update their beliefs according to evidence, observing an audit would increase their expected audit probability
- Field evidence broadly agrees

(Kleven et al., 2011; Pomeranz, 2015; Advani et al., 2023)

- But lab experiments frequently find the opposite: compliance drops after an audit
  - Named the “bomb-crater effect” (Mittone, 2006; Kastlunger et al., 2009)
  - Attributed to gambler’s fallacy, loss repair, or morale erosion

What the Literature Has Found

Lab experiments:

- Fined audits raise compliance while audits that come out clean can lower it
  (Kasper and Alm, 2022b)
- Audit rates, fines, and tax rates all shape post-audit behavior
  (Kasper and Alm, 2022a; Alm and Malézieux, 2021)
- Exogenous audit rules decrease compliance while endogenous audit rules increase it
  (Kasper and Rablen, 2023)

Field studies:

- Audits have positive, durable compliance effects
  (Kleven et al., 2011; Pomeranz, 2015; Gemmell and Ratto, 2012)
- Third-party information and enforcement spillovers amplify positive impact
  (Pomeranz, 2015)
- Effects are heterogeneous by audit outcome and income source
  (Advani et al., 2023)

The lab–field gap persists even with perfectly random audits and fixed parameters.
We show the estimation strategy can explain the discrepancy.

What We Find

We use a lab experiment with **randomized i.i.d** audits.

1. Standard method replicates the bomb-crater: $\beta = -2.8$ pp ($p = 0.007$)

At first glance, our data also suggest audits backfire

2. Due to an estimation choice, rather than real behavior: the control group in the standard method is contaminated by earlier audits

3. Non-contaminated control group causal methods: audits **raise** compliance by +5 to +15 pp, and the gain **persists for at least 10 rounds**

4. Who responds: pre-audit **evaders** drive the effect (specific deterrence); with no causal bomb-crater effect for already-compliant taxpayers

The Experiment

40-round Tax Evasion Game with 410 college students in Bogotá, Colombia.

Each round a player:

- Earns income from a real-effort task
- Declares income (any amount $\in [0, z_t]$)
- Pays tax on declared amount
- Faces a 20% audit draw — i.i.d., announced, fixed
- If audited and under-declared: pays 4× the evaded tax

Why this setting is ideal:

- Audits are purely random each round — no targeting, no behavior dependence
- 40 rounds: sufficient variation to trace dynamic responses
- Parallel trends hold by design

Avg. payoff $12 USD ≈ 8.4× Colombia minimum hourly wage

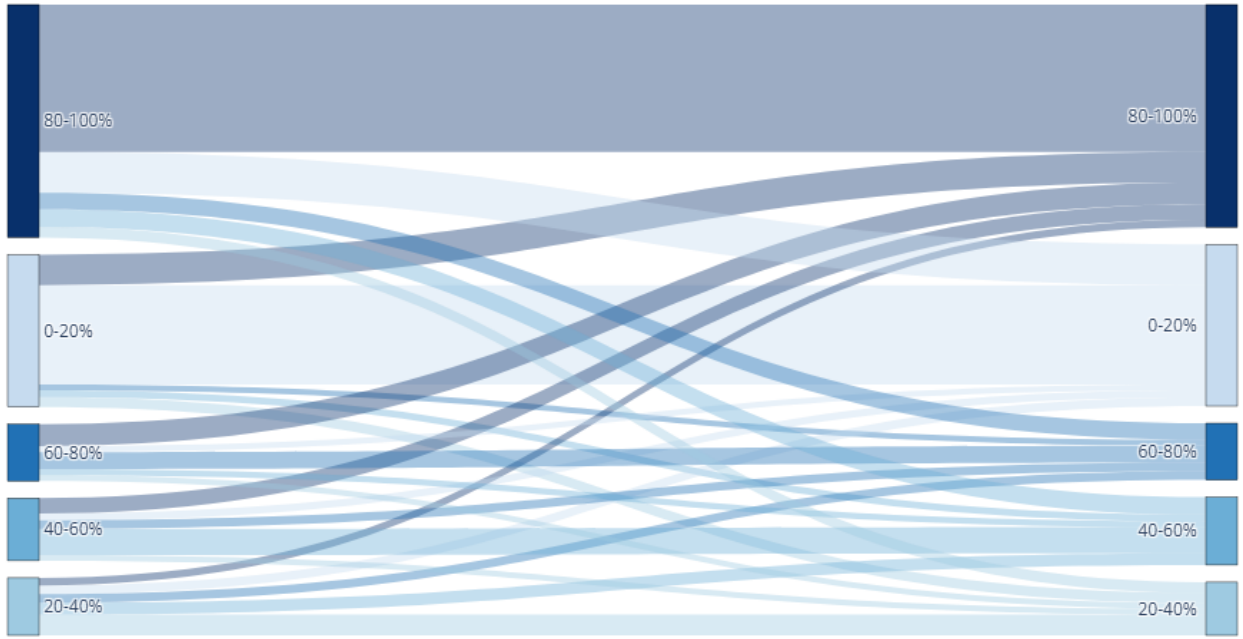
After an Audit, Behavior Moves in Every Direction

Compliance transitions from audit round (t) to next round (t + 1):

Some increase compliance
Some decrease
Some are unchanged

No dominant direction — foreshadows heterogeneous effects

Among evaders: 44.6% increase compliance
Among compliers: 47.9% *decrease*



Left: compliance quintile at audit round. Right: quintile in t + 1.

