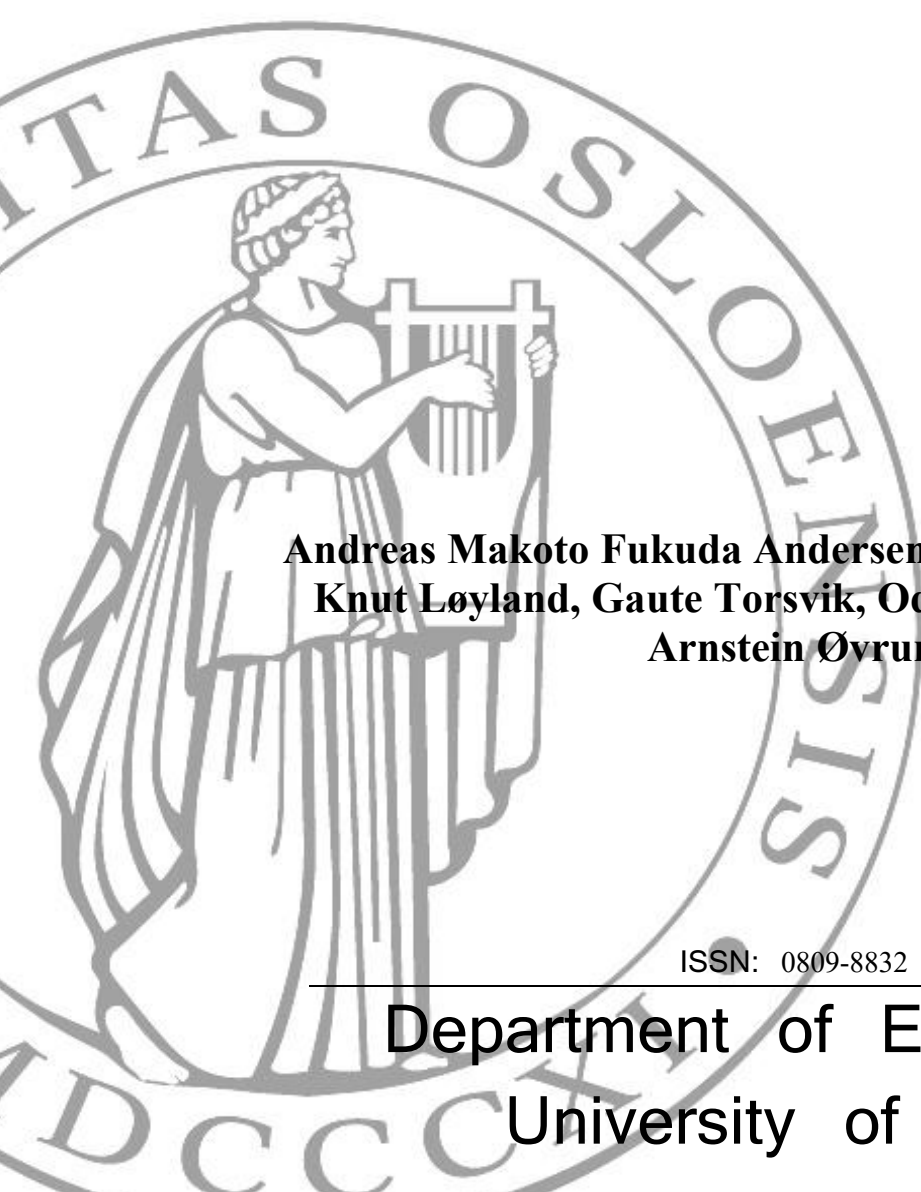


MEMORANDUM

No 02/2026

May 2026

The Limits of Low-Intensity VAT Enforcement: Experimental Evidence from Norway



**Andreas Makoto Fukuda Andersen, Andreas Kotsadam,
Knut Løyland, Gaute Torsvik, Oddbjørn Raaum and
Arnstein Øvrum**

ISSN: 0809-8832

**Department of Economics
University of Oslo**

This series is published by the
University of Oslo
Department of Economics

P. O.Box 1095 Blindern
N-0317 OSLO Norway
Telephone: + 47 22855127
Fax: + 47 22855035
Internet: <http://www.sv.uio.no/econ>
e-mail: post@econ.uio.no

In co-operation with
**The Frisch Centre for Economic
Research**

Gaustadalleén 21
N-0371 OSLO Norway
Telephone: +47 22 95 88 20
Fax: +47 22 95 88 25
Internet: <http://www.frisch.uio.no>
e-mail: frisch@frisch.uio.no

Last 10 Memoranda

No 01/2026	Ragnhild Schreiner, Gaute Torsvik and Gjermund Sterri Yndesdal: Who Bears the Burden of Payroll Taxes at the Top?
No 03/2025	Edwin Leuven and Ingrid Mikkelsen Semb: Entrepreneurial returns to college
No 02/2025	Zachary Grossman, Tony Hua, Jo Thori Lind and Karine Nyborg Unwillingly Informed: the Prosocial Impact of Third-Party Informers
No 01/2025	The mixed effects of making contribution behaviors observable Noel Z. Hoven og Karine Nyborg
No 03/2024	Effects of Compressed Work Schedules on Sickness Absence Elisabeth Fevang, Andreas Fidjeland, Karen Hauge, and Otto Lillebø
No 02/2024	Tore Adam Reiremo The Anatomy of Mismatch in the College Market
No 01/2024	Gunnar Bårdsen and Ragnar Nymoen U.S. wage-price dynamics, before, during and after COVID-19, through the lens of an empirical econometric model
No 04/2023	Andreas Bjerre-Nielsen, Lykke Sterll Christensen, Mikkel Høst Gandil and Hans Henrik Sievertsen Playing the system: address manipulation and access to schools
No 03/2023	Gunnar Bårdsen and Ragnar Nymoen Dynamic time series modelling and forecasting of COVID-19 in Norway
No 02/2023	Hans A. Holter, Serhiy Stepanchuk and Yicheng Wang Taxation and Entrepreneurship in the United States

Previous issues of the memo-series are available in a PDF® format at:
<https://www.sv.uio.no/econ/english/research/publications/memoranda/>

The Limits of Low-Intensity VAT Enforcement: Experimental Evidence from Norway*

Andreas Makoto Fukuda Andersen¹, Andreas Kotsadam², Knut Løyland³,
Gaute Torsvik¹, Oddbjørn Raaum², and Arnstein Øvrum³

¹University of Oslo

²Ragnar Frisch Centre for Economic Research

³Norwegian Tax Administration

May 12, 2026

Abstract

We study the effects of low-intensity tax enforcement on VAT compliance among newly registered firms in Norway using a large-scale randomized field experiment. Around 20,000 firms were randomly assigned to desk-based audits, informational letters, phone-based guidance, or a control group. Desk-based audits uncover non-compliance in over nine percent of cases, corresponding to an estimated VAT gap of about 6.3 percent. However, results show no evidence of persistent increases in reported net VAT following any of the three interventions, and the estimates are sufficiently precise to rule out even small compliance effects. The results suggest that low-intensity VAT enforcement, in isolation, is insufficient to generate behavioral change in the firms' average VAT reports.

*This study was reviewed and approved by the Norwegian Tax Administration. Filed at the AEA RCT Registry (AEARCTR-0008688). Authorship is listed alphabetically. We thank participants at various seminars, workshops and conferences for valuable comments. Funding from the Research Council of Norway (project No. 341250) is gratefully acknowledged.

1 Introduction

Value-added tax (VAT) is a central source of government revenue in most OECD countries and accounts for nearly 20 percent of total tax revenue in Norway ([The Norwegian Tax Administration, 2024](#)). It is also an area of taxation where compliance challenges frequently arise. VAT rules are often complex, particularly for newly established firms. Combined with reliance on self-reporting and limited invoice-level third-party reporting, this creates scope for both unintentional errors and deliberate evasion. These features make effective enforcement essential for safeguarding the VAT base.

Tax administrations, therefore, employ a wide range of enforcement strategies to improve compliance. Some interventions are soft, such as informational letters, reminders, or telephone guidance, and aim to reduce errors by improving taxpayers' understanding. Others are more forceful, such as desk-based audits or, at the more intensive end, full audits, which seek to detect, correct, and deter underreporting.

Beyond their immediate effects, both soft and hard interventions can shape subsequent compliance through two main channels. First, they may induce deterrence by increasing the perceived likelihood of future detection and sanctions, thereby raising the expected cost of evasion. Second, they can provide informational benefits that help honest taxpayers interpret and apply the rules correctly.

In this paper, we examine how three enforcement interventions (desk audits, phone-based guidance, and informational letters) affect subsequent VAT reports. Desk audits account for the bulk of VAT enforcement activity at the Norwegian Tax Administration (NTA), a point we return to in [Section 2.1](#). The interventions targeted newly registered VAT-liable firms in Norway. We implemented a large-scale field experiment in which around 20,000 eligible firms were randomly assigned to one of the three treatments or to a control group. This random assignment eliminates the selection bias inherent in risk-based enforcement and enables us to identify both the immediate fiscal effects and the longer-term behavioral responses in VAT reporting. Since the enforcement strategies are used for the same sample of firms, the effects are directly comparable.

Results show that desk-based audits uncover initial non-compliance in just over 9 percent of the cases examined. On average, however, the audited firms did not report net VAT differently compared to the control firms over the next year. The phone-based guidance and informational letters had no significant impact on VAT reporting, and we can reject fairly small effects. Using two one-sided equivalence tests at the 5 percent level, we can rule out audit effects larger than about NOK 3,000 per firm per bi-monthly period, or less than 7 percent of the control-group mean.

