CESifo
CONFERENCES 2021

12th Norwegian German Seminar on Public Economics

Munich, 5 – 6 November 2021

Does perceived risk of future audits explain the behavioral effects of tax enforcement?

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

CESifo NHH OFS Oslo Fiscal Studies

Does perceived risk of future audits explain the behavioral effects of tax enforcement?*

Andreas Kotsadam†, Knut Løyland‡, Oddbjørn Raaum§, Gaute Torsvik¶ and Arnstein Øvrum‖

Abstract

While audits have lasting effects on the subsequent reporting of audited taxpayers, the evidence on mechanisms is scarce. We compare the effectiveness of a correspondence audit and a letter encouraging tax filers to take a second look at their deductions. We find that both treatments lower tax deductions, also in the next year. A subsequent survey documents that the audit exposure raised the perceived risk of future audits, but not enough to explain the behavioral response to being audited. We conclude that the increase in future compliance is mainly due to other channels such as improved knowledge about tax rules.

*We thank seminar participants at several universities and conferences, Matteo Alpino, Annette Alstadsæter and Ole Rogeberg for useful comments and suggestions.

†Ragnar Frisch Centre for Economic Research, Oslo; e-mail: andreas.kotsadam@frisch.uio.no

‡The Norwegian Tax Administration

§Ragnar Frisch Centre for Economic Research

¶University of Oslo

‖The Norwegian Tax Administration

1 Introduction

Tax authorities interact with the public in different ways to increase compliance. While audits explicitly check if the tax filing is correct, softer, less intrusive and cheaper enforcement policies such as information campaigns, reminders, and encouragements intend to enhance compliance without control. In addition to detect and correct noncompliance on the spot, audits also have lasting effects on subsequent reporting among audited taxpayers (Kleven et al., 2011; Advani et al., 2021; DeBacker et al., 2018; Hebous et al., 2020; Løyland et al., 2019). The evidence for the softer policies is mixed, even in the short run (see Alm, 2019; Slemrod, 2019 and Pomeranz and Vila-Belda, 2019).

In this paper we compare immediate and subsequent effects of hard and soft enforcement policies by means of a randomized controlled trial. The RCT contains two alternative treatments; a desk based correspondence audit and a letter encouraging tax filers to take a second look at their itemized self-reported income tax deductions within a target population of 15 000 personal Norwegian taxpayers with relatively high self-reported income tax deductions. Both the audit and the letter treatment lowered self-reported deductions in the year of the intervention.[1] We also find that both policies reduced self-reported deductions in the subsequent tax year. The accumulated effect over the intervention year and the subsequent year is largest for audit. Although the letter is a cheaper enforcement policy, the audit generates more net tax revenue.

Our main contribution is to explicitly test a plausible mechanism for why individual enforcement policies have persistent effects on taxpayer reports. Specifically, we test the relevance of taxpayer updating within the standard Allingham and Sandmo (1972) framework of non-compliance by asking taxpayers - after the intervention - to assess the probability that they will be audited in the future. If taxpayers intentionally misreport, we expect the behavioral effects of enforcement treatments to arise from updated beliefs about the risk of future audits. We find that audit exposure raises the perceived risk of future audits, but not enough to explain the large drop in self-reported deductions in the following year. Tax payers who received the encouragement letter reported the same audit probability as the control group. To our knowledge, we are the first to combine a study of tax enforcement exposure with a survey to assess how the treatment affects individual taxpayer's perceptions of future audits.[2]

We also contribute to the literature on soft enforcements. Previous research has shown that infor-

[1] These short run effects were included in the pre-analysis plan, where we also specified how to investigate persistent behavioral compliance effects and heterogeneity, as well as the mechanisms behind the effects on future tax filings. The pre-analysis plan was registered at the AEA RCT Registry (#0004817) before we received the post treatment period data. The plan is added to Appendix Section F. All deviations from the plan are explicitly stated in the text.

[2] Bérgolo et al. (2021) test the effects of letters to firms and find that letters signaling the audit probability decreased the perceived risk of audit, yet the letters still decreased evasion.

mation making audits and the detection probability more salient often affect reporting in the short run (Slemrod et al., 2001; Kleven et al., 2011; Fellner et al., 2013; Bott et al., 2020), as does information about public disclosure and prison sentences (Holz et al., 2020). Many studies appealing to social norms or tax morality find limited effects on tax compliance (Hallsworth, 2014). While some recent papers report increased compliance (Hallsworth et al., 2017; Holz et al., 2020; Bott et al., 2020), others suggest that such appeals may even backfire (De Neve et al., 2021). Using survey data from Uruguay, Bérgolo et al. (2020) find that tax evasion is essentially uncorrelated with tax morale. In a meta-analysis of 40 studies, Antinyan and Asatryan (2020) find that non-deterrence nudges are ineffective whereas deterrence nudges seem to have effects.[3] Our results, that a letter simply encouraging tax filers to take a second look at their tax deductions affect misreporting, show that also non-deterrence interventions can have persistent effects on tax compliance.

Unlike most studies, we compare the effectiveness of alternative policies to improve tax compliance within the same population of personal taxpayers typically targeted by operational audits. Such an explicit comparison of alternative strategies is rare as previous studies typically focus on a single policy. Our combined findings suggest that improved knowledge about tax rules is an important mechanism for why exposure to individual tax enforcement policies have persistent compliance effects. The perceived risk is barely affected by the audit and audits in Norway do not generally come with substantial pecuniary punishments.[4] The letter improved compliance without affecting the perceived risk of future audits and there was no moral suasion in our letter.

A final contribution of the paper is that we use machine learning (ML) methods to detect heterogeneity in compliance effects. To our best knowledge, we are the first study testing heterogeneous audit effects with ML methods. De Neve et al. (2021) is the only tax evasion study using ML methods, but they identify effects of different letters. Tax administrations regularly use predictive models to identify high risk individuals, but they seldom target individuals based on their behavioral responses to tax enforcement policies. Robust evidence on heterogenous compliance effects is potentially useful for policies. If some taxpayers are particularly responsive to enforcement policies, tax revenues may increase if the authorities target these taxpayers.[5] We find, however, no heterogeneity in subsequent responses.

[3] They also document some worrying signs about the literature to date which suggest selective reporting of results, in particular they find that larger studies tend to have smaller effects and that marginally significant effects are more likely to be reported than marginally insignificant effects.

[4] Hebous et al. (2020) document that less than one percent of the audited Norwegian taxpayers in their sample received a penalty.

[5] Enforcement tagging based on compliance responses may, however, clash with ethical principles such as e.g. horizontal equity.

2 Institutional Background, Data and Experimental Design

2.1 Tax filing, timeline and experimental design

Our design is closely linked to the sequence of actions and the information exchange between the The Norwegian Tax Administration (NTA), third party institutions and personal taxpayers (See Appendix Table A.1 for a detailed timeline of tax returns). Employers report employee earnings to the NTA and withhold stipulated taxes. Other individual income sources, e.g. interests and financial capital gains, are also reported by third parties, as are some deductions, including interest paid on mortgages and donations to charitable organizations. Based on the third-party information, tax returns for year t are pre-filled and distributed by the NTA to taxpayers at the beginning of April in year t+1. Employees and pensioners can then make corrections to their tax returns and self-report income and/or deductions until April 30, while self-employed must file their personal tax report before the end of May.[6] Our study focuses on self-reported deductions. The most common self-reported income tax deduction items are interests on debt, personal work-related expenses on costs related to stays away from home, childcare deductions and expenses from lending out property (Løyland et al., 2019).

Tax audits are carried out during May–December year t+1. Since 2014, two main types of audits have been used to check the self-reported itemized tax deductions of personal taxpayers. First, a traditional targeted audit is based on computer-generated flags that pop up if there are irregularities on specific items. The second type of audit is based on a broader set of information where every taxpayer is given a risk-score based on individual characteristics, recent filing and historical records.[7]

Our study is based on two different treatments for the tax year 2017, where the NTA calculated a risk-score for all personal taxpayers after they had filed their report by the end of April 2018 (end of May for the self-employed). From this distribution, around 15 000 individuals with the highest score were selected to constitute the experiment population.