Who Adjusts After an Audit? Correlational Evidence

Linear probability model: what predicts the direction of compliance change in t + 1 after an audit at t?

Pre-audit evader status is by far the strongest predictor:

+44.6 pp probability compliance goes **up** ($p < 0.001$)
−21.4 pp probability compliance goes **down** ($p < 0.001$)

Higher risk appetite reduces the probability of compliance going up after an audit — consistent with risk theory

	Goes up	Goes down	No change
Evader at audit round (t)	0.446***	−0.214***	−0.232***
Risk measure (1–10)	−0.015**	0.006	0.009
Age	0.004	−0.008**	0.004
Observations	*3,266 audit events*		

*Significant rows only. SEs clustered at individual level. ***p<0.001, **p<0.01, *p<0.05*

Standard Method: Audits Appear to Backfire

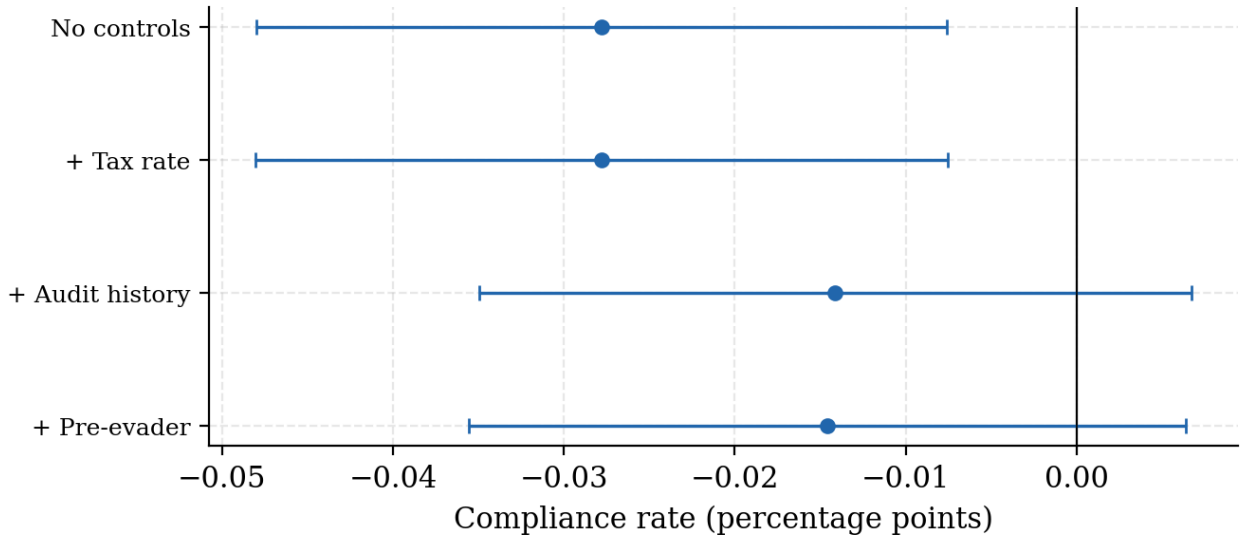
Compare rounds preceded by an audit vs. rounds not preceded by an audit, controlling for individual and round fixed effects.

Result: **β = −2.8 pp** (p = 0.007)

"Being audited last round is associated with lower compliance"

But: as we add controls for **cumulative audit history**, the estimate shrinks toward zero and becomes insignificant.

▶ Full regression table (appendix)



Coefficient on "audited last round" across specifications. Controls added sequentially. 95% CI.