Our study contributes to a growing body of experimental work on tax enforcement (reviewed in [Slemrod, 2019](#)). While audits and informational nudges have been shown to raise compliance among personal income taxpayers (e.g., [Hebous et al., 2023](#); [Løyland et al., 2024](#); [Gemmell and Ratto, 2012](#); [Advani et al., 2023](#); [DeBacker et al., 2018](#)), the evidence on firms, and on VAT specifically, is thinner and more mixed. [Pomeranz \(2015\)](#) shows that third-party paper trails are central to VAT compliance in Chile; [Carrillo et al. \(2017\)](#) document limits to VAT enforcement via third-party cross-checks in Ecuador; [Naritomi \(2019\)](#) uses consumer incentives to uncover firm misreporting in Brazil. [Best et al. \(2021\)](#) document substantial detected evasion among Pakistani VAT firms but little subsequent deterrence. [Bjørneby et al. \(2021\)](#) find that unannounced on-site payroll audits in Norway raise reported wages by about 18 percent, which suggests that in this institutional setting, the audit tool can shift reporting when it is intensive enough. Evidence on VAT-firm behavior under the less intensive audit tools that tax administrations deploy at scale is sparse.

We contribute in two ways. First, we randomize three active enforcement modes within the same firm population: a desk-based audit, an informational letter, and phone-based guidance. [Boning et al. \(2020\)](#) compare letters and in-person visits at the IRS but include no audit modality beyond the visit; [Ortega and Scartascini \(2020\)](#) compare letters, email, and in-person visits among individual delinquents in Colombia; [Almunia et al. \(2025\)](#) run a three-arm letter design at the VAT-network level in Uganda but include no audit arm. To our knowledge, no prior firm-level VAT experiment

combines an audit arm with two softer interventions within a single randomization. Second, we restrict the sample to firms that have recently registered for VAT. [Gangl et al. \(2014\)](#) is the only prior experiment we found that restricts itself to newly founded firms, and it studies Austrian tax payments broadly rather than VAT; [Brockmeyer et al. \(2019\)](#) is related but targets already-registered non-filers rather than new registrants.

Our audit arm (desk-based correspondence audit) is the workhorse verification tool of modern VAT administration: desk-based correspondence audits of single VAT returns, widely used for pre-refund verification of credit claims, are a standard instrument across OECD jurisdictions ([Harrison and Krelove, 2005](#); [OECD, 2025](#)), and credit-return verification has been described as the principal enforcement margin in the VAT system ([Brockmeyer et al., 2024](#)). The randomized experimental literature on VAT enforcement, however, has focused on adjacent instruments. Our experiment fills that gap.

The desk-based correspondence audit is excluded from the welfare analysis of IRS audits in [Boning et al. \(2025\)](#) on the grounds that it has a fundamentally different cost structure; [Li and Wang \(2026\)](#) provide quasi-experimental evidence on this modality using a Chinese VAT audit threshold. The absence of behavioral effects is consistent with [Best et al. \(2021\)](#), who find null deterrence effects of comprehensive on-site audits among Pakistani VAT firms; our result extends that finding to a high-capacity Nordic tax environment and to newly registered firms. Our finding fits a now-three-country picture for VAT desk audits: [Best et al. \(2021\)](#) find null deterrence in Pakistan, where weak state capacity limits collection; [Li and Wang \(2026\)](#) find a net negative effect in China, where strong state capacity is offset by aggressive avoidance among large firms; we find a precise null in Norway, where strong state capacity meets a population of newly registered firms with limited scope for the kind of avoidance substitution Li and Wang document.

The design was preregistered at the AEA RCT Registry before any outcome data was accessed (AEARCTR-0008688).

The remainder of the paper is structured as follows. Section 2 describes the institutional setting, data, and interventions. Section 3 presents the empirical strategy. Section 4 reports the results, and Section 5 concludes.

2 Institutions, Data and Treatments

Each bi-monthly VAT return covers a two-month period, yielding six *reporting periods* per year. Each return contains aggregated transaction data for the period, reported across several transaction categories. Norway applies five VAT rates (25 percent, 15 percent, 12 percent, 11.11 percent, and 0 percent) and also allows for transactions outside the scope of the VAT Act that carry no VAT liability. VAT-registered entities must report both input VAT (i.e., purchases) and output VAT (i.e., sales) for each category. Net VAT liability is calculated as the difference between total reported output and input VAT. Because transactions are not subject to third-party reporting, all information in these returns is self-reported.

Returns are submitted with a delay: The deadline for filing is the 10th day of the second month after the end of the reporting period. For example, reporting period 1 covers all transactions made between 1 January and the end of February, with a submission deadline of 10 April. The only exception is the deadline for reporting period 3 (covering May and June) which is set at 31 August. This will have implications later on when we discuss treatment timing.

2.1 Desk Audits and the Additional Tax Regime

Norwegian VAT enforcement distinguishes between two main audit modalities. A desk audit is a correspondence audit of a single bi-monthly VAT return: the NTA notifies the firm that a specific return has been selected, reviews the reported balances, and, when errors are suspected, requests and inspects supporting documentation. The audit is confined to one reporting period and is conducted from the tax office without an on-

site visit. This contrasts with an extended audit, which can span multiple periods and several tax types, may involve on-site inspection, and typically takes several months to complete ([Auditor General of Norway, 2025](#)).

Desk audits are the workhorse of Norwegian VAT enforcement, but the volume has fallen sharply over the past decade, from 4.5 percent of registered VAT entities in 2013 to 1.6 percent in 2024 ([Auditor General of Norway, 2025](#)). In routine practice, desk audits are typically risk-based and target credit returns; in 2024 the hit rate was 51 percent for risk-scored cases and 75 percent for cases opened manually by caseworkers. These population-level rates are well above the 9.3 percent detection rate we observe in our randomly assigned sample of newly registered firms.

When a desk audit establishes that a firm has provided incorrect or incomplete information that could have produced a tax advantage, the NTA shall impose additional tax under the Tax Administration Act. The standard rate is 20 percent of the established advantage and is, by statute, mandatory whenever the conditions are met. In practice, however, [Auditor General of Norway \(2025\)](#) reviewed a sample of desk audits with upward VAT adjustments and concluded that additional tax should have been imposed in 20 of 25 cases under the statute but was not. The sanction component of a desk audit is therefore, on average, weaker than the statute prescribes, a feature we return to when interpreting the null in Section 4.4.

The timing of the two NTA actions also differs. The firm receives notice of the audit one bi-monthly period after selection, while the first-instance assessment with any additional tax typically follows two to three periods after that. Any deterrence channel running through the sanction therefore cannot operate at $t + 1$, and only partly at $t + 2$.

2.2 Data

The target population consists of 50,495 businesses (either limited liability companies or sole proprietorships) with bi-monthly VAT returns and newly registered for VAT

between 1 May 2019 and 30 April 2021. We focus on this group for three reasons. First, around 75.5 percent of all VAT-registered entities in Norway have this combination of organizational form and filing frequency, ensuring broad coverage of the general VAT-paying population. Second, newly registered companies are assumed to face a higher risk of non-compliance due to unfamiliarity with VAT rules, which may increase the effectiveness of enforcement interventions. Third, the homogeneous filing frequency across these entities allows us to construct a panel dataset without requiring temporal aggregation.

Our dataset is initially structured as an unbalanced panel¹ containing all VAT returns from reporting periods 3/2020 through 6/2021. We also have access to a rich set of entity-level data, including information on industry, number of employees, and whether the firm engages a certified auditor.

2.3 Treatments and Sampling

Firms in the target population were randomly assigned to one of four conditions: a no-intervention control, an informational letter, phone-based guidance, or a desk-based correspondence audit. The letter arm informed the firm of its VAT reporting obligations and included a phone number to a dedicated NTA caseworker who could be contacted for further information. The phone arm consisted of an outgoing call from an NTA caseworker in which the firm was offered guidance and given the opportunity to ask questions. The audit arm corresponds to the desk audit modality described in Section 2.1, applied to reporting period 4/2021.

The design was stratified by firm age at t_0 : 1–12 months since VAT registration (stratum 1) and 13–24 months (stratum 2). The letter and telephone arms were drawn from both strata; the audit arm was drawn exclusively from stratum 2, because operational constraints limited the audit rollout to firms registered at least 12 months before t_0 . In stratum 2, the assignment produced 1,647 letter firms, 813 telephone firms, 1,069

¹This is mainly due to companies being liquidated during the experiment period.

audit firms, and 18,366 control firms, summing to the 21,895 estimation-sample total.

Because the audit treatment exists only in stratum 2, our main estimation sample is this stratum, comprising 21,895 firms registered between 1 May 2019 and 30 April 2020. All three treatment arms are represented in this sample and all treatment-effect estimates in Section 4 are reported within it. Table A1 in the appendix shows that pre-treatment firm characteristics are balanced across arms within this estimation sample, as expected under random assignment.

When determining treatment timing, we identify the last reporting period whose deadline falls before any treatment is announced or implemented to exclude any anticipatory effects.² Defining treatment timing is particularly challenging due to the time required to complete all planned audits. Since all audit results and adjustments are included in the VAT reports of reporting period 4/2021, we set t_0 , the last pre-treatment period, to be reporting period 3/2021. Since the audits were not completed until later periods, post-treatment effects measured from t_0 may be diluted for firms audited late, a limitation we return to in Section 4.4.