One third of the taxpayers was drawn for a standard low-cost office-based audit. The audit checked for suspicious itemized tax deductions and asked for documentation from the taxpayer if needed.[8] The

[6] Over the next months, and actually up to three years under the current tax law, personal taxpayers can reopen the file and adjust their reported items.

[7] In 2013, the NTA singled out 310 000 taxpayers claiming self-reported deductions above an (unofficial) threshold of Z Norwegian kroner on one or two items from a list of 29 specified expenses. A random sample was checked to train and test a gradient boosting machine learning algorithm to predict a binary classifier of compliance/noncompliance. In 2014-2016, the model provided a risk score for every taxpayer and those with a risk-score above a year-specific threshold were selected for audit, (Løyland et al., 2019).

[8] The NTA also run firm audits. Using data from randomly assigned on-site audits among 2 462 Norwegian firms, Bjørneby et al. (2018) provide evidence of collusive tax evasion whereby employers and employees collude to keep transactions off the books.

taxpayer was only notified if the auditor found irregularities with the claimed deductions or were asked for additional documentation. Hence, all taxpayers who had their deductions adjusted by the NTA knew they had been audited, but we do not have exact information on whether the compliant taxpayers knew they were audited. The auditors did not check other items such as income reporting.

Another third received the softer treatment; a letter encouraging them to reconsider and check their self-reported itemized deductions. The letter was sent to the taxpayers between May and October 2018. The letter (reproduced in Appendix Figure A.1) asks the taxpayers to take a second look at their itemized self-reported deductions and states that "Random checks performed by the NTA show that 6 out of 10 taxpayers in your situation make mistakes when claiming this kind of deductions". Finally, the letter reminds the taxpayers that documentation must be provided upon request. The letter was sent via an electronic personal information platform used by Norwegian authorities ("Altinn.no"). The taxpayers were notified once by an e-mail or SMS that there is a letter from the NTA in the personal inbox. While our letter had no moral suasion or explicit statement that should increase their perceived detection probability, it may still trigger changes in compliance related to both these mechanisms. Of course, the encouragement also induces taxpayers to take a second look and correct unintentional mistakes.

The final third had their tax reports checked by the standard procedures. As in most RCTs, some participants finally received a treatment different from the one that was assigned (cross-overs). In our case, since ordinary desk audits of suspicious filing carried on as usual irrespective of the RCT, a small minority of the control group and the letter group experienced a flag audit. The taxpayers were not given any information about the audit selection mechanism, and since they followed the same protocol we assume equal compliance effects of the two types of audits. We can therefore estimate the effect of audit using the RCT assignment as an instrument. Moreover, a small fraction of the taxpayers selected to the letter treatment did not actually receive the message. The policy relevant treatment is sending a letter (intention to treat), but the effectiveness of this enforcement depends on the extent to which the message is received and opened, which in turn is affected by delivering technology and individual effort. In Appendix Section B we discuss and present several ways of dealing with cross-overs, all of which lead to even larger behavioral effects of actually receiving the treatments.

2.2 Outcomes

2.2.1 Tax administration registers

The short run outcomes are deduction adjustments by the NTA (in the case of audit) or by the taxpayers themselves (in the case of letter). As described in the registered pre-analysis plan, compliance effects in the next year tax filing are measured by self-reported deductions as well as total claimed deductions. The total claimed deductions contain pre-filled and self-reported deductions and will therefore capture the behavioral effects of the enforcement treatments. In the absence of audits, claimed and final deductions are equal.[9]

2.2.2 Survey

To improve our understanding of the behavioral responses to the these tax enforcement interventions, we hired a Norwegian poll agency to conduct a phone survey. To avoid any impact of the survey itself on the taxpayers' reports, the survey took place just *after* the filing of the tax information for 2019, one year after the treatment. The survey focused on the taxpayers perceived probability of being audited in the future, an empirical equivalent to the detection probability in the Allingham Sandmo tradition. Our main question of interest is whether the audit and the letter affected the Perceived Probability of Future Audit (PPFA)[10] based on the following question: "What do you think the probability is that the tax authorities will control your reported taxes in 2019?". The answer categories range from 1 "Not likely at all (0 percent)" to 7 "Certainly (100 percent)" and we retain the continuous coding of this variable (1-7). The respondents were told that the survey was conducted on behalf of NTA, but that answers are anonymous and that the survey would take around 3.5 minutes. We tried to contact all individuals in the main sample but were not able to obtain the phone numbers for everyone and not everyone responded. Attrition is not correlated with treatment and is discussed in Appendix Section C.

[9] We cannot use final total deductions as an outcome since the field experiment was implemented in such a way that the group that received the letter treatment in May/June 2018 were audited later on. Since these audits were done in July-October 2019 (unexpected and after the taxpayer filing) they did not affect self-reported deductions.

[10] In the pre-analysis plan we called this "Subjective Detection Risk (SDR)", but this label is imprecise.

3 Empirical specifications and hypotheses

3.1 Compliance effects

To find the average effects of the two tax enforcement policies we estimate the following regression using ordinary least squares:

$$Y_{i,t} = a + b_t Letter_{i,t_0} + c_t Audit_{i,t_0} + u_{i,t} \tag{1}$$

We estimate the effects on deductions in the year of the interventions ($tax\ year = t = t_0 = 2017$) and the year after ($t = t_0 + 1$).

Taxpayers had already filed their tax report for the income year 2017 when exposed to either the letter or the audit treatment. Both treatments can, however, alter the final deductions for year t_0. While the letter encouraged the taxpayer to reopen their report and adjust self-reported tax deductions, the audit would lead to an adjustment by the NTA in case of any irregularity. We expect this short run adjustment effect to be largest for the audit ($c_{t_0} < b_{t_0} < 0$). There are reasons to be genuinely uncertain about the short-term effect of the letter. We know that information hinting at increased deterrence tend to have stronger effect on tax compliance than letters appealing to tax morality or civic duty (Slemrod, 2019). The letter is fairly neutral, there is no explicit mention of injunctive or descriptive norms, and it does not openly threaten that the itemized deductions will be audited unless action is taken. Notwithstanding, those who receive the letter will now be aware that they are on the radar of the tax authorities. Furthermore, among the taxpayers who want to pay their due taxes it is reasonable to assume that a fraction have mistakenly filed too high deductions, and the letter encourages them to check the rules more thoroughly. Hence, we expect that a fraction of those with letter treatment will reopen their files and self-adjust their report. But as long as far from every taxpayer with irregular tax deductions self-adjust, the average after adjustment deductions will be lower than among the audited taxpayers.

With respect to future compliance effects, existing empirical evidence makes us expect that audit exposure leads to lower self-reported itemized tax deductions the year following the treatment ($c_{t_0+1} < 0$). There are, however, potential mechanisms that may contribute to higher deductions. First, some lab evidence suggests that an audit today can reduce the perceived future audit probability (a "bomb-crater effect") (Mittone et al., 2017). Moreover, even if the risk of being audited is adjusted upwards, the assessed probability that non-compliance will be detected may go down for those audited without

consequence (Gemmell and Ratto, 2012; Mittone et al., 2017). Finally, the audits can also lower future compliance by showcasing that the penalties for non-compliance are low. For the letter, we see no role for these mechanisms and expect increased compliance ($b_{t_0+1} < 0$). It is not clear a priori which wpolicy will have the largest effect. The audit adjustment is more intrusive and forceful, but some of the audited will not know that the tax administration has checked their files because they were not adjusted or asked for documentation. In contrast, the letter was sent to everyone assigned to this treatment.