Why the Standard Method Is Misleading

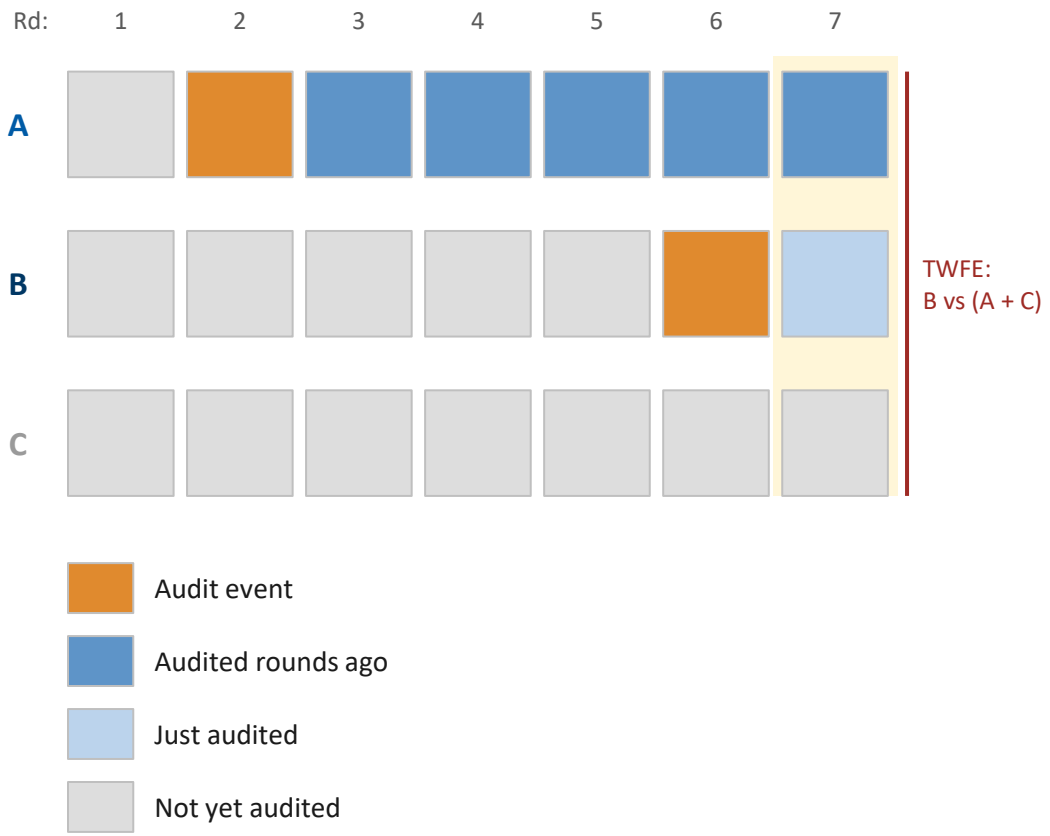
TWFE "control" = not audited *last* round

Problem: A player-round is only seen as treated in the round after the audit event. Later rounds count as "controls" while compliance may have risen from earlier audits (darker blue).

At round 7: B was just audited (round 6) ⇒ TWFE marks B as treated. A and C are control. But A (audited at round 2) is also in the control group with high compliance.

True counterfactual: **C** (never audited, gray)

Fix: use only not-yet-audited (**C**) as controls



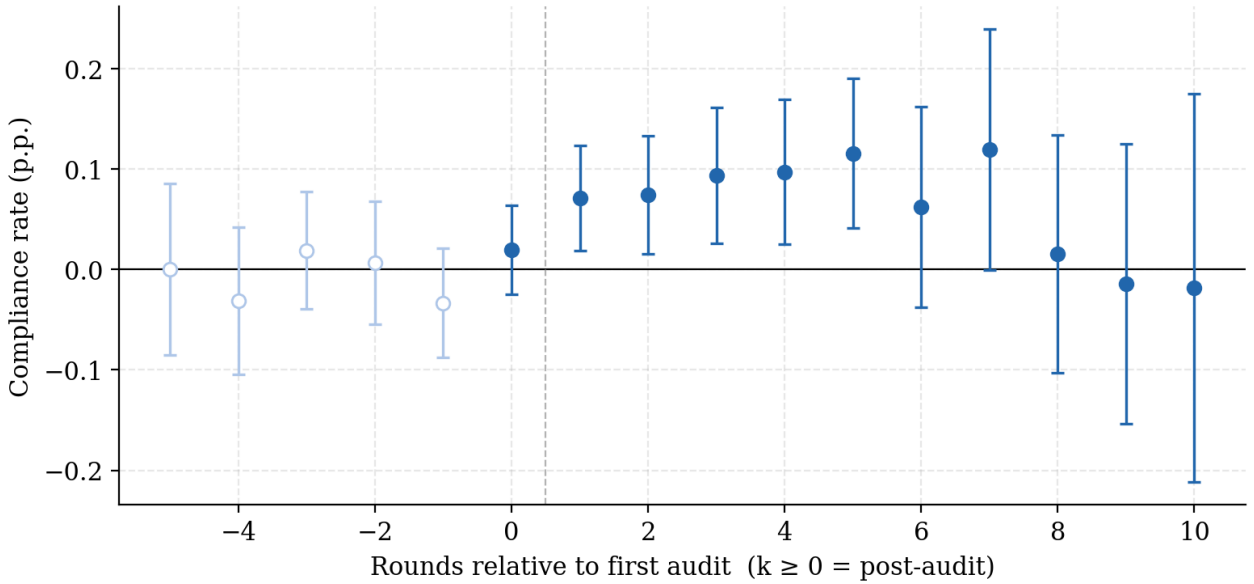
Under an Event-Study Approach: Audits Raise Compliance

Event-study estimator (Callaway and Sant'Anna, 2021):

- Treats the first audit as a treatment event; observations remain treated until their second audit
- Control: participants not yet audited
- Pre-event placebo checks pass ✓

Result: compliance rises by **+7 to +12 pp** in the five rounds after the first audit ($p < 0.05$, $k = 1–5$)

A 10 pp gain ≈ one-sixth of the average compliance rate



Event-study estimates. Control: not-yet-audited. 95% CI.