²Although entities are technically allowed to amend returns filed up to three years prior, we fix all pre-treatment periods as of 20 August 2021 to avoid contamination.

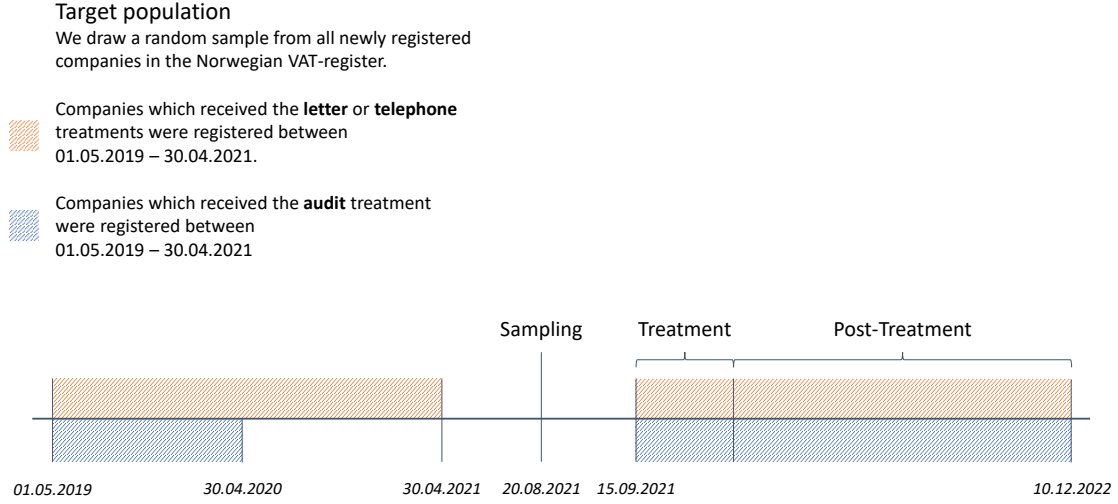


Figure 1: Timeline of sampling and treatment

Notes: The figure shows the experimental timeline from sampling through post-treatment reporting. Bi-monthly VAT reporting periods are indicated along the horizontal axis; t_0 denotes the last pre-treatment period (reporting period 3/2021).

Letter and telephone interventions were delivered between 15 September and 10 October 2021; audits of reporting period 4/2021 were conducted from October 2021 and completed in later periods. Source: Authors' rendering based on Norwegian Tax Administration operational records.

2.4 Outcomes and Sample Trimming

Since our goal is to investigate and compare changes in non-compliance across treatment groups over time, we use *net VAT* as the primary outcome of interest. This captures compliance effects on both input and output VAT and has clear implications for government revenue.

In the post-treatment periods, some outcome observations will be missing. This can occur because entities are closed, temporarily fail to report, or are removed from the VAT register by the NTA as a result of the audit. Since these missing observations are likely related to the treatment, excluding them would introduce selection bias. Therefore, missing post-treatment outcomes are coded as zero and included in the analyses, converting our unbalanced panel into a *balanced* panel. All period outcome

variables are winsorized at p_1 and p_{99} .

3 Empirical Strategy

For each post-treatment period t , we estimate the average treatment effect (ATE) of arm j by an OLS model of the following specification:

$$Y_{i,t} = \alpha_t + \beta_j \text{Treatment}_{i,j} + \gamma' \mathbf{X}_{i,t_0-2} + \varepsilon_{i,t}, \quad (1)$$

where $Y_{i,t}$ is firm i 's reported net VAT in period t ; $\text{Treatment}_{i,j}$ is an indicator equal to one if firm i was assigned to arm $j \in \{\text{letter, phone, audit}\}$ and zero if assigned to control; $\mathbf{X}_{i,t_0-2}$ is a vector of pre-treatment firm characteristics measured at $t_0 - 2$; γ is the conformable coefficient vector; and α_t is a period intercept. We estimate the specification separately for each arm j , restricting the sample in each regression to that arm and the control group and excluding firms assigned to the other two arms. Because treatment is randomly assigned, it is orthogonal to potential outcomes, and β_j identifies the arm-specific ATE. Standard errors are clustered at the firm level.

A major challenge in our setting is statistical power. When treatment and control are balanced on $\mathbf{X}$, including controls is not necessary for identification, but because there is substantial heterogeneity in net VAT across firms and fluctuations between periods, including $\mathbf{X}$ may increase precision by soaking up residual variation. Including irrelevant controls, however, may reduce statistical power. We therefore select controls using a doubly robust LASSO procedure (Belloni et al., 2014; Ahrens et al., 2018). This procedure selects variables that are correlated with both the treatment and the outcome, together with variables that are strongly correlated with the outcome. We supply the full vector $\mathbf{X}$ as candidates.

Because our main findings are null, we also quantify the economic significance of the estimates using equivalence tests. We apply two one-sided tests (TOST) at the 5% level, which is equivalent to rejecting effects larger in magnitude than a threshold

Δ when the corresponding 90% one-sided confidence bound lies below Δ . All rule-out statements in the results that quantify economic magnitudes are based on this 90% equivalence criterion.

4 Results

4.1 Audit Outcomes

The audits disclosed misreporting with a detection rate of just under 9.3 percent, see Figure 2. Among these 99 cases, the median VAT adjustment was NOK 10,317 with an interquartile range of NOK 5,070 to NOK 32,381; the mean (NOK 30,694) exceeds the third quartile, reflecting a right-skewed distribution driven by a small number of large adjustments. The average audit adjustment across the whole treatment group (including the 970 cases where no error was found) was NOK 2,842, giving an estimated tax gap for the target population of around NOK 62.2 million for the reporting period, or roughly 6.3 percent of reported net VAT. This falls within the 3.0 to 8.9 percent range that the European Commission’s *EU VAT Gap Report* reports for comparable Nordic countries in 2024, consistent with Norway’s high-capacity tax environment. In short, the audits function as a detection tool: they uncover non-compliance at rates in line with the published tax-gap literature. Any audit requires resources, mostly in terms of staff time. The estimated average cost of the audits is about NOK 2,700, almost identical to the average net VAT adjustment of NOK 2,842. Scaled to FTE-equivalents, our random audits produce roughly NOK 0.85 million in detected revenue per caseworker year, close to break-even against the FTE cost of about NOK 0.8 million.³ By comparison, [Auditor General of Norway \(2025\)](#) reports that risk-based audits yield about NOK 7.4 million

³Based on internal NTA time-use data from the 2021 field study, which records an average of 0.6 working days per random audit. The 2021 cost per FTE for caseworkers was about NOK 800,000 (salary plus social costs of about NOK 715,000, plus office, IT, and supplies of about NOK 85,000), and the operational norm of 180 effective working days per FTE gives a per-day cost of about NOK 4,500. Risk-based audits use roughly 2.5 working days per audit (rising to about 3.5 by 2024 after new procedural requirements); the lower number for our experiment reflects the absence of post-detection follow-up steps for cases without adjustments.

in detected revenue and additional tax per FTE-year (2023–2024 average). The gap reflects both the higher hit rate (40–45 percent vs. our 9.3 percent) and the larger size of firms in the risk-based pool, relative to our random sample of newly registered firms. The next question is whether they changed the reporting practice of firms.

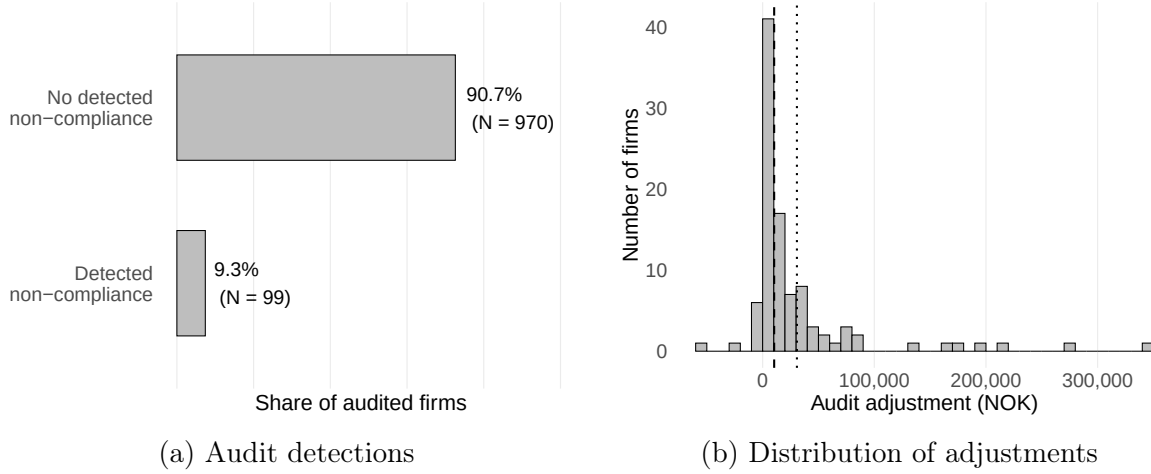


Figure 2: Audit detection rate and adjusted net VAT

Notes: Sample is the 1,069 firms assigned to the audit arm. Panel (a) shows the share of audited firms with and without detection in reporting period 4/2021 (the number of firms is indicated in parentheses). Panel (b) shows the distribution of adjustment amounts (in NOK) among the 99 firms where non-compliance was detected, binned at NOK 10,000 intervals. Adjustment amounts are unwinsorized and include both positive (firm owed more VAT) and negative (firm owed less) adjustments. The distribution is right-skewed with a small number of large adjustments. The dashed line indicates the median (NOK 10,317); the dotted line indicates the mean (NOK 30,694).