3.2 Effects on perceived probability of future audit

One reason why an enforcement intervention at t_0 may have an effect on tax filings at t_0+1 is because it affects the perceived probability of being audited in the future. To test this mechanism we study the core question in our survey related to the Perceived Probability of Future Audit (*PPFA*); "What do you think the probability is that the tax authorities will control your reported taxes in 2019?" and estimate the following OLS regression:

$$PPFA_i = e + dLetter_i + fAudit_i + \epsilon_i \tag{2}$$

We expect that $d \geq 0$ since we cannot think of any reasons why a letter with this content should induce recipients to lower their perceived probability of an audit. For the effect of audit, the sign of f is a-priori uncertain. While most studies of long-term compliance effects of audits lead us to expect $f > 0$, some argue that being audited today may reduce the perceived probability being picked out in future audits (Mittone et al., 2017).

4 Estimated treatment effects

4.1 Short run effects

In Panel A of Table 1, we show that the pre-treatment outcomes are balanced across treatments. The short run average treatment effects are reported in the two last columns of Panel B as entries from separate linear OLS regressions controlling for pre-treatment claimed deductions (row c) and reports the treatment dummy with standard error in parenthesis.

The audits disclosed extensive illegitimate deductions as two in three taxpayers (65.3%) had their

report adjusted by the auditor. Among the taxpayers without a treatment, only 6% were adjusted via the ordinary flag based audit. Hence, the effect of the audit treatment was to increase the fraction who had their deductions adjusted by 59 percentage points. The average audit adjustment -29 538 NOK (1 USD= 8.3 NOK in 2017), or 43% of the average self-reported deductions. Among those adjusted, on average 50 064 NOK of the deductions was not approved by the NTA.

Turning to the letter treatment, 11% reopened their files and lowered their self-reported deductions. As close to none did so in the no treatment group, the short run letter effect is estimated to -0.105 and highly significant. The self-adjustment effect of the letter is -3 584 NOK. The self-adjustment among those who responded (compliers) was nearly ten times larger and estimated to -34 133 NOK. However, the letter effect on the final total deductions is smaller; -2 503 NOK. This is because 6% of the letter and control group had their deductions adjusted by the NTA through flag audit. Thus, a substantial part of the mistakes corrected by the self-adjustment would have been discovered by the standard procedures. For final total deductions, the audit adjustment by far exceeds that the self-adjustment from the letter (-30 159 NOK vs -2 503 NOK). As expected, far from every taxpayer respond to soft measures that encourage them to check their reports and follow the rules. When we combine an audit hit rate of 0.653 with a self-adjustment share of 0.105, the evidence suggests that about one in six taxpayers who had made a mistake did respond to the letter.

4.2 The effect on future tax compliance

While only the letter allows for any behavioral responses in the year of the treatment, both the letter and the audit potentially affect future compliance. In Panel C of Table 1 we report deductions for the following tax year, reported about ten months after the treatments. First, the pre-filled deductions from third parties are slightly lower for the two treatment groups, but there are no significant differences compared to the no treatment group. Turning to the treatment effects on taxpayer's self-reported deductions, we see that both interventions significantly affected the extensive margin and reduced the fraction with self-reported deductions. The audit lowered the share with non-trivial self-reported deductions by 12 percentage points (pp). The letter effect is lower (5 pp), but statistically and economically significant.

The effects on self-reported deductions are also significant.[11] The audit effect is -10 128 NOK, or 29% of the self-reported deductions in the no treatment group. The letter effect is smaller (- 3 900

[11] The self-reported deductions are clearly lower than in the previous year for all three groups. This mean-reversion reminds us that treatment effects are hard to identify from operational audits triggered by "suspicious" reporting.

NOK), but also statistically significant. Even if the future compliance effect is considerably larger for audits, they are more similar than the short-run adjustment effects.

Taxpayers with a spouse will typically not make filing decisions in isolation. Some deductions are household specific and can potentially be transferred from one spouse to the other as a response to the treatment. Spouses may also update their knowledge about tax rules, and audit probabilities, when their partner has been subjected to the audit or the letter. Both mechanisms suggest that spousal reporting is a part of the future compliance effects. Estimates based on household outcomes are very similar to the individual effects reported in Panel C of Table 1. Our calculations in Appendix section E show that, despite the letter being substantially cheaper, the audit generates approximately five times the net tax revenue.

One might ask to what extent the large and persistent response to the letter is explained only by those who responded in the short run and whether there are delayed behavioral responses even among those who did not self-adjust directly. A persistent reduction is expected if the taxpayer found the deductions to be illegitimate, or decided not to report excessively after receiving the letter. It is also possible that some taxpayers prefer to wait until next year to make a correction as they think it looks suspicious to respond directly, or if there are other fixed costs associated with opening up and correct an already filled in deduction. The question of persistence also exists for audits. As a complementary analysis, not specified in the pre-plan, we look at differential responses in year t_0+1 for those who had an adjustment and those who did not. We restrict the samples to include the respective treatment sample and the pure control group. In equation (3), we let Treatment be either Audit or Letter

$$Y_{i,t_0+1} = \alpha + \beta_t Treatment_{i,t_0} + \gamma_t Treatment * Short\,run\;adjustment\,dummy_{i,t_0} + \rho Y_{i,t_0} + \eta_{i,t_0+1} \quad (3)$$

where Y_{i,t_0+1} will be the self-reported deductions (in NOK) or an extensive margin dummy for having at least 1 000 NOK in self-reported deductions. The $Short\ run\ \ adjustment\ dummy_{i,t_0}$ is equal to one for individuals with a self adjustment as direct response to the letter treatment. For audits, the $Short\ run\ \ adjustment\ dummy_{i,t_0}$ is equal to one if the tax administration made an adjustment as part of the audit. To account for selective adjustment we also include pre-treatment self-reported deductions (Y_{i,t_0}).[12]

[12]Note that in the analysis of the letter treatment, any main effect for the $Short\ run\ \ adjustment\ dummy_{i,t_0}$ is not defined due to collinearity with treatment (i.e. there is no self-adjustment in the control group). In the audit treatment some individuals in the control group are audited (flag audits) and the results are similar if we include a full set of interactions.

In Table 2 we report the OLS regression coefficients. Panel A shows the differentials for Letter and we see in column (1) that both groups report significantly lower deductions in year t_0+1 than the control group. The drop in self-reported deductions is, however, much larger for those who responded the first year.

In interpreting this heterogeneity it is important to note that we condition on an endogenous outcome. For the letter, we find no evidence of pre-treament differences since the self reported deductions do not differ by self-adjustment in the intervention year (column 3). In columns (4) and (5), we control for the continuous pre-intervention values of the self-reported deductions and the significant responses for both groups remain. Even if pre-treatment deductions are balanced, there might be other unobserved characteristics that may bias the differential response by short run adjustment.

In Panel B, we present the conditional effects for the audit treatment. We find lower deductions in both audit groups in year t_0+1 (column 1) compared to the control group, but the probability of self-reported deductions is not reduced for those without audit adjustment in the treatment year (column 2). Column (3) of Table 2 shows that the pre-treatment self-reported deductions are higher among those who were adjusted in the subsequent audit than among non-adjusted taxpayers. The non-adjusted taxpayers constitute a selected group of taxpayers, who, based on their lower pre-treatment level of self-reported deductions will tend to have lower self-reported deductions in the post-audit year as well. When we control for pre-treatment self-reported deductions, we find no difference between the non-adjusted and the control group (columns 4 and 5). For the taxpayers who did not know they were audited, we expect no behavioral responses and there appears to be zero effect even for those without adjustment who knew they were audited.

This exercise indicates that the letter had a (delayed) effect also on those who did not adjust their filing just after receiving it. The audit effect in contrast, seems to be driven entirely by those who had their filing adjusted by the tax administration. However, since adjustment is not random we cannot rule out that unobserved factors also influence these conditional means.