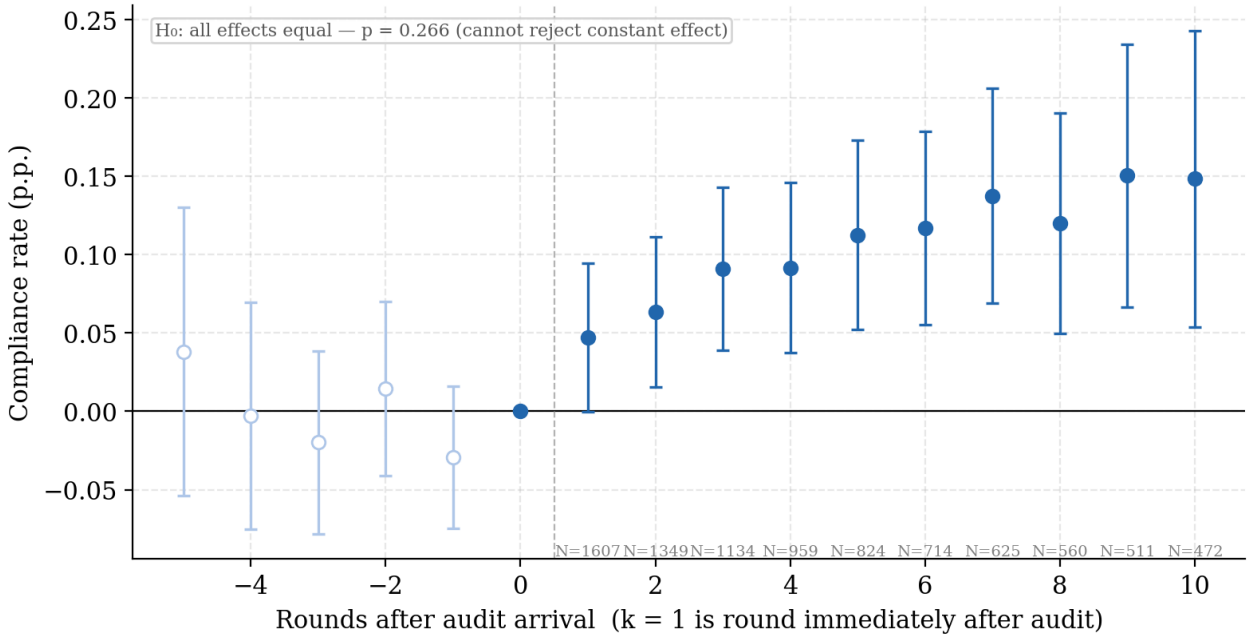
The Compliance Gain Is Durable

Intertemporal DiD (De Chaisemartin and d'Haultfœuille, 2024):

- Treats each audit arrival as a transient shock
- Traces the compliance gap over 10 rounds
- Uses all audit events (not just first)

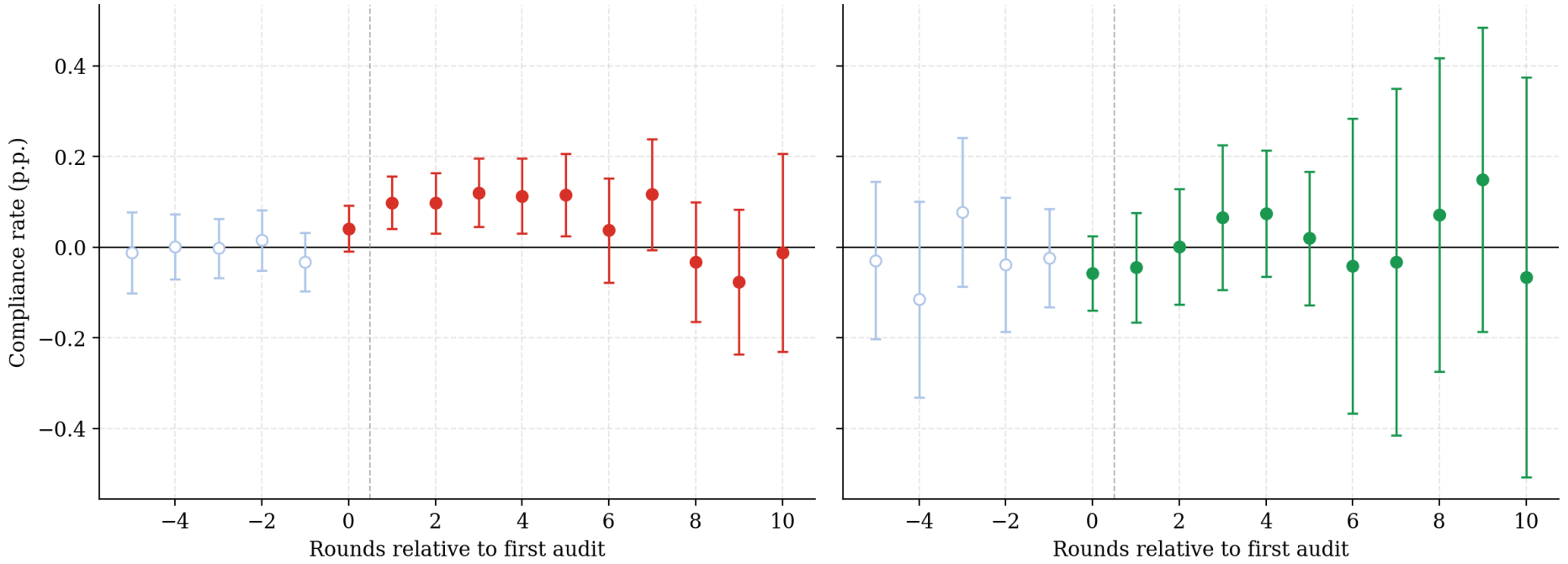
Result: **Effect does not decay.** Estimates grow from ≈ 5 to ≈ 15 pp.