Source: Norwegian Tax Administration audit records.

4.2 Behavioral Effects on VAT Reporting

In short, we find no indication that the audit affected firms' future reporting. The main specification results in Figure 3 show that none of the OLS estimators are statistically significantly different from zero in any of the post-treatment periods, indicating that the audit had no discernible effect on firms' VAT reporting. The coefficient measured

at reporting period 4/2021 includes the audit adjustments themselves, which likely explains why it is distinctly higher than the other post-treatment point estimates. Figure 4 shows the LASSO-controlled estimates by post-treatment period in panel (a) and pooled across post-treatment periods in panel (b); the pooled estimate is more precise than the period-by-period estimates. Table A2 shows that the post-treatment point estimate is essentially unchanged when moving from the unadjusted specification (column 1) to the LASSO-controlled specification (column 3), while the standard error is roughly halved; columns 2 and 4 report the corresponding pre-treatment estimates. Using two one-sided equivalence tests at the 5 percent level, we rule out audit effects larger than about NOK 3,000 per firm per bi-monthly period, or less than 7 percent of the control-group mean.

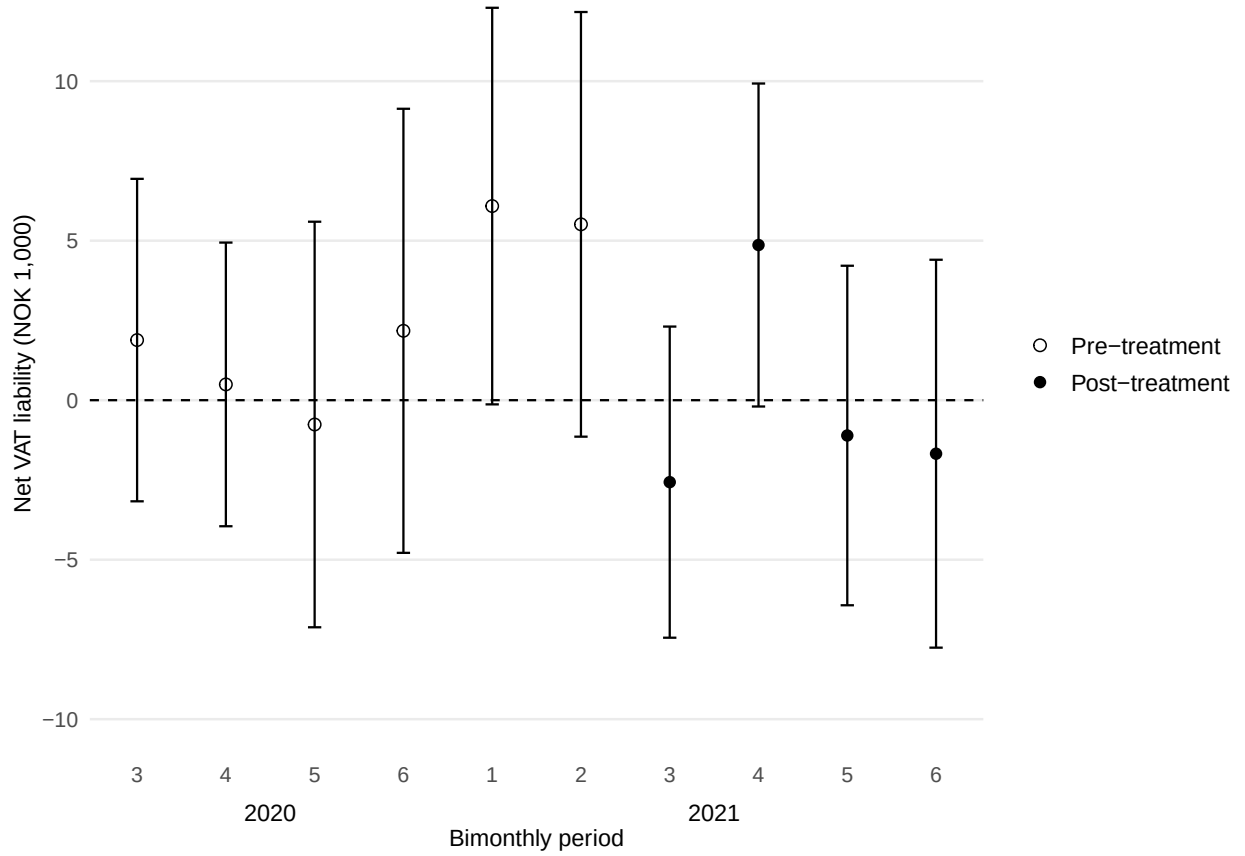


Figure 3: Period-wise effects of audits on reported net VAT

Notes: The figure plots period-specific estimates of the audit treatment effect on net VAT liability (NOK 1,000). Each point comes from a separate cross-sectional regression on firms in stratum 2 (13–24 months old at t_0), excluding those assigned to the letter or phone arm, with controls selected by post-double-selection LASSO.

Vertical lines indicate 95 percent confidence intervals based on heteroskedasticity-robust standard errors (equivalent to firm-level clustering with one observation per firm per period). Periods 3/2020 through 3/2021 are pre-treatment; the audit arm was assigned to period 4/2021 and the outcome in that period includes audit adjustments. Source: Norwegian Tax Administration administrative data.

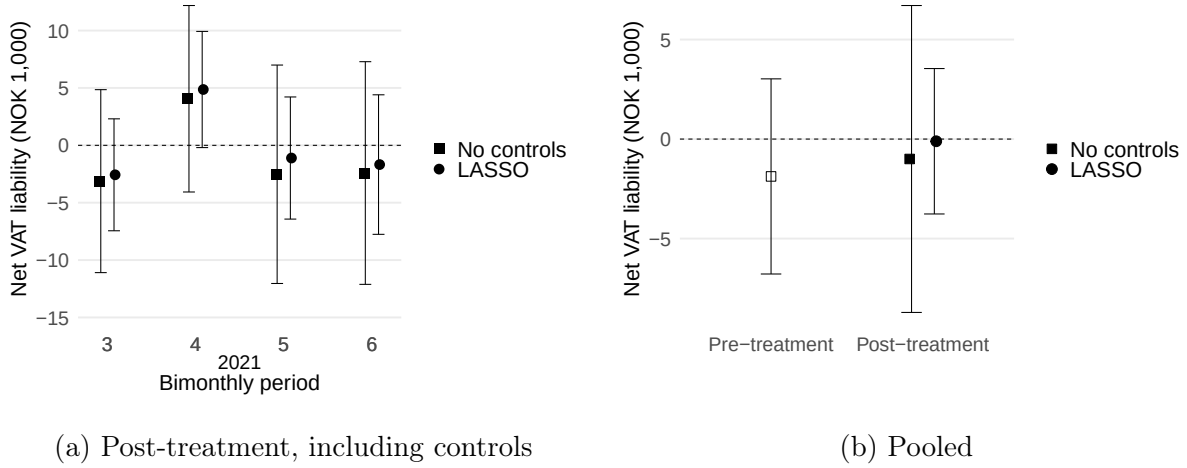


Figure 4: Period-wise and pooled effects, including control variables

Notes: Point estimates of the audit treatment effect on net VAT liability (NOK 1,000), with controls selected by post-double-selection LASSO. Panel (a) shows estimates by post-treatment reporting period, each from a separate cross-sectional regression; standard errors are heteroskedasticity-robust. Panel (b) pools pre- and post-treatment periods in a firm-level panel, with standard errors clustered at the firm level. Estimation sample in both panels is stratum 2 (13–24 months old at t_0), excluding firms assigned to the letter or phone arm. Vertical lines indicate 95 percent confidence intervals. Source: Norwegian Tax Administration administrative data.

4.3 Letter and Telephone Arms

Table 1 reports pooled post-treatment estimates for all three arms, using the same LASSO specification and estimation sample (firms in the older stratum). All three point estimates are small, negative, and statistically indistinguishable from zero. The letter arm yields an estimate of -0.48 (SE 1.70) on a control-group mean of NOK 45,090, and the equivalence test rules out letter effects larger than about NOK 2,300 per firm per bi-monthly period at the 5 percent level. The telephone arm yields -2.71 (SE 2.14); although larger in magnitude than the audit or letter point estimates, it is not statistically significant, and the 95 percent two-sided confidence interval extends from NOK $-6,900$ to NOK $+1,500$.

Table 1: Pooled post-treatment effects by arm

	(1)	(2)	(3)
	Net VAT	Net VAT	Net VAT
Audit	-0.11 (1.86)		
Letter		-0.48 (1.70)	
Telephone			-2.71 (2.14)
Mean in control group	45.09	45.09	45.09
N	77520	79816	76496
Arm	Audit	Letter	Telephone

Notes: Pooled post-treatment coefficients on net VAT (NOK 1,000) for each treatment arm, estimated by post-double-selection LASSO. The outcome is winsorized at p_1 and p_{99} ; missing post-treatment returns are coded as zero. Estimation sample is stratum 2 (13–24 months old at t_0), in which all four arms are represented, pooling reporting periods 3/2021 through 6/2021. Each column excludes the other two treatment arms and estimates the effect of the named arm against the stratum-2 control group. The LASSO candidate set is described in Section 3. Standard errors clustered at the firm level in parentheses.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

4.4 Interpretation

Several mechanisms are consistent with the null effect of the audit on future VAT reporting. We organize them around two questions: what does the audit teach the firm, and what does it cost the firm when misreporting is detected?