4.3 Heterogeneity

We use the "Generic ML" approach method by Chernozhukov et al. (2018) to test for heterogeneous treatment effects (see Appendix Section D for details). Table 3 presents the conditional average treatment effect and the heterogeneity parameter (HET) from preferred methods for the different samples. Together they form a weighted linear prediction of the outcomes. By separating their influence

we get a test of the heterogeneity in the data. The significant heterogeneity parameters for short run effects suggest that the adjustments effects differ across taxpayers. Those who respond strongest to the audit and the letter interventions are individuals with high risk-scores and high prior deductions (Appendix Section D). For the letter we also find larger self-adjustment among those born in Norway. The audit adjustment is highest among the younger, labor immigrants, single, and men. In contrast to the adjustment heterogeneity, we find no support for differential persistent compliance effects as the heterogeneity parameters for the subsequent tax year are very close to zero. For enforcement policies, it appears sufficient for the Norwegian Tax Authorities to target individuals based on short run effects, although the precision can be improved by adding more characteristics in the model that calculates the risk-score.

5 Effects on perceived probability of future audit

We show the distribution of survey responses to the question "What do you think the probability is that the tax authorities will control your reported taxes in 2019?" in Figure 1. The distribution of the subjective audit risk for Audit (left panel) is tilted to the right of the no treatment distribution, but there is also a slightly larger share answering 1 "Not likely at all (0 percent)". The distributions for Letter and no treatment are very similar. Table 4 shows the OLS estimates. The audit effect is positive, but small, corresponding to about 5 percent of the no treatment mean, and only statistically significant at the 10 percent level. Using an equivalence testing approach of two one-sided t-tests (TOST), and a 5 percent significance level, we can reject that the effect of audit is larger than 0.42, or 10 percent of the mean. We find no indications that the letter affected the perceived probability of future audit in any way. The results are similar if we include controls (second column).

6 Conclusion

In a large scale field experiment we find that two alternative enforcement policies affect the filing of self-reported deductions by personal taxpayers, both in the year of intervention and the subsequent tax year. The effect of a desk based correspondence audit is larger than the effect of a letter encouraging tax filers to take a second look at their tax deductions. Even if audits are more costly, they generate higher net tax revenue than sending a letter.

The reasons why exposure to tax enforcements such as audits have persistent effects on self-reporting

are not obvious. From a deterrence perspective, audits will have persistent compliance effects if they make people believe that future audits are more likely. In a survey of the treated taxpayers, we find indications that exposure to audit raises the perceived probability of future audit, but this can only explain a minor part of the behavioral response to the audit. We find no letter effect on the perceived audit probability, even if the encouragement to take a second look actually lowers taxpayers' self-reported deductions. Alternative mechanisms for compliance effects of individual tax enforcement measures include improved knowledge about tax rules and punishments, or more salient and effective norms. Improved knowledge about tax rules appears to be a plausible mechanism since the letter contained no moral suasion and because correspondence audits of personal taxpayers in Norway do not generally come with substantial pecuniary punishments. When taxpayers do not spend sufficient time to study detailed tax rules, tax compliance can be substantially improved by providing better and simpler information that make it easier to for individuals to report correctly. In this light we find the immediate behavioral response to the letter interesting since the encouragement to have a second look was sent after the first filing of deductions by the taxpayer.

With increased digitized tax reporting, there are numerous ways for the tax authorities to communicate information and encouragements to the taxpayers based on their reporting behavior. In recent years, tax administrations have introduced predictive models to target enforcement towards those with the riskiest profiles based on short run effects (OECD, 2017). From a tax revenue perspective, both short and long run effects should be included. Our heterogeneity analyses show that targeting based on short run adjustments cannot be improved by including future behavioral responses. Whether this also holds for other outcomes and countries needs to be investigated in future work.

References

Advani, A., W. Elming, J. Shaw, et al. (2021). The dynamic effects of tax audits. Technical report, Institute for Fiscal Studies.

Allingham, M. G. and A. Sandmo (1972). Income tax evasion: A theoretical analysis. *Journal of public economics 1*(3-4), 323–338.

Alm, J. (2019). What motivates tax compliance? *Journal of Economic Surveys 33*(2), 353–388.

Antinyan, A. and Z. Asatryan (2020). Nudging for tax compliance: A meta-analysis.

Bérgolo, M., M. Leites, R. Perez-Truglia, and M. Strehl (2020). What makes a tax evader? *NBER Working Paper* (w28235).

Bérgolo, M. L., R. Ceni, G. Cruces, M. Giaccobasso, and R. Perez-Truglia (2021). Tax audits as scarecrows: Evidence from a large-scale field experiment. Technical report, National Bureau of Economic Research.

Bjørneby, M., A. Alstadsæter, and T. Kjetil (2018). Collusive tax evasion by employers and employees: Evidence from a randomized field experiment in norway. *CESifo Working Paper No. 7381*.

Bott, K. M., A. W. Cappelen, E. Ø. Sørensen, and B. Tungodden (2020). You've got mail: A randomized field experiment on tax evasion. *Management Science 66*(7), 2801–2819.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernandez-Val (2018). Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research.

De Neve, J.-E., C. Imbert, J. Spinnewijn, T. Tsankova, and M. Luts (2021). How to improve tax compliance? evidence from population-wide experiments in belgium. *Journal of Political Economy*.

DeBacker, J., B. T. Heim, A. Tran, and A. Yuskavage (2015). Once bitten, twice shy? the lasting impact of irs audits on individual tax reporting. *Journal of Financial Economics 117*(1), 122–138.

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), 1–35.

Fellner, G., R. Sausgruber, and C. Traxler (2013). Testing enforcement strategies in the field: Threat, moral appeal and social information. *Journal of the European Economic Association 11*(3), 634–660.

Gemmell, N. and M. Ratto (2012). Behavioral responses to taxpayer audits: evidence from random taxpayer inquiries. *National Tax Journal 65*(1), 33.

Hallsworth, M. (2014). The use of field experiments to increase tax compliance. *Oxford Review of Economic Policy 30*(4), 658–679.

Hallsworth, M., J. A. List, R. D. Metcalfe, and I. Vlaev (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of public economics 148*, 14–31.

Hebous, S., Z. Jia, K. Løyland, T. O. Thoresen, and A. Øvrum (2020). Do audits improve future tax compliance in the absence of penalties? evidence from random audits in norway.

Holz, J. E., J. A. List, A. Zentner, M. Cardoza, and J. Zentner (2020). The 100 million nudge: Increasing tax compliance of businesses and the self-employed using a natural field experiment. Technical report, National Bureau of Economic Research.

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*(3), 651–692.

Løyland, K., O. Raaum, G. Torsvik, and A. Øvrum (2019). Compliance effects of risk-based tax audits.

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.

OECD (2017). *The Changing Tax Compliance Environment and the Role of Audit*.

Pomeranz, D. and J. Vila-Belda (2019). Taking state-capacity research to the field: Insights from collaborations with tax authorities. *Annual Review of Economics 11*, 755–781.

Slemrod, J. (2019). Tax compliance and enforcement. Technical Report 4.

Slemrod, J., M. Blumenthal, and C. Christian (2001). Taxpayer response to an increased probability of audit: evidence from a controlled experiment in minnesota. *Journal of public economics 79*(3), 455–483.

Table 1. Treatment effects.

	Audit	Letter	No	Audit	Letter
		Means		Equation (1) coefficients	
Panel A. Pre-treatment balance					
Pre-filled deductions	129 270	127 569	128 310	960	--811
				(1 083)	(1 072)
Self-reported deductions	66 623	68 080	68 365	-1 741	-284
				(938)	(950)
Claimed deductions	195 893	195 649	196 674	-731	- 1 025
				(1 324)	(1 039)
Panel B. Short run					
Share with self-adjustment	0.014	0.110	0.005		0.105***
					(0.00)
Self-adjustment	-135	- 3 084	474		-3 584***
					(293)
Share audit adjusted	0.653	0.062	0.061	0.59***	
				(0.00)	
Audit adjustment	-34 071	-3 845	-4 674	- 29 538***	
				(731)	
Final total deductions	161 687	189 079	192 474	- 30 159***	-2 503***
				(738)	(573)
Panel C. Future compliance					
Pre-filled deductions	136 754	136 431	137 600	-846	-1 169
				(1 119)	(1 108)
Share with self-reported ded. $>$ 1 000 NOK	0.539	0.610	0.660	- 0.12***	- 0.05***
				(0.01)	(0.01)
Self-reported deductions	25 155	31 382	35 283	-10 093***	-3 900***
				(898)	(922)
Claimed deductions	161 909	167 814	172 883	-10 825***	-5 070***
				(1 351)	(1 360)
Sample sizes					
Panel A	4 151	4 130	4 178	8 329	8 308
Panel B	3 918	3 890	3 962	7 880	7 852

Note: All numbers are in NOK, except for fractions. Panel A includes pre-treatment items. Panel B has outcomes for the tax year 2017 and Panel C has outcomes for year 2018. In columns four and five, each entry is from a separate linear OLS regression and reports the treatment dummy estimate with standard error in parenthesis. In Panel B and C, total pre-treatment deductions is included as control in the regressions (based on pre-plan Table 4). *** : significance at the 1%-level.