Cannot reject equal effects across 10 horizons (p = 0.267) — a stable or growing gap, not a fading one.



Impulse response to audit arrival.
On-switchers vs. not-yet-switched. 95% CI.

Who Responds? Evaders Deter; Compliers Are Unaffected



Pre-audit evaders (≥ 50% evasion pre-audit):

- **+10 to +12 pp** ($p < 0.05$, $k = 1–5$)

Fines trigger meaningful compliance updating

Pre-audit compliers (already largely compliant):

- All estimates not statistically different from **zero**

Summary: What Different Methods Tell Us

Method	Avg. effect	Key finding
TWFE (no history)	**−2.8 pp (p = 0.007)**	Bomb-crater pattern
TWFE (with audit history)	**−1.4 pp (p = 0.18)**	Shrinks to zero once history controlled
Callaway–Sant'Anna	**+7 to +12 pp**	Significant k = 1–5; correct control group
De Chaisemartin–D'Haultfœuille	**+5 to +15 pp**	Non-decaying gap over 10 rounds
CS: Pre-audit evaders	**+10 to +12 pp**	Specific deterrence
CS: Pre-audit compliers	**Mixed, all zero**	No causal bomb-crater established

Policy Implications

1. Audits work. The bomb-crater effect is a measurement artifact. Standard TWFE is biased when audit effects persist beyond one period.

2. The behavioral return is durable. A single audit generates compliance gains that persist for at least 10 rounds. Evaluations capturing only the immediate post-audit period understate the full value of enforcement.

3. Target evaders. Compliance gains are concentrated among non-compliant taxpayers. Auditing already-compliant taxpayers yields no measurable causal gain.

4. Method matters even under random assignment. Dynamic effects contaminate the TWFE control group. Event-study methods with not-yet-audited controls recover the correct sign and magnitude.