The first set of mechanisms concerns belief updating. A standard deterrence model predicts that an audit raises expected compliance costs through two channels: the perceived probability of audit and the sanction conditional on detection. Both channels are weaker in our setting than the statute would suggest. The standard additional tax is set at 20 percent of the established tax advantage and is, by statute, mandatory whenever the statutory conditions are met. In practice, however, the sanction is applied incon-

sistently: [Auditor General of Norway \(2025\)](#) reviewed 25 risk-based desk-audit cases that produced upward VAT adjustments and concludes that additional tax should have been imposed under the statute in 20 of them but was not. From the perspective of a newly registered firm, then, an audit conveys that the agency can produce a one-off correction without a financial penalty beyond the principal, and that even when misreporting is detected, the sanction-conditional-on-detection is well below the statutory rate. A firm that misreported and is corrected without sanction has reason to revise its expected penalty downward, not upward. Direct experimental evidence on the punishment channel in Norway comes from [Kotsadam et al. \(2026\)](#), who randomize the timing of additional tax assignment among detected noncompliers in a separate Norwegian experiment and find large behavioral effects of realized punishment, while audits alone have only small effects. Our null is consistent with that work in the sense that, where the sanction is not credibly applied, the deterrence channel does not operate.

A related mechanism operates among firms that misreported but were not detected. The desk-based audit reviews a single bi-monthly return against the documentation the agency requests; some misreporting will pass through. A firm that bent the rules and emerged from an audit with no adjustment learns that the conditional detection probability is lower than it had assumed. The probability of audit is revised upward by the experience of being audited, but the probability of detection given audit is revised downward. Whether the net effect on expected costs is positive or negative depends on which channel dominates, and our null is consistent with the two roughly canceling. This is the false-negative version of the so-called bomb-crater effect ([Mittone et al., 2017](#), see also the discussion in [Gemmell and Ratto, 2012](#)).

The second set of mechanisms concerns what the audit actually finds. The modest size of most detected adjustments (see Figure 2) suggests that the audits correct mechanical reporting errors rather than reveal deliberate evasion. Correcting a mechanical error in one period is unlikely to change subsequent reporting behavior because the firm was not optimizing against the audit margin in the first place. To the extent that

the audited population is dominated by such firms, behavioral persistence is small not because deterrence fails but because there is little misreporting behavior to deter.

A different mechanism, not based on belief updating, is the buyer–seller self-enforcement that [Pomeranz \(2015\)](#) documents in the Chilean VAT. The same paper-trail incentive operates in Norway, where buyers want invoices to deduct input VAT, with the well-known exception of final-stage transactions. Where this self-enforcement is already at work, the marginal value of any one desk-based audit is correspondingly smaller. Our null is consistent with that interpretation: among newly registered firms with mostly business-to-business transactions, much of the compliance margin is already covered by buyers’ own incentives.

Two comparisons sharpen the interpretation. The contrast with [Bjørneby et al. \(2021\)](#), whose unannounced on-site payroll audits in the same institutional setting shift reported wages by about 18 percent, suggests that audit intensity and the credibility of the attached sanction matter together: the more intensive on-site modality moves behavior; the lower-intensity desk modality, with sanctions applied inconsistently, does not. The contrast with [Li and Wang \(2026\)](#) runs the other way. In their Chinese setting, audited firms increase both reported output and input VAT credits, with a net negative effect on remittances driven by large-firm avoidance through consultants, related-party transactions, and fake invoicing. Our population of newly registered firms has limited scope for that kind of substitution. Our null is therefore cleaner evidence on the deterrence channel itself, not masked by offsetting margins of behavioral response.

5 Conclusion

This paper studies the effects of low-intensity tax enforcement interventions on VAT compliance among newly registered firms in Norway, using a large-scale randomized field experiment. Around 20,000 firms were randomly assigned to desk-based audits, informational letters, phone-based guidance, or a control group, allowing us to credibly estimate both immediate fiscal effects and longer-run behavioral responses in VAT

reporting.

Results show that desk-based audits uncover non-compliance in a non-trivial share of cases, with detected errors in just over 9 percent of audited firms and an implied tax gap of approximately 6.3 percent for the target population. The audit costs are very similar to the increased VAT collection. However, despite this detectable level of non-compliance, we find no evidence of persistent increases in reported net VAT following the audit. Estimated effects are small and statistically indistinguishable from zero, even when using post-selection methods to improve precision. Informational letters and phone-based guidance had similarly negligible effects.

Taken together, these results suggest that low-intensity enforcement tools, when implemented in isolation and with limited sanctions, may be insufficient to generate durable improvements in VAT compliance among new firms. One interpretation is that the informational content of desk-based audits is too weak to materially update beliefs about detection risk, particularly in an institutional environment without comprehensive third-party reporting. Indeed, an audit without a credibly applied sanction may convey the wrong information: a firm that is corrected without penalty, or that emerges from an audit with undetected misreporting intact, has reason to revise its expected cost of non-compliance downward. Our null findings on behavioral effects are not driven by low statistical power: equivalence tests rule out audit effects larger than about NOK 3,000 per firm per bi-monthly period, or less than 7 percent of the control-group mean, and the letter and telephone bounds are tighter still.

From a policy perspective, these findings highlight the importance of enforcement design rather than enforcement intensity per se. Desk-based audits are noticed by firms and generate first-stage corrections in 9.3 percent of cases, yet remain too weak to affect subsequent reporting. The policy-relevant question is therefore not whether to audit, but what form of audit, at what frequency, and paired with what follow-up, produces durable compliance. Useful next steps include stronger sanctions attached to detected non-compliance, repeat contact with the same firm over multiple periods, and adaptive

targeting rules that use flags from prior returns to allocate audit capacity.

An additional perspective offered by our findings is that the compliance effects of audits may differ systematically between firms and individuals. Much of the experimental evidence documenting persistent deterrence effects comes from audits of individual or self-employed taxpayers, where reporting is comparatively simple, and audits may more directly shift perceived detection risk (e.g., [Kleven et al., 2011](#); [DeBacker et al., 2018](#); [Hebous et al., 2023](#)). In contrast, evidence on audits of firms, particularly in the context of VAT, points to more limited and context-dependent behavioral responses. [Best et al. \(2021\)](#) document substantial detected evasion among firms in Pakistan but find little evidence of subsequent deterrence, attributing this to limited fiscal capacity and weak enforcement follow-up. Despite uncovering non-trivial non-compliance among newly registered firms, our results indicate that low-intensity desk-based audits do not generate medium-term increases in reported VAT. Although two studies cannot establish a general pattern, the near-zero point estimates in both the Norwegian high-capacity and the Pakistani lower-capacity settings suggest that when firm audits lack strong follow-up, meaningful sanctions, or complementary information structures, their deterrence effects may be limited across a range of institutional environments.

By providing a precisely estimated null effect in a large-scale randomized setting, our study identifies conditions under which audits of firms do not induce lasting behavioral change. Showing that three standard low-intensity enforcement tools leave reported VAT unchanged among newly registered Norwegian firms provides a tight benchmark for what this class of interventions can and cannot do. Future work that raises the intensity of enforcement, or targets a different margin, can be evaluated against this benchmark. The institutional setting we study (a high-capacity tax administration operating without comprehensive third-party invoice reporting on small newly registered VAT-liable firms) is shared with several OECD economies, suggesting these results extend beyond the Norwegian context.

References

- ADVANI, A., W. ELMING, AND J. SHAW (2023): “The Dynamic Effects of Tax Audits,” Review of Economics and Statistics, 105, 545–561.
- AHRENS, A., C. B. HANSEN, AND M. E. SCHAFFER (2018): “PDSLASSO: Stata module for post-selection and post-regularization OLS or IV estimation and inference,” <https://ideas.repec.org/c/boc/bocode/s458459.html>.
- ALMUNIA, M., J. KNEBELMANN, D. NAKYAMBADDE, AND L. TIAN (2025): “Firm networks and tax compliance: Experimental evidence from Uganda,” IFS Working Paper 25/51, Institute for Fiscal Studies.
- AUDITOR GENERAL OF NORWAY (2025): “The Auditor General’s Investigation of the Tax Administration’s Control of Value-Added Tax [Riksrevisjonens undersøkelse av Skatteetatens kontroll med merverdiavgift],” Document 1 (2025–2026), Auditor General of Norway (Riksrevisjonen), Oslo.
- BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2014): “Inference on treatment effects after selection among high-dimensional controls,” The Review of Economic Studies, 81, 608–650.
- BEST, M. C., J. SHAH, AND M. WASEEM (2021): “Detection without Deterrence: Long-Run Effects of Tax Audit on Firm Behavior,” Working paper. Randomized audits of Pakistani VAT filers, 2013–2017. The paper’s own description of the audit procedure: “All random audits in our sample are comprehensive audits... Auditors spend considerable time with taxpayers going through their records, visiting their premises.”
- BJØRNEBY, M., A. ALSTADSÆTER, AND K. TELLE (2021): “Limits to third-party reporting: Evidence from a randomized field experiment in Norway,” Journal of Public Economics, 203, 104512.