Table 2. Behavioral responses by short run compliance

Period	Self-reported deductions t_0+1		t_0	t_0+1	
	Continuous	Extensive margin	Pre-treatment	Continuous	Extensive margin
Panel A. Letter treatment					
Treatment (Letter)	-2 917***	-0.0433***	68	-2 940***	-0.0434***
	(951)	(0.011)	(1 008)	(887)	(0.011)
Treatment*Self Adjustment t_0	-8 881***	-0.0624**	-2 871	-7 910***	-0.0592**
	(2 086)	(0.025)	(2 212)	(1 948)	(0.024)
Self reported deductions t_0				0.338***	0.000001***
				(0.010)	(0.000)
Constant	35 283***	0.660***	68 204***	12 205***	0.585***
	(649)	(0.025)	(688)	(909)	(0.011)
Observations	7 852	7 852	7 852	7 852	7 852
Panel B. Audit treatment					
Treatment (Audit)	-3 009**	-0.0173	-6 675***	-843	-0.0096
	(1 246)	(0.015)	(1 339)	(1 170)	(0.015)
Treatment*Audit Adjustment t_0	-10 921***	-0.159***	7 650***	-13 403***	-0.168***
	(1 332)	(0.016)	(1 431)	(1 250)	(0.016)
Self reported deductions t_0				0.325***	0.000001***
				(0.010)	(0.000)
Constant	35 283***	0.660***	68 204***	13 146***	0.581***
	(631)	(0.008)	(677)	(893)	(0.0112)
Observations	7 880	7 880	7 880	7 880	7 880

Table 3. Average treatment effects and test of the degree of heterogeneous treatment effects.

	Audit		Letter	
	Short run	Future compliance	Short run	Future compliance
ATE	- 29 684	- 7 597	- 3 834	- 3 949
	(-31 631,-27 731)	(-10 516,-4 678)	(-4 736,-2 944)	(-6 947,-985)
	[0.000]	[0.000]	[0.000]	[0.018]
HET	0.851	0.002	0.552	-0.003
	(0.728, 0.975)	(-0.136, 0.138)	(0.283, 0.815)	(-0.147, 0.145)
	[0.000]	[1.000]	[0.000]	[1.000]
Method	Elastic Net	Neural Net	Elastic Net	Neural Net

Note: ATE refers to the average treatment effects. HET is the heterogeneity parameter. All numbers are medians over 100 splits. 90 percent confidence interval in parenthesis. P-values for the hypothesis that the parameter is equal to zero in brackets.

Figure 1. Perceived Probability of Future Audit (PPFA) by treatment.

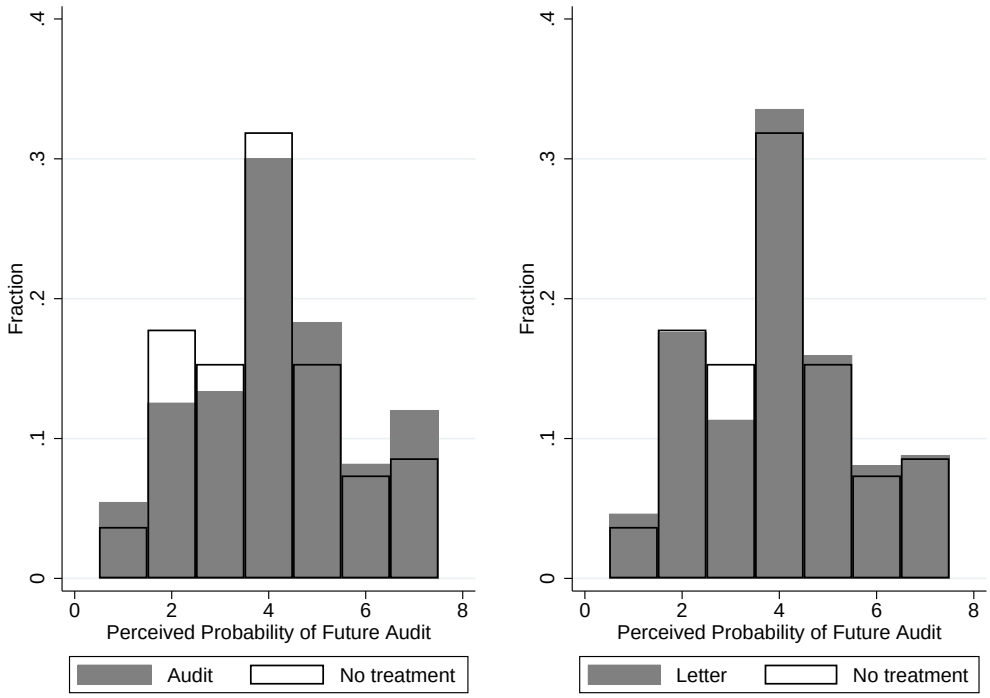


Table 4. Effects of audit and letter on perceived probability of future audit.

	(1)	(2)
Audit (f)	0.221*	0.227*
	(0.120)	(0.120)
Letter (d)	0.070	0.0538
	(0.114)	(0.115)
Female		-0.0705
		(0.106)
Age 30-39		0.028
		(0.149)
Age 40-49		0.144
		(0.144)
Age 50-59		0.426***
		(0.142)
Age 60+		0.358*
		(0.205)
Self-employed		-0.086
		(0.115)
Risk-score level 2		-0.288
		(0.194)
Risk-score level 3		-0.404**
		(0.149)
Risk-score level 4		-0.283
		(0.144)
Risk-score level 5		-0.112
		(0.271)
Mean PPFA in no treatment group	3.94	3.94
R squared	0.003	0.019
Sample size	1 121	1 121

Note: Dependent variable is Perceived Probability of Future Audit (PPFA). ***/**/* indicate significance at the 1%/5%/10% level.

A Tables and figures referred to in the text

Table A.1. Timeline 2018. Personal Tax Returns for Tax Year 2017.

	Standard procedures (business as usual)	**Actors**	**Field Experiment Treatments**	**Outcome short run**
January-February	Third party reporting	Employers and Financial Institutions		Income, interests, wealth
March	Pre-filled tax returns distributed	Norwegian Tax Administration (NTA)		Income by source, deductions, gross wealth, debt
April	Check, correct and self-report if relevant	Employees and pensioners		Acceptance of pre-filled or ***self-reported*** deductions and income
May	Check, correct and self-report if relevant	Self-employed		Acceptance of pre-filled or ***self-reported*** deductions and income
May-October	Programmed audit routines (flags)	NTA to taxpayers	**Letter (L=1)**	***Self-adjustment*** by taxpayers
	Programmed audit routines (flags)	NTA	**Audit (A=1)**	Approval or ***audit-adjustment*** by the NTA
	Programmed audit routines (flags)	NTA	**Non-treatment (A=L=0)**	Approval or audit-adjustment by the NTA
October-December	Final assessment	NTA		***Final total deductions***, taxable income and wealth

Skatteetaten

Return address:
P O Box 6499 Etterstad, N-0606 OSLO

Our date | Your date | Executive officer

Your reference | Telephone

Our reference | Postal address

Name
Address
Postcode and town

Check your deductions

In your tax return for the 2017 income year, you have claimed a deduction under item

- x.x.x. (item name/deduction)
-

From the 2016 income year, you're required to assess the basis for your tax calculation yourself. Random checks performed by the Norwegian Tax Administration show that 6 out of 10 taxpayers in your situation make mistakes when claiming this kind of deduction.

