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 he distribution across treatments and strata.

Table 1. Samples

	Audit		Letter		Telephone	
Strata (age at t_0-2)	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*(Total sales – 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 (TOTSALES, 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 (TOTSALES 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_o-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_o-13 and t_o-2.

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

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

Reactivated: Dummy variable for whether the entity has been it 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) $Y_{it0-k} = \alpha_{t0-k} + \beta_j Treatment_j + u_{t0-k}$

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 (Gemmel 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;

$$(2)\ Y_{t0+s} = \alpha_{t0+s} + \beta_j\, Treatment_j + \gamma_{t0-2} Controls + \ u_{t0+s}$$

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 $t0+2$.

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) =a*i/m . To illustrate, with three hypotheses and a significance level (a) of 0.05, the critical p- value would be 0.016 for the one with the lowest p-value (0.05* 1/3, which is the same as a Bonferroni correction. For the second hypothesis, the critical p- value is 0.033 (0.05*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.