- BONING, W. C., J. GUYTON, R. HODGE, AND J. SLEMROD (2020): “Heard it through the grapevine: The direct and network effects of a tax enforcement field experiment on firms,” Journal of Public Economics, 190, 104261.
- BONING, W. C., N. HENDREN, B. SPRUNG-KEYSER, AND E. STUART (2025): “A welfare analysis of tax audits across the income distribution,” Quarterly Journal of Economics, 140, 63–112.
- BROCKMEYER, A., M. HERNANDEZ, S. KETTLE, AND S. SMITH (2019): “Casting a wider tax net: Experimental evidence from Costa Rica,” American Economic Journal: Economic Policy, 11, 55–87.
- BROCKMEYER, A., G. MASCAGNI, V. NAIR, M. WASEEM, AND M. ALMUNIA (2024): “Does the Value-Added Tax Add Value? Lessons Using Administrative Data from a Diverse Set of Countries,” Journal of Economic Perspectives, 38, 107–132.
- CARRILLO, P., D. POMERANZ, AND M. SINGHAL (2017): “Dodging the taxman: Firm misreporting and limits to tax enforcement,” American Economic Journal: Applied Economics, 9, 144–164.
- DEBACKER, J., B. T. HEIM, A. TRAN, AND A. YUSKAVAGE (2018): “Once bitten, twice shy? The lasting impact of enforcement on tax compliance,” The Journal of Law and Economics, 61, 1–35.
- GANGL, K., B. TORGLER, E. KIRCHLER, AND E. HOFMANN (2014): “Effects of supervision on tax compliance: Evidence from a field experiment in Austria,” Economics Letters, 123, 378–382.
- GEMMELL, N. AND M. RATTO (2012): “Behavioral Responses to Taxpayer Audits: Evidence From Random Taxpayer Inquiries,” National Tax Journal, 65, 33–57.
- HARRISON, G. AND R. KRELOVE (2005): “VAT Refunds: A Review of Country Experience,” IMF Working Paper WP/05/218, International Monetary Fund, Washington, DC.

HEBOUS, S., Z. JIA, K. LØYLAND, T. O. THORESEN, AND A. ØVRUM (2023): “Do Audits Improve Future Tax Compliance in the Absence of Penalties? Evidence from Random Audits in Norway,” Journal of Economic Behavior & Organization, 207, 305–326.

KLEVEN, H. J., M. B. KNUDSEN, C. T. KREINER, S. PEDERSEN, AND E. SAEZ (2011): “Unwilling or Unable to Cheat? Evidence From a Tax Audit Experiment in Denmark,” Econometrica, 79, 651–692, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA9113>.

KOTSADAM, A., K. LØYLAND, K. R. MADLAND, O. RAAUM, G. TORSVIK, AND A. ØVRUM (2026): “The Effect of Punishment on Tax Compliance: Evidence from a Randomized Experiment,” Memorandum XX/2026, Department of Economics, University of Oslo, Oslo.

LI, H. AND X. WANG (2026): “Firm responses to tax audits: Regression discontinuity evidence from a threshold-based audit program,” The Economic Journal, 136, 560–601.

LØYLAND, K., O. RAAUM, G. TORSVIK, AND A. ØVRUM (2024): “Evaluating compliance gains of expanding tax enforcement,” Economica, 91, 142–162.

MITTONE, L., F. PANEBIANCO, AND A. SANTORO (2017): “The bomb-crater effect of tax audits: Beyond the misperception of chance,” Journal of Economic Psychology, 61, 225–243.

NARITOMI, J. (2019): “Consumers as tax auditors,” American Economic Review, 109, 3031–3072.

OECD (2025): Tax Administration 2025: Comparative Information on OECD and Other Advanced an
Paris: OECD Publishing.

ORTEGA, D. AND C. SCARTASCINI (2020): “Don’t blame the messenger. The delivery

method of a message matters,” Journal of Economic Behavior & Organization, 170, 286–300.

POMERANZ, D. (2015): “No Taxation without Information: Deterrence and Self-Enforcement in the Value Added Tax,” American Economic Review, 105, 2539–2569.

SLEMROD, J. (2019): “Tax compliance and enforcement,” Journal of Economic Literature, 57, 904–954.

THE NORWEGIAN TAX ADMINISTRATION (2024): “Årsrapport 2023,” .

Online Appendix

Table A1: Balance across treatment arms

	Control	Letter	Telephone	Audit	<i>p</i>
Age (weeks)	94.07	93.80	94.29	93.82	0.809
Employees	2.89	2.91	2.39	2.48	0.218
Total outside sales	10.55	13.85	14.59	13.44	0.051
VAT return: to pay	0.63	0.63	0.63	0.62	0.857
VAT return: credit	0.26	0.26	0.24	0.26	0.380
VAT return: zero	0.11	0.10	0.13	0.12	0.314
Reactivated	0.03	0.04	0.03	0.03	0.957
Certified auditor	0.18	0.19	0.17	0.19	0.456
Self-employed (ENK)	0.37	0.36	0.38	0.38	0.773
Audits last 2y	0.09	0.09	0.08	0.10	0.742
Flags last 2y	0.99	1.03	0.94	1.03	0.540
Detections last 2y	0.02	0.03	0.02	0.02	0.915
Exporter ($t_0 - 2$ to $t_0 - 7$)	0.07	0.06	0.07	0.07	0.919
Importer ($t_0 - 2$ to $t_0 - 7$)	0.25	0.27	0.26	0.27	0.316
Industry dummies	10 industries, joint F -tests $p \geq 0.107$				
County dummies	10 counties, smallest joint F -test $p = 0.074$				
N	18,366	1,647	813	1,069	

Notes: Means of pre-treatment firm characteristics by treatment arm, measured at $t_0 - 2$ (reporting period 1/2021). Sample is stratum 2 (13–24 months old at t_0), in which all four arms are represented. The final column reports the p -value from regressing each covariate on three treatment-arm dummies (control omitted) with heteroskedasticity-robust standard errors and testing the joint null that the three dummies are zero. For brevity, the ten industry and ten county dummies are summarized in two lines reporting the smallest joint F -test p -value within each group. No covariate shows imbalance at conventional levels. Source: Norwegian Tax Administration administrative data.

Table A2: Table of coefficients, main results

	(1)	(2)	(3)	(4)
	Net VAT	Net VAT	Net VAT	Net VAT
Audit	-1.00 (3.93)	-1.09 (2.95)	-0.11 (1.86)	2.62 (2.00)
Mean in control group	45.06	34.39	45.09	34.43
N	77740	116262	77520	116138
Period	Pooled post	Pooled pre	Pooled post	Pooled pre
Controls	No	No	Yes	Yes

Notes: Estimates of the audit effect on net VAT (NOK 1,000), pooling firm-level observations across reporting periods. The outcome is winsorized at p_1 and p_{99} ; missing post-treatment returns are coded as zero. Columns 1 and 3 pool the four post-treatment periods (3/2021 through 6/2021); columns 2 and 4 pool six pre-treatment periods (3/2020 through 2/2021). Columns 1 and 2 are OLS without controls. Columns 3 and 4 apply post-double-selection LASSO, with a different candidate set in the two cases: the post-treatment specification uses the full covariate vector described in Section 3, while the pre-treatment specification is restricted to pre-period lagged outcomes. Estimation sample is stratum 2 (13–24 months old at t_0), excluding firms assigned to the letter or phone arm. Standard errors clustered at the firm level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

A.1 Pre-Analysis Plan

Pre-analysis plan for: Tax enforcement interventions to improve VAT-reporting

Andreas Kotsadam, Knut Løyland, Oddbjørn Raaum, Gaute Torsvik and Arnstein Øvrum

December 10 2021

Abstract

Using a large scale field experiment, we compare the effects of three different tax enforcement policies to increase compliance with VAT reporting for newly established firms. In this pre-analysis plan we describe the treatments and lay out some important decisions with respect to coding of variables, definitions of samples and the empirical strategy we will apply.

1 Introduction

Any legal unit in Norway with more than 50, 000 NOK in yearly sales is obliged to submit a tax return for VAT. Many start-ups fail to report their net VAT correctly, either because they are confused about the rules or deliberately try to bend/avoid the rules.

This pre-plan describes how to analyse future data from a field experiment that aims to uncover how different tax enforcement policies affect the compliance with the VAT regulation among young businesses. We will compare the effect of different policies, both preventive information treatments and a reactive control/audit treatment.

2 Data and treatments

The population is businesses with first time registration in the VAT register in the period from 01.05.2019 until 30.04.2021 and bi-monthly VAT tax returns, including around 50,000 entities. These entities are randomly allocated to four different groups; a control group (with no special intervention), two information treatments (letter and telephone) and one audit treatment. The businesses are either limited companies ("AS") or self-employed ("Enkeltmannsforetak").

2.1 The treatments and sampling

The **Letter** treatment informs the entity that a dedicated National Tax Authority (NTA) caseworker can be contacted if need to sort out any question and give contact information the him/her. We code a dummy variable, **letter**, which is equal to one if the entity was assigned to the letter treatment.

The **Telephone** goes one step further as it involves a call from a NTA caseworker where the entity is offered guidance if needed. We code a dummy variable, **telephone**, which is equal to one if the entity was assigned to the telephone treatment.