What do I have to do?

You have to check that you've given us the right information, and that you can document the deductions you've claimed in your tax return.

You have to enter the correct information in your tax return. If you find errors, correct them and submit your tax return at https://www.skatteetaten.no/en/person.

If we ask you to, you must be able to show documentation (receipts etc.) for the deductions in your tax return.

Please disregard this letter if the amount entered in your tax return is correct.

Do you have any questions?

You can read more about the items at https://www.skatteetaten.no/en/person/taxes/tax-return/find-item/. Call us on 800 80 000 if you have any questions.

Yours sincerely,

Figure A.1. Letter to taxpayer. Check deductions

B Accounting for treatment cross overs

As in most RCTs, some participants received a treatment different from the one they were randomly assigned to.[13] First, a small minority of the no treatment group experienced a flag audit. The flag audits follow the same protocol. Since the taxpayers did not know why they were selected (risk-score threshold or single items with a flag), we assume they have equal behavioral effects on future compliance and use assigned to audit as an instrument for any audit. This basically scales the effect of audit in Table 1 by the the inverse of the increase in the share with audit due to the random assignment. About 22% of the non-treatment group were audited due to flags. Therefore, the IV estimate of actual audit in Table A.2 is somewhat larger (in absolute numbers) than the effect of assigned audit in Table 1.

Table A.2. IV estimates accounting for cross overs.

	Audit	Letter (sent)	Letter (sent and opened)
Short run effects **(Panel A of Table 1):**			
Self-adjustment (NOK)	not relevant	-3 751***	-3 871***
		(306)	(316)
Audit adjustment (NOK)	-37 750***	not relevant	not relevant
	(965)		
Sample size	8 329	8 308	8 308
Compliance effects **(Panel B of Table 1):**			
Self-reported deductions (NOK)	-12 949***	-3 918***	-4 037***
	(1 140)	(932)	(962)
Claimed deductions(NOK)	-13 887***	-4 760***	-4 905***
	(1 396)	(1 137)	(1 172)
Sample sizes	7 880	7 852	7 852

Second, the letter was not sent to all assigned taxpayers due to some failing administrative procedures. For about 3.7%, no letter was sent. We can use estimate the effect of letter sent by instrumenting with assigned letter and find that the compliance effect estimate increases slightly. The effect of letter sent does also depend on the share of taxpayers who actually gets the message. From a behavioral insight perspective we would like to know the average effect of receiving the message of the letter. Again, we can use letter assignment as an instrument for sent and opened. Since 93.4% received the message, this effect is slightly larger than the intention-to-treat estimate in Table 1.

[13] This is typically called non-compliance, but this label has another meaning in this paper.

C Samples, pre-treatment balance and attrition

To analyze the short run treatment effects we use data from the initially submitted Spring 2018 tax return for the income year 2017, and any self-adjustments by the taxpayer or audit-adjustments by the NTA after the initial submission. Data on the initial submission of tax returns for the income year 2017 were extracted from the data warehouse of the NTA as of May 17 2018. The gross sample counts 14 902 with an equal share for the treatment groups (Table A.3). Data were missing for 826 taxpayers because they had not yet submitted their tax return (mainly self-employed taxpayers). Tax return data were also missing or incomplete for another 1 108 taxpayers at this date. For these taxpayers, the main reason for missing data was that they had submitted their tax return on paper (non-electronically). In such cases, tax return data are entered manually into the tax systems by the NTA, and this was done after the extraction date (May 17 2018). Due to the overwriting of previous versions of tax returns in the data warehouse, initial tax return data on these 1 934 taxpayers are not available. Taxpayer filing data contain outliers stemming from different sources of measurement error and/or extreme random numbers. Previous studies have used different strategies to deal with these problems. While DeBacker et al. (2015) winsorize at 90 percent, Kleven et al. (2011) trim income changes (post treatment) at −200 000 and +200 000 kroner "to get rid of extreme observations that make estimates imprecise" and Advani et al. (2021) "trim the top 1 percent to avoid outliers having an undue impact on the results". We are following the practice of trimming, and exclude taxpayers with values above the 99th percentile for one or more of the four variables; pre-filled deductions, pre-treatment claimed deductions, post-treatment claimed deductions and post-treatment final deductions. Our net sample for the analyses of short run treatment effects (and future compliance effects) therefore consists of 12 459 taxpayers.

Table A.3. Gross and net samples, attrition and sample exclusion across groups.

Criteria	Observations				Comments
	Audit	Letter	No	All	
Gross sample	4 964	4 945	4 993	14 902	
- Missing data due to late subs	285	285	256	826	Mostly self-employed
- Missing other reasons	376	366	366	1108	Manual submission, delayed handling by NTA
- Trimming	151	164	193	409	Taxpayers with deductions (4 items) above p99
Short run effect sample	4 151	4 130	4 178	12 459	Table 1
-No info 2018	52	49	43	144	Not present in the NTA tax liability register in 2019
-Technical attrition	104	109	94	307	System changes in the NTA
-Trimming	77	82	79	238	Taxpayers with deductions above p99
Compliance effect sample	3 918	3 890	3 962	11 770	Table 1
Gross survey sample	2 773	2 779	2 817	8 369	With identified telephone number
-Refusal	737	669	707	2 113	Refused to answer the survey
-Timout callback	1 356	1 014	1 1 418	3 788	Timeout callback
other reasons	270	607	257	1 134	Other reasons
Interviewed	410	489	435	1334	Individuals that were interviewed
- Missing info on characteristics	58	74	81	213	
Final survey sample	352	415	354	1 121	Table 4

Note: The short run effect sample and the sample exclusions follow the pre-plan.

We test for balance on a set of pre-treatment values of relevant variables. The results for the short run effects sample are shown in Table A.4. We test whether there is a difference between the audit group and the control group (Non-treat) and then between the letter group and the control group. First we test each variable separately and then we conduct an F-test of whether the variables jointly can predict treatment status. The F-tests are passed for both treatments and the individual variables seem balanced across groups. In the F-test we code eventual missing observations as zero and include a dummy variable for missing status in order not to lose observations.

Turning to the compliance effects sample in the year following the treatments, all analytic choices follow our pre-plan unless otherwise stated. There will be missing values and attrition due to trimming, deaths, and migration. In particular, there are a large number of migrants in our initial sample and many of them are likely to have left Norway. While the decision to stay may be affected by the treatment we do not view this as very likely. We tested and rejected that there is a difference between the groups in the probability of being present in the tax register. We further follow the same trimming practice for restricting the gross sample as above.

Finally, with respect to the PPFA mechanism, the sample is further reduced by non-response to the survey. For that sample we test attrition in the following way:

$$InPPFASample = g + hLetter_i + kAudit_i + u_i \tag{4}$$

Where *InPPFASample* is a binary variable for being in the survey sample with valid information on the question about perceived audit probability. There coefficients for h *or* k are not statistically significant.

Table A.4. Short run effects sample. Balance across treatment assignments. Means, standard deviation and tests.

Characteristic	Audit	Letter	No	Audit vs No	Letter vs No
	Mean (std dev)			t-test p-value	
Women	0.27	0.28	0.27	0.819	0.374
Age	39.8	39.9	39.9	0.921	0.696
	(10.6)	(10.7)	(10.6)		
Married	0.27	0.26	0.26	0.612	0.807
Norwegian citizen	0.62	0.52	0.61	0.595	0.476
Self employed	0.04	0.04	0.04	0.770	0.897
Risk-score 2017	0.697	0.694	0.694	0.078	0.972
	(0.079)	(0.0797)	(0.079)		
Final total deductions 2015	140 636	140 460	140 742	0.949	0.863
	(70 582)	(69 416)	(71 413)		
Final total deductions 2016	144 398	144 324	144 269	0.929	0.969
	(62 833)	(62 420)	(63 125)		
Pre-treatment deduction 2017	195 893	195 649	196 674	0.555	0.437
	(60 878)	(60 282)	(59 961)		
Observations	4 151	4 130	4 178		
All variables				F-test	F-test
				F(9, 8 316) = 0.48	F(9, 8 294) = 0.26
				p-value = 0.88	p-value = 0.98

Note: From pre-plan Table 3.