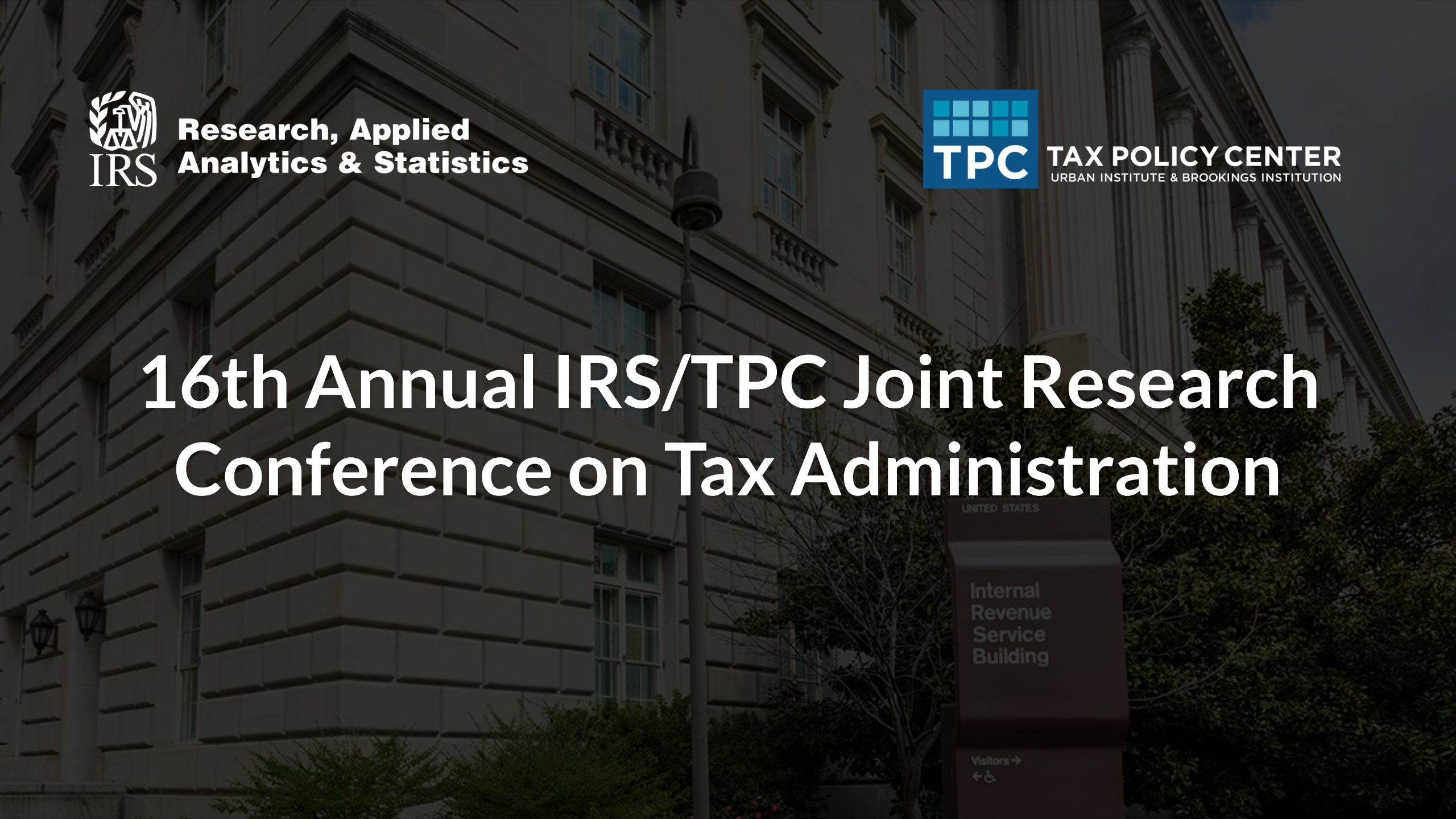
Conclusion

- We exploit a perfectly randomized audit design — no selection, no targeting
- Standard TWFE gives −2.8 pp (bomb-crater). It disappears once audit history is controlled: the control group was contaminated by earlier audits
- Causal estimators recover a positive, persistent effect: +5 to +15 pp over a 10-round window
- Effects are driven entirely by pre-audit evaders; no causal bomb-crater even for the most plausible candidate subgroup
- **Bottom line: audits produce durable behavioral improvements. A single well-timed audit changes behavior for far longer than previously thought.**

Thank you!

carlossc@iadb.org · andrealo@iadb.org · gabrielamej@iadb.org

www.cscartascini.org



Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

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

Discussant Comments: Getting the Most out of Audits

Brian Erard

LOO Estimation: Core Contributions

- **Important Issue:** Undetected noncompliance on NRP audits accounts for more than half of the overall tax gap.
- **Nonparametric Approach:** Introduces a data-driven leave-one-auditor-out jackknife pipeline that wraps around any predictive surface, bypassing rigid parametric constraints.
- **Advances Beyond Tobit:** Proposes more flexible two-stage frameworks that separately model detection and amount, though these extensions remain restricted to the appendix.
- **Actionable Multiplier Framework:** Provides a practical algorithm for discovering elite examiner reference groups and generating empirical multipliers to scale observed adjustments into detection-adjusted tax-gap estimates.

Methodological Limitations: Part 1

- **Conflating Evasion and Detection (the Case-Steering Problem)**: Lacks structural parameters to separate taxpayer compliance from auditor capacity. Because the predictive model cannot untangle the presence of noncompliance from the skill to detect it, the estimator is entirely defenseless against institutional steering of more complex and productive cases to selected examiners.
- **Baseline Contamination (False Zeros)**: Because less-skilled examiners generally leave unclassified items untouched, the baseline is flooded with undetected cheating masked as "compliance". This leaves a mass of noncompliance completely unaccounted for, heavily biasing the final tax-gap multipliers
- **Failing to Disaggregate:** Difficult to scale methodology to permit tax gap estimation by income source due to severe data sparsity in small individual caseloads.

Methodological Limitations: Part 2

- **Signal vs. Noise:** NRP application in Table 5 with *k = 1* yields a 20% top examiner pool based on an average caseload of just 22 cases, meaning a single fortunate case can elevate an examiner to the top tier group.
- **Coronation of Posers:** The "elite" pool risks being systematically populated by aggressive auditors who take unsupportable positions that produce overly large recommended tax adjustments
- **Advanced Models Never Touch Real Data (YET):** The Tobit and two-stage models developed specifically to handle censoring are confined to the appendix. All actual NRP data results in Tables 5 and 6 rely on an oversimplified linear model.