The **Audit** is similar to a desk based correspondence audit. If an entity is selected for an audit, NTA is obliged to inform the entity ahead of inspection. Thus, all audited entities are aware that they are audited. The main aim of the audit is to check whether the balances in the VAT tax return are correct. If errors are suspected, further documentation of transactions are collected and inspected by an NTA employee. Two months after the deadline of reporting,

about 80 percent of the audits are finished. We code a dummy variable, *audit*, which is equal to one if the entity was assigned to the audit treatment.

Each VAT tax return to the NTA in our data is based on transactions over a two month period, with a lag. The deadline for reporting VAT is the 10th in the second month following the end of the period.¹ Period 1 refers to entities that report VAT for the period January and February and so on. Hence, the deadline for reporting period 1 VAT is April 10.

We sampled entities on August 20 2021 that have a first time registration in the VAT register in the period from 01.05.2019 until 30.04.2021. That is, all entities are in the data in period 2 in 2021 and registered during the previous 24 months. The Treatments start on September 15, which is before the deadline for reporting of period 4 (i.e. October 10). Period 4 in 2021 is therefore our t_0 and period 2 is t_0-2 .

The **Telephone** and **Letter** treatments are finished on October 10 so t_0 is our first outcome period for any behavioural effects of these treatments. As some have been contacted very close to this reporting date, it is unclear if they have had time to respond already at t_0 .

The **Audit** treatment is still ongoing. The audit is of period 4 in 2021, but some may be treated before October 10 if they deliver before the deadline (they may then get notified of an impending audit). We will therefore see if there is a direct audit effect in t_0 . The audit may disclose non-compliance and this direct effect is a part of the treatment effect. Any behavioural effects, however, will not be seen before t_0+1 .

2.2 Strata and samples

Entities for letter and telephone treatments were randomly drawn from two strata, one for businesses in their infancy (1-12 months old) and one for entities in their early childhood (13-24 months) at the date of the sampling (August 20 2021). We code a dummy variable, *Strata2*, which is equal to one for entities that are 13-24 months old.

The businesses selected for audit were drawn randomly from early childhood entities (13-24 months), i.e. from the same population as the second strata for letter and phone interventions, also on August 20 2021.

¹ An exemption is given for the 3rd period (May and June) for which the deadline is August 31th.

Our total sample consists of 50,495 entities out of which 44,528 are in the control group and 5,967 are treated. Table 1 shows the distribution across treatments and strata.

Table 1. Samples

Strata (age at t_0-2)	Audit		Letter		Telephone	
	1-12	13-24	1-12	13-24	1-12	13-24
Treatment	0	1 059	1 641	1 639	814	814
Control	0	18 389	26 139	18 389	26 139	18 389
Total	0	19 442	27 780	20 022	26 953	19 197

2.3 Outcomes and sample trimming

According to the regulations, an entity should file a VAT tax return and the main items are total sales and net VAT. In the simplest case, net VAT equals $0.25 \times (\text{Total sales} - \text{Input costs})$ where input is material and intermediate products/services used in the production. In a given period the net VAT can be negative and the business will have it transferred directly. Thus, the firm has an incentive to underreport sales (for which it have collected VAT) and overreport input costs. The report is net VAT and not the two components.

Our aim is to investigate and compare how each of the three different treatments, the phone, the letter and the audit influence direct and subsequent VAT behaviour. Our post treatment time frame will allow for an estimation of direct effects in the 4th period (t_0) and at least one year of “behavioural” post treatment effects for the audit. That is, covering the 6 post-treatment VAT reporting periods starting with the 5th term, Sept-Oct 2021 (t_0+1). The report deadline for the 5th term in 2021 is December 10. We consider three different outcomes in each post-treatment period.

- Net VAT to pay (NETVAT)
- Total turnover (TOTSALLES, within the VAT-domain)
- Number of formal errors on task (checked by mechanical control) (ERRORS)

All of these variables are continuous and will not be recoded in any way. NETVAT is our main outcome variable for the **Audit** treatment while ERRORS is the main outcome for the **Telephone** and **Letter** treatments.

In the post-treatment periods, there will be missing observations on the outcomes. Entities are closed down and some will temporarily fail to report. Some entities will also have missing observations because they are removed from the VAT register by the NTA as a result of the audit (or the communication triggered by the letter of telephone). From the firms' perspective, being part of the register gives some advantages (e.g. when NETVAT is negative). The missing observations are not expected to be random or unaffected by treatment. Excluding missing observations would thus create a selection bias. Thus, missing observations on the post treatment outcomes will be coded as zero and included in the analyses (without a dummy variable for missing status). All period outcome variables will be winsorized at p1 and p99.

Data trimming practice

To avoid estimates that are driven by extreme values we will winsorize the first two outcome variables (TOTSALLES and NETVAT in 1, 000 NOK) by replacing observations in the two tails ($< p1$ and $> p99$), with the minimum ($p1$) and maximum ($p99$) values. This winsorization is done separately for every observation period.

2.4 Coding of covariates

In addition to the outcome variables we will code a number of variables used for balance tests, to improve precision, and in the search for heterogeneous treatment effects. All of these variables are predetermined and fixed at t_0-2 .

Lagged outcomes: Continuous values of six lags of the dependent variables (t_0-2 to t_0-7). Note that we do not use t_0-1 as the reporting for this period may have changed retrospectively due to the treatment.

Age: Continuous variable of age of the entity in weeks.

Counties: Dummy variables for county location (10) of the firm. The two smallest northern counties are grouped together.

Industry affiliation: Based on NACE and the national standard used by the NTA, called "Hovednæring". All industries with less than 5% of the observations are allocated to a category that we label "other".

Number of employees in the entity: Continuous

Certified auditor: A dummy variable for whether the entity has an external certified auditor to check its accounts.

Self-employed: A dummy variable if the business is a single-person entity ("ENK" in the register data) and not a limited company ("AS").

Exporting: A dummy variable if the business has exported at least once during (t_0-2 to t_0-7).

Importing: A dummy variable if the business has imported at least once during (t_0-2 to t_0-7).

Total outside sales: Total sales outside the VAT-domain (VAT legislation domain)

Tax return to pay: Dummy variable for whether the tax return is net VAT to pay

Tax return credit: Dummy variable for whether the VAT tax return includes net VAT credit

Tax return zero: Dummy variable for whether the VAT tax return is zero

Number of flags: Number of times the entity has been flagged for suspicious VAT tax returns between t_0-13 and t_0-2 .

Number of audits: Number of times the entity has been audited between t_0-13 and t_0-2 .

Number of detections: Number of times evasion has been detected between t_0-13 and t_0-2 .

Reactivated: Dummy variable for whether the entity has been in the VAT register before and is reactivated during the period that defines our sample.

We refer to this vector of controls as **X**.

We also have the following Strata variable:

Strata2: A dummy variable for whether the entity belongs to the oldest (13-24 months=1) or the youngest (1-12 months=0) entities in the target population. This variable is only relevant for Telephone and Letter.

Additional variables used for heterogeneity analyses

We will also likely code a set of extra variables that will be used in the heterogeneity analyses. These are characteristics of the daily managers, owners, and board members. We also hope to get data on external accountants. We do not yet know which variables we can access here and we are therefore not able to provide details of the coding for these variables. As such, analyses with these variables will be seen as more exploratory. We refer to these variables as the vector **H**.

3 Pre-treatment analysis

We will test if the randomization achieved balance across groups. For the outcome of the entities i in the last pre-treatment period t_0 (and earlier) we will estimate:

$$(1) Y_{it_0-k} = \alpha_{t_0-k} + \beta_j Treatment_j + u_{t_0-k}$$

$k=2,...,7$, separately by treatment and outcome. For the **Letter** and **Telephone** treatments, we also include a strata control (i.e. a **Strata2** fixed effect). We will also check that all covariates in **X** are balanced by running regressions such as (1) for the variables both individually and together. We will base our judgement of the randomization on an F-test of all the **X** variables included together after controlling for strata fixed effects (when applicable).

4 Post-treatment effect analyses

Main hypotheses

We expect a non-negative average effect on the post-treatment NETVAT reported by the entities. Audits will disclose mistakes that are likely to be biased towards underreporting and the improved knowledge about the rules is expected to raise NETVAT reporting in the post-audit period. It's not obvious, however, for how long such an effect will last. Moreover, there might be a (small) fraction of "false-negatives" in the audit who update their expectations and actually lower their compliance. Since they got away with misreporting this time, they might reduce their NETVAT next time (Gemmell and Ratto 2015).

For the letter and telephone interventions, which are more of the "intention to treat" type, the entities can choose to take actions that (presumably) will raise reported NETVAT. The idea is that businesses have strong incentives to (already) learn their rights to reduce the NETVAT they should pay. Therefore, the additional information triggered by the interventions are likely to be tilted in favor of the tax authorities. For these treatments our main interest is, however, on ERRORS as the goal of these interventions was to reduce costs due to unnecessary mistakes and make the collection of VAT more efficient.

Empirical strategy

At the date of filing this report, no post-treatment data has been accessed. This section describes how will analyze the data and how to test the effects of the tax enforcement interventions. For each post-treatment outcome in period (t_0+s) we estimate the average treatment effects (ATE) by an OLS model of the following specification;

$$(2) Y_{t_0+s} = \alpha_{t_0+s} + \beta_j Treatment_j + \gamma_{t_0-2} Controls + u_{t_0+s}$$

As time (s) goes by it, is of core interest to see to what extend any causal impact prevails.