Table A.5. Compliance effects sample. Balance across treatment assignments. Means, standard deviation and test.

Characteristic	Audit	Letter	No	Audit vs No	Letter vs No
	Mean (std dev)			t-test p-value	
Women	0.276	0.280	0.273	0.741	0.450
Age	40.0	40.0	39.8	0.587	0.424
	(10.7)	(10.7)	(10.5)		
Married	0.272	0.269	0.265	0.481	0.653
Norwegian citizen	0.617	0.618	0.608	0.406	0.377
Self employed	0.037	0.037	0.037	0.888	0.842
Risk-score 2017	0.697	0.694	0.694	0.168	0.931
	(0.079)	(0.077)	(0.080)		
Final total deductions 2015	140 667	139 668	140 029	0.701	0.826
	(69 026)	(67 660)	(70 183)		
Final total deductions 2016	144 439	144 211	144 094	0.810	0.935
	(61 008)	(60 892)	(61 853)		
Pre-treatment deduction 2017	195 107	194 561	195 355	0.851	0.548
	(59 320)	(58 720)	(58 331)		
Observations	3 918	3 890	3 962		
All variables				F-test	F-test
				F(9, 7 867) = 0.48	F(9, 7 839) = 0.23
				p-value = 0.89	p-value = 0.99

Table A.6. Survey sample. Balance across treatment assignments. Means, standard deviation and tests.

Characteristic	Audit	Letter	No	Audit vs No	Letter vs No
	Mean (std dev)			t-test p-value	
Women	0.281	0.304	0.274	0.720	0.390
Age 30-39	0.223	0.195	0.280	0.114	0.127
Age 40-49	0.287	0.270	0.220	0.608	0.296
Age 50-59	0.225	0.289	0.249	0.265	0.400
Age 60+	0.074	0.080	0.088	0.292	0.668
Self-employed	0.162	0.145	0.180	0.549	0.230
Risk-score level 2	0.349	0.318	0.359	0.614	0.683
Risk-score level 3	0.261	0.277	0.282	0.435	0.810
Risk-score level 4	0.241	0.272	0.201	0.994	0.410
Risk-score level 5	0.080	0.070	0.093	0.421	0.478
Observations	352	415	354	706	769
All variables				F-test	F-test
				F(10, 695) = 1.00	F(10, 758) = 1.86
				p-value =0.4453	p-value = 0.0482

D Using machine learning to explore heterogeneous treatment effects

We pre-specified that we would use machine learning techniques with all control variables to automate the search for heterogenous treatment effects. In particular, we thought we would be using the random causal forests (R package grf, Wager and Athey, 2017). We also wrote: "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." We have now decided to use a newer method by Chernozhukov et al. (2018), the "generic ML" approach. This method has several advantages. First of all, it uses several machine learning methods in addition to random forests 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 Y, 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 can then measure the average treatment effect in each group (GATES) and examine how different the treatment effects are in the different groups. Next one can describe the covariate characteristics of units in the least and most affected group (CLAN) to understand the treatment heterogeneity. For instance, the share of men in the least affected group versus the share in the most affected group.

D.1 Short run effects of Letter on self adjustment

We first go through the heterogeneous treatment effects analysis for the short run effects of Letter on self-adjustment in 2017. The first result regards what ML method to rely on. We run the analysis using four different prediction methods and pick the best performer among these four. Table A.7 presents the statistics to pick the best ML method. The decision criteria is to maximize the "Best BLP" and "Best GATES" parameters. In this case, the Elastic Net performs better on both dimensions. We therefore use these predictions in the analysis.

Table A.7. Best ML method to predict short run self adjustment heterogeneity.

V1	Elastic Net	Boosting	Nnet	Random Forest
Best BLP	18 014 827	16 599 655	15 862 319	16 808 273
Best GATES	1 853	1 448	754	1 722

Note: Medians over 100 splits in half.

A useful feature of the Generic ML method is that there is a direct test of the degree of heterogeneous treatment effects in the data. In Table A.10 we present the estimates of the conditional average treatment effect and the heterogeneity parameter (HET). Together they form a weighted linear prediction that predicts self-adjustment. By separating their influence we get a first test of the heterogeneity in the data: If the heterogeneity parameter is different from zero there is significant heterogeneity.

We see that the ATEs is in the same ballpark as the one estimated using OLS in Table 4.1 for both methods. Furthermore, in both cases we can reject the null hypothesis of zero heterogeneity.

In Table A.8 we show the ATE for the 20 percent least and most affected groups (the Group Average Treatment Effects) and we note that the difference is large.

Table A.8. Self-adjustment from letter. ATE and GATES of 20 percent least and most affected groups.

Elastic Net			
ATE	Least Affected	Most Affected	Difference
-3 834	-2 681	-7 236	4 533
(-4 736,-2 944)	(-4 667,-684)	(-9 211,-5 193)	(1 712,7 363)
[0.000]	[0.016]	[0.000]	[0.003]

Note: Medians over 100 splits. 90 percent confidence interval in parenthesis. P-values for the the hypothesis that the parameter is equal to zero in brackets.

In comparing the average characteristics of the most and least affected units in a classification analysis (CLAN) we see that the most affected have higher risk-scores, higher previous deductions (going back to 2015) and are more likely to be born in Norway.

D.2 Short run effects of audits on adjustment

We conduct a similar analysis for the short run effects of audits on adjustment in 2017. The method chosen is again the Elastic Net. In Table A.10 we see that the difference in ATE for the 20 percent least and most affected groups is large.

Table A.9. Best ML method to predict short run audit adjustment heterogeneity.

V1	Elastic Net	Boosting	Nnet	Random Forest
Best BLP	1 094 918 447	1 043 373 227	1 007 501 057	1 078 184 786
Best GATES	12 899	12 366	11 507	14 500

Note: Medians over 100 splits in half.

Table A.10. Audit-adjustment. ATE and GATES of 20 percent least and most affected groups.

Elastic Net			
ATE	Least Affected	Most Affected	Difference
- 29 684	-14 533	-54 910	40 749
(-31 631, -27 731)	(-18 787, -10 255)	(-59 293, -50 524)	(34 719, 46 764)
[0.000]	[0.000]	[0.000]	[0.000]

Note: Medians over 100 splits. 90 percent confidence interval in parenthesis. P-values for the the hypothesis that the parameter is equal to zero in brackets.

As compared to the least affected, the most affected have higher risk-scores, higher previous deductions (going back to 2015), are younger, and are more likely to be labor immigrants, single, and men.

For none of the longer run compliance effects do we find any significant heterogeneity in the first test (as reported in Table 3).

E Cost-effectiveness and optimal enforcement policy

Net tax revenue To get the tax revenue generated by Letter and Audit we multiply the total reduction in deductions caused by these interventions with the relevant tax rate, which is 23 %. The numbers are shown in Table A.11. Although it costs much more to conduct a correspondence audit than sending an encouragement letter, 1 333 NOK versus 266 NOK, the net tax revenue is considerably higher for the audit. Both the audit and the letter generate a positive net tax revenue in this population, but the audit approximately five times the net tax revenue generated by the letter.