Suggestions

- **Incorporate Generalized Propensity Weighting:** Introduce a high-dimensional propensity score (e.g., multinomial logit) step with inverse probability weighting across case characteristics to control for case-steering and ensure examiners are evaluated on a balanced case mix.
- **Replace Simulation Tuning with Empirical Validation**: Reject tuning *p* and *k* via artificial simulations. Instead, test the thresholds directly on real data by injecting known, artificial auditor effects into a test sample to evaluate exactly how accurately the algorithm flags them.
- **Account for Employee Classification Tiers**: Separately control for examiner types (TCOs, TAs, and RAs) within the estimator. Because these statuses govern the baseline complexity of the assigned caseload, failing to segment them causes the algorithm to confound a high-level job title with an individual's exceptional detection skill.

Refresh Matters: Key Contributions

- **Important Issue:** Systematic attempt to optimize audit selection model retraining frequency
- **Large Revenue Implications:** Finds large benefit of more frequent refreshes: 26% ROI from monthly updating paired with 10% Epsilon-Greedy exploration rule relative to baseline approach
- **Sample Delay Can Be Costly:** Demonstrates that a static one-shot model actually outperforms an update schedule that waits 3 years for complete data, confirming the severe revenue costs of waiting for a complete audit sample under concept drift

Methodological Limitations

- **Overly General Model: E**valuates a single feature vector across the entire return population, omitting exam class-stratified IRS selection models
  - Stratified selection helps ensure appropriate coverage across taxpayer types and improves selection.
- **Pooled Audit Types:** Conflates field examinations and correspondence audits despite distinct ROI profiles, case durations, and response rates
- **Data Sparsity Concerns:** Annual NRP samples plummeted from 12,000 to 4,000 cases, increasing small-sample variance and early-closure fixation risk
- **Unmodeled Operational Constraints**: Assumes a perfectly fungible workforce across auditor types (TCO, TA, RA) and omits the immediate revenue cost of a 10% random audit diversion

Suggestions

- **Downsample Earlier Years:** Uniformly downsample pre-2013 annual datasets to the 4,000-case post-2018 density to isolate data sparsity while preserving the full timeline
- **Evaluate Case-Mix Impacts:** Examine how high-cadence selection shifts the demographic, income, or geographic profiles of audited taxpayer
- **Disaggregate Selection Models:** Test how optimal retraining cadences vary when models are tailored strictly to independent examination classes

Rethinking Post-Audit Compliance: Context

- Many lab experiments (and a growing number of field studies) have examined how subjects change their reporting behavior after experiencing an audit.
- The results typically find that audits serve as a specific deterrent to noncompliance, although estimates of the magnitude of the impact and its trajectory over time vary across studies.

Bomb Crater Hype:

- A relatively small number of lab experiments have found that audits have a counter-deterrent effect on future compliance behavior.
- Main explanations offered for such an effect:
  - Bomb Crater Hypothesis: Misperception of future audit risk.
  - Loss Repair: Think of a gambler who doubles down after losing big

Key Contribution

- **The Invalidation of Standard TWFE:** The authors demonstrate that the classic "bomb-crater" effect is largely a statistical illusion.
  - Standard regression models (TWFE) fail when an audit has a lingering, multi-period impact on behavior because the "unaudited" periods become heavily contaminated by multiple past treatments.
  - The authors show that even a crude control for a participant's prior audit history causes this negative effect to disappear.

Limitation #1: No Incentive to Cheat

- With a fixed audit probability of 20% and a penalty of 4 times the amount evaded, there is no incentive to cheat!
  - A risk-neutral income-maximizer would be completely indifferent about how much tax to report
  - A risk-averse utility maximizer would choose fully to comply