The regressions for **Telephone** and **Letter** will always control for the strata variable **Strata2**. We use robust standard errors in all estimations unless otherwise stated.

If we have missing values on explanatory variables we will code the variables as zero and include dummy variables controlling for missing status so that we do not lose observations. Remember that missing observations on the post treatment outcomes will be coded as zero and included in the analyses (without a dummy variable for missing status).

A major challenge in our setting is statistical power. Since the treatments are relatively cheap, even small effects on the NETVAT-reporting and on ERROR are of interest to the tax authorities. There are large fluctuations in NETVAT by period, including both positive and negative numbers over time, as well as huge heterogeneity. We will therefore see if we can increase precision by adding control variables.

To the extent that the samples are balanced across treatments it is not necessary to include the vector **X** in the estimation. On the other hand, including controls may increase power by soaking up residual variation. We expect the lagged values of the dependent variables to be especially important in this respect. Including too many and non-relevant controls may, however, lead to less power. We will therefore use a doubly robust LASSO procedure to select optimal control variables (Belloni et al. 2014; Ahrens et al. 2018). This procedure selects variables that are correlated with both the treatment and the outcomes, if any, and otherwise only variables that are strongly correlated with the outcomes. As there is no limitation on how many variables that can meaningfully be included in the LASSO regressions we will include all variables in **X**. We may also add variables from **H** to explore if precision can be improved

further. Given the expected explanatory power of the lagged variables we do not expect large differences of including **H**. In any case, our main specification will be results where we limit the potential control set to **X**.

As we are concerned that the precision of the estimates will be low due to large fluctuations we will also explore the distribution of effects by testing if there are effects on being in the group above the median and above and below different quartiles of the distribution (see e.g. Pomeranz, 2015).

For the **Telephone** and **Letter** treatments we will explore whether it is useful to scale the effects by the share of people that the NTA managed to contact. This will be done in an IV regression, instrumenting contact with the treatment assignment to estimate the causal effect of letter received and telephone contact.

Comparisons of effects across treatments

For audits, there is just one sample of young businesses 13-24 months old at t_0 , *i.e.* one parameter per outcome-period. For letter and telephone there are potentially differential treatment effects between young (13-24 months) and very young businesses (1-12 months). If treatment effects differ by age, we would also like to compare treatment effects conditional on being a young firm. Hence, we will also estimate versions of equation (2) with an interaction term (Treatment* **Strata2**) for **Telephone** and **Letter**. If the interaction term is not statistically significantly different from zero we will proceed to compare the effects using the estimates from the full samples. If at least one of the interaction terms is statistically different from zero, however, we will also compare the effects when restricting the sample to young (13-24 months) businesses only.

5 Heterogeneous effects

Investigating heterogeneous effects is interesting in order to understand if there are differences in effects for different groups and to understand mechanisms. For optimal audit targeting it is also key to know how different groups respond to being audited. Testing for treatment effect heterogeneity entails its own set of problems, however, and there are pitfalls of naively splitting the data to test for effects across subgroups. Multiple hypotheses are

tested which is likely to lead to false positives if not corrected for. Without a pre-specified analysis plan it is impossible to know how many non-reported tests were conducted by the researchers. Specifying all possible tests is, however, difficult, especially in a setting where there is an inherent uncertainty with respect to treatment heterogeneity. With few covariates, one can simply include treatment covariate interactions between all covariates and estimate treatment heterogeneity using traditional regression methods, but usually the number of covariates is high relative to the sample size, which renders the traditional approach susceptible to overfitting. Overfitting is a more general problem in heterogeneity analysis as well and is induced by testing too many hypotheses unless p-values are adjusted. With adjustment of p-values, most analyses are not sufficiently powered to test many hypotheses. Hence, there is a tradeoff between pre-specifying hypotheses and learning about heterogeneity from the data.

In our setting it is also important as there may be effects that go in different directions. In particular, some firms may realize that the controls are not as good or extensive as they feared and may therefore report less VAT after they have been checked. Other firms may think it is unlikely that they will be checked again, e.g., due to beliefs about non-replacement in random audits. These latter effects have been labelled “bomb-crater” effects in the literature (Mittone et al. 2017).

We will solve both of the problems described above by using credible machine learning methods to detect heterogeneity. There are many different types of machine learning algorithms and we have decided to use the “Generic ML” approach by Chernozhukov et al. (2018). As this field is moving rapidly, however, it is possible that there will be other techniques that are relevant for us once we start analyzing the data.

The “Generic ML” approach has several advantages as compared to other approaches. First of all, it uses several machine learning algorithms and selects the ones that are most appropriate for the data at hand. Secondly, it provides an omnibus test of heterogeneity in the data. Thirdly, it accounts for partitioning uncertainty. ML results can be sensitive to the specific partitioning into training and test data set. Thus, with a single data-split, there is a risk that the results are non-typical for the universe of possible results from different splitting. Chernozhukov et al. (2018) solve this problem by repeating the procedure above for a large number of partitions and report the median estimates across the sample splits.

The approach consists of the following steps. First we partition the data into training and test data set. Then we use the training set to predict the outcome, given the covariates and treatment status. From these regressions we derive the conditional average treatment effects (CATEs). The predictions are made using standard ML methods and a procedure is used to select the ML method that produces the most accurate predictions in the test data set. This test is based on comparing the Best Linear Predictor (BLP) and best predictions for Group Average Treatment Effects (GATES). For the chosen best method, we classify units into groups based on the CATEs. One type of grouping is to split the units into five groups based on their CATE, and set the splits so that they explain as much variation in the CATEs as possible. We then measure the average treatment effect in each group (GATES) and examine how different the treatment effects are in the different groups. Next we will describe the covariate characteristics of units in the least and most affected group (CLAN) to understand the treatment heterogeneity.

The approach will tell us whether there is heterogeneity in the treatment effects overall. If the omnibus test suggests that there is no heterogeneity, we will not present further results from the method. It will also show if there are some groups for which the treatment effect is going in the negative direction (the “least affected” groups may have negative treatment effects). Finally, it will allow for a characterization of which types of firms are most affected while avoiding the pitfalls of regression based heterogeneity analyses. We will conduct the main heterogeneity analyses with all the variables in **X**, but may explore additional heterogeneity by adding variables in **H**.

For ERROR it will also be interesting to explore whether the effects are different after a new system of VAT reporting (MEMO) is introduced in period t_0+2 .

6 Power

We have different sample sizes and different shares of treated entities in the different treatment arms. We have fewest observations for the **Audit** treatment since this treatment only covers entities in **Strata2**. We also have the lowest share of treated entities for this treatment since it requires more NTA resources. We have 19,442 observations but only 1,059 assigned to audit. At the conventional level of significance of 0.05 and a power of 0.8, this would allow for a minimum detectable effect of 0.09 standard deviations.

We will also adjust the p-values for the fact that we are testing the impact on three outcomes. We follow the recommendations of Fink et al (2014) and use a method developed by Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) to minimize the false non-discovery rate. The main advantage of the method is that it is limiting the risk of false discoveries while only adjusting the critical values based on other true hypotheses. The false discovery rate method developed by Benjamini and Hochberg (1995) implies that the m p-values of the i hypotheses are ordered from low to high and that the critical value of the p-value is then $p(i) = \alpha \cdot i / m$. To illustrate, with three hypotheses and a significance level (α) of 0.05, the critical p-value would be 0.016 for the one with the lowest p-value ($0.05 \cdot 1/3$, which is the same as a Bonferroni correction. For the second hypothesis, the critical p-value is 0.033 ($0.05 \cdot 2/3$). For the third hypothesis the critical value is just 0.05. The minimum detectable effect if our variable with the lowest p-value is **Audit** after accounting for multiple hypothesis testing ($p=0.016$) is below 0.12 standard deviations. In addition, we expect that the control variables will explain a large part of the residual variation with R-squared values up to 0.6. We would then be powered to detect effects of around 0.1 standard deviations. We conclude that our experiment is reasonably well powered.

7 Archive and data disclosure

The pre-analysis plan is archived before any post treatment data has been looked at and the audit treatment is still ongoing. The latest data we have looked at corresponds to (t_0-2) . We archive it at the registry for randomized controlled trials in economics held by The American Economic Association: <https://www.socialscienceregistry.org/> on December 10 2021, which is the same day as the VAT-reporting deadline for period (t_0+1) . This data will be available from December 13 and expect to start to analyze it sometime in 2022.

8 References

- Ahrens, A., Hansen, C. B., & Schaffer, M. (2018). PDSLASSO: Stata module for or post-selection and post-regularization OLS or IV estimation and inference.
- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608-650.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289-300.
- Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165-1188.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68.
- Fink, G., McConnell, M., & Vollmer, S. (2014). Testing for heterogeneous treatment effects in experimental data: false discovery risks and correction procedures. *Journal of Development Effectiveness*, 6(1), 44-57.
- Gemmell, N. and Ratto, M. (2012). Behavioral responses to taxpayer audits: Evidence from random taxpayer inquiries. *National Tax Journal*, 65(1):33.
- Pomeranz, D. (2015). No Taxation without Information: Deterrence and Self-Enforcement in the Value Added Tax. *American Economic Review*, 105(8): 2539–2569.
- Mittone, L., Panebianco, F., & Santoro, A. (2017). The bomb-crater effect of tax audits: Beyond the misperception of chance. *Journal of Economic Psychology*, 61, 225-243.