Limitation #2: Lucky Survivor Effect

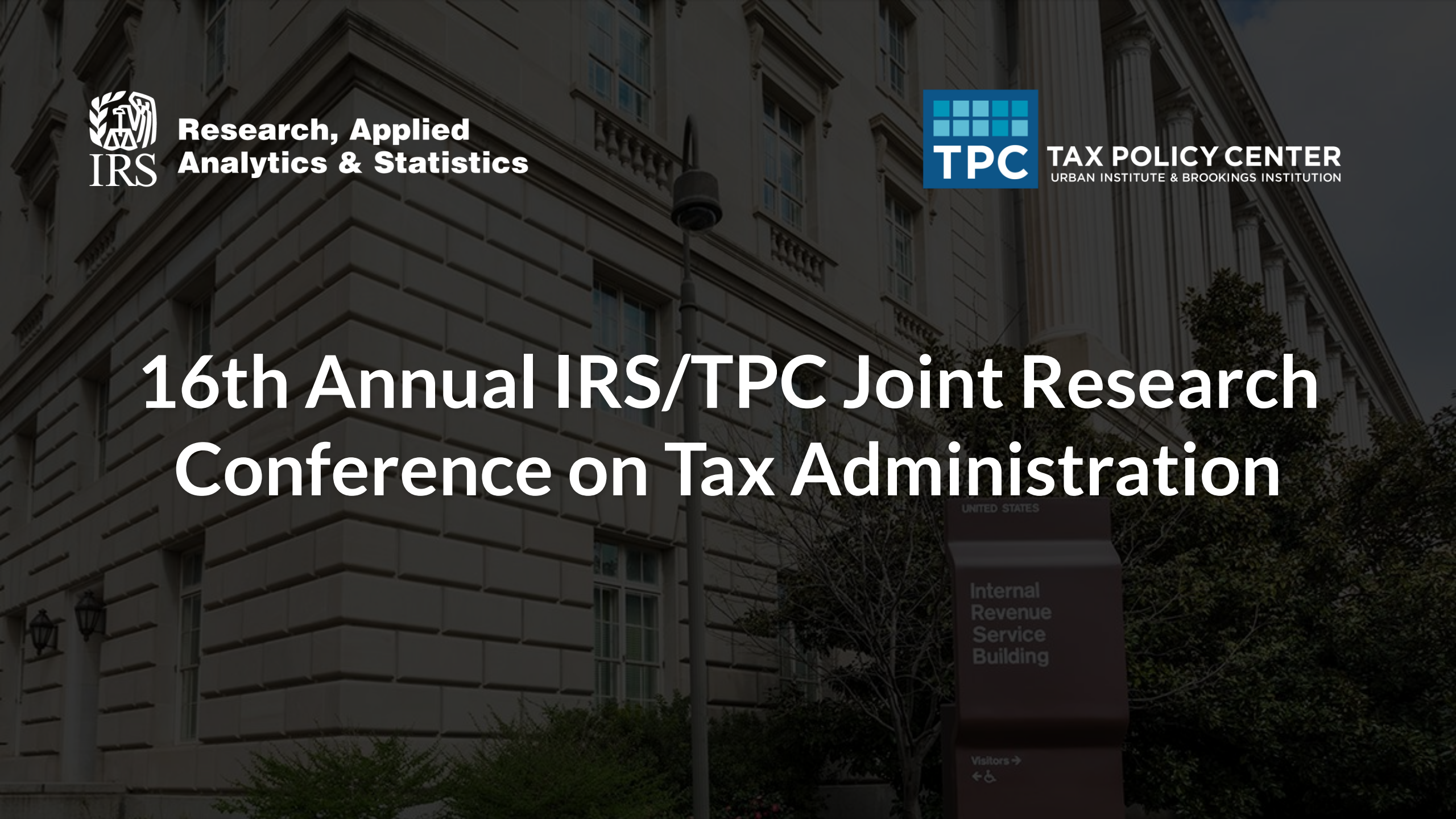
- The CSDID method does technically estimate a valid Average Treatment Effect (ATE) of a first-audit; however, it does so by ruthlessly censoring the data
  - Participants who were first audited in round *t* as well as those who had not yet experienced an audit are dropped from the analysis as soon as they experience an audit in a subsequent round.
  - Therefore, a new Local ATE is computed in each subsequent round restricted to the ever-shrinking pool of lucky survivors who have not yet been eliminated from the sample.
  - By round 10, 90% of the sample has been tossed out, leaving an unrepresentative sliver of "hyper-lucky" participants whose risk psychology no longer mirrors the general population.

Limitation #3: DcDH Woes

- The authors incorrectly claim that their advanced De Chaisemartin & d'Haultfœuille (DcDH) model compares groups facing the same future audit risk.
  - The software mechanically forces the control group to be "stayers"—meaning they must not have experienced any audits up through the specified round. As soon as a member of the control group gets audited, that member is eliminated from the group for all subsequent rounds.
  - Meanwhile, the treated group accumulates random re-audits at a 20% rate. The estimated 15-percentage-point compliance gap is an artifact of this massive treatment-intensity imbalance, not a behavioral response to the lingering memory of a single first-time audit pulse.

Proposed Alternative: ITT Event Study

- **Preserves Natural Randomization:** Avoids the ruthless data censoring of CSDID and the rolling 0% audit-free filter of DcDH by conditioning only on a participant's baseline history up to period t.
- **Static Treatment Assignment:** Restricts the sample to units with zero prior audits up to period t-1, defining a time-invariant indicator $\boldsymbol{FirstAudit_t} = 1$ if audited at period t, and 0 if not.
- **Uncensored Longitudinal Tracking:** Tracks both groups forward through period t+k without dropping units when they receive subsequent audits.
- **Guarantees Symmetric Future Risk:** Because the 20% random audit lottery resets every round, both groups naturally accumulate future audits at the exact same expected rate.



Research, Applied Analytics & Statistics

TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

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