Table A.11. Net tax revenue effects

	Audit	Letter
a. Short run	- 29 538	-3 584
b. Future compliance	-10 128	-3 900
c. Total compliance effect (a+b)	-39 666	-7 484
d. Tax revenue (0.23*c)	9 123	1 721
e. Unit Costs	1 333	266
Net tax revenue (d-e)	7 790	1 455

F Pre-analysis plan

A Pre-Analysis Plan for "Do Think Twice, it´s All Right: Effects and Mechanisms of Tax Enforcement Policies"

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

3 October 2019

Abstract

We compare the tax compliance effects of two different tax enforcement policies, a relatively costly desk/correspondence audit and a cheap information treatment with a letter asking filers with "risky" filing behavior to take a second look at their tax returns. To assess the total effect of these enforcement policies, we need to take account of both their immediate effects and their long term effects on future filing behaviour. Audits will detect and correct non-compliance on the spot, but may also change subsequent filing behavior. A letter will not disclose non-compliance, but (some) taxpayers may adjust their filings. A letter may also have long term effects on tax compliance. Based on the short run effects of both treatments, which are easy to measure, we describe how we will test the behavioural effects on future filing. We describe the intervention and lay out some important decisions with respect to coding of variables, definitions of samples and the empirical strategy we will apply.

1 Introduction

It is important to ensure that taxpayers follow the tax rules and pay their due taxes, since non-compliance creates both efficiency and fairness losses. To enforce compliance, the tax authorities carry out audits. Audits are costly and targeted towards individuals with indications of risky filing behavior. It has been demonstrated in several studies that in addition to detect and correct infringement on the spot, audits may have lasting effects on taxpayers subsequent compliance with the tax code, [Advani et al., 2017, DeBacker et al., 2015, Løyland et al., 2019]. A softer, and less costly tax enforcement strategy than audits, is to send taxpayers with high risk profile a letter encouraging them to take a second look at their files to see if they have done a mistake. Such a letter may have both immediate effects, as some taxpayers may choose to self-correct this years tax files, as well as long term effects on tax compliance in subsequent years.

In this study we work together with the Norwegian National Tax Authorities (NTA) to compare the effectiveness of these two enforcement policies; a desk based correspondence audit and a "take a second look" letter, referred to as the Audit and the Letter hereafter. We expect the Letter to have a more moderate effect on tax compliance than the Audit. From a cost-benefit perspective, however, the Letter can be the more efficient policy since the costs are negligible compared to the Audit.

We randomize the Letter and the Audit in a population of nearly 15 000 personal taxpayers who claimed relatively high self-reported income tax deductions. A random third of this population got the Letter, another third were exposed to the Audit, and one third where not exposed to any intervention. We will consider both the effect on the treated taxpayers and any spillovers to the spouse.

In order to get a better understanding of the mechanism behind potential behavioural responses to the interventions, we conduct a survey to assess how and whether the interventions influenced taxpayers perceived probability of being audited in the future. According to standard theory, an increase in the perceived probability that evasion will be detected will raise compliance. It is, however, debated how an audit affect the perceived probability of future audits, and also how an audit affects the likelihood that evasion is detected if they are audited [Gemmell and Ratto, 2012, Mittone et al., 2017]. For the Letter on the other hand it is hard imagine that it can lower the recipients perceived probability of future audits.

The structure of this pre-analysis plan is as follows. Section 2-4 describes the institutional details, the treatments, samples and short run effects from data already accessed. From section 5 onwards, we lay out strategies for estimating subsequent compliance effects and tests of mechanisms based on data that will be available to us after the filing of this plan.

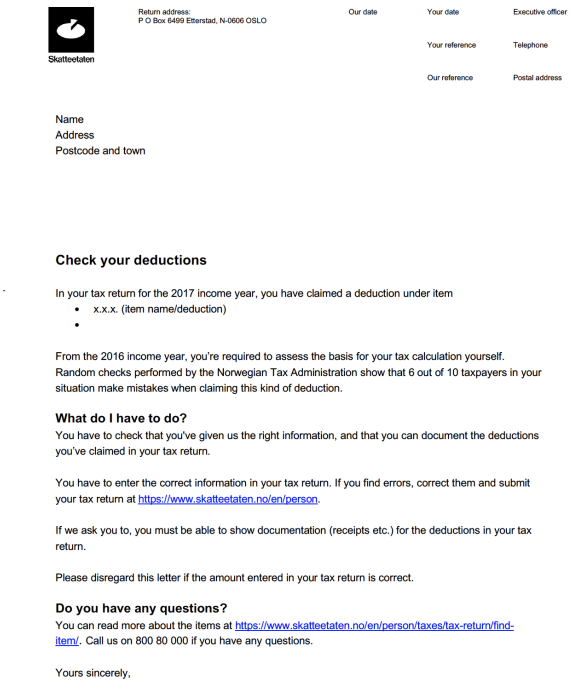
2 Treatments and survey

We consider two treatments. The enforcement interventions are conducted on a selected population of taxpayers with a risky profile, ie. where individual characteristics indicate a high likelihood that they are not complying with the tax code. We provide more details of this population in section 3.2.

One third of this population is selected (randomly) to be audited. The audits are standard low-cost office-based audits, commonly labeled correspondence audits [Hodge et al., 2015]. We define the treatment as being audited. In this type of audit, taxpayers are only notified that they are audited

it the auditor find some irregularities with the claimed deductions. Hence only those who obtain an adjustment can respond in subsequent years to being audited. It is also of some interest, although not directly policy relevant, to check the effects on future filings of those who obtained an adjustment of self reported deductions.

Another third received a letter encouraging them to reconsider and check whether their filed deductions were correct. The letter that was sent to the taxpayers after they self-reported deductions is presented in Figure 1.



Skatteetaten

Return address:
P O Box 6499 Etterstad, N-0606 OSLO

Our date | Your date | Executive officer
Your reference | Telephone
Our reference | Postal address

Name
Address
Postcode and town

Check your deductions

In your tax return for the 2017 income year, you have claimed a deduction under item

- x.x.x. (item name/deduction)
-

From the 2016 income year, you're required to assess the basis for your tax calculation yourself. Random checks performed by the Norwegian Tax Administration show that 6 out of 10 taxpayers in your situation make mistakes when claiming this kind of deduction.

What do I have to do?
You have to check that you've given us the right information, and that you can document the deductions you've claimed in your tax return.

You have to enter the correct information in your tax return. If you find errors, correct them and submit your tax return at https://www.skatteetaten.no/en/person.

If we ask you to, you must be able to show documentation (receipts etc.) for the deductions in your tax return.

Please disregard this letter if the amount entered in your tax return is correct.

Do you have any questions?
You can read more about the items at https://www.skatteetaten.no/en/person/taxes/tax-return/find-item/. Call us on 800 80 000 if you have any questions.

Yours sincerely,

Figure 1. Letter to taxpayer. Check deductions

The taxpayer is asked to take a second look at the self-reported deductions and check if the filed information is correct. The letter also refers to evidence from previous audits where 6 of 10 were found to "make mistakes". Finally, the letter reminds the taxpayer that documentation must be provided upon request. The policy relevant treatment is defined as a letter being sent.

The letter was sent via an electronic personal information platform used by different Norwegian authorities. The taxpayer is notified by an e-mail or SMS that there is letter in the personal inbox from the NTA. There is only this one message, no reminder. It is likely that some of the recipient will not log on to the system and read the letter. Again it has some interest to also estimate the future filing effects of those who logged on the system and opened the letter.

The remaining one third of the sample is subject to business as usual. These taxpayers may obtain ordinary tax audits that are not part of our experiment. These audits are also correspondence audits, but audit selection is based simple rules, or "audit flags" (i.e., not risk scores). These audit flags are typically related to thresholds for self-reported data on specific income or deduction items. These other audits are independent of the random audit we consider, and will therefore not interfere

Table 1. Stylized Timeline of Employee Tax Returns for Tax Year t

Period Year $t+1$	**Action (business as usual)**	**Actors**	**Field Experiment Treatments**	**Outcome short run**
January-February	Third party reporting	Employers and Financial Institutions		Income, interests, wealth
March	Pre-filled tax returns distributed	Norwegian Tax Administration (NTA)		Income by source, deductions, gross wealth, debt
April	Check, correct and self-report if relevant	Taxpayers		Acceptance of pre-filled or ***self-reported*** deductions and income
May-October	Programmed audit routines (flags)	NTA to taxpayers	**Letter (L=1)**	***Self-adjustment*** by tax payers
	Programmed audit routines (flags)	NTA	**Audit (A=1)**	Approval or ***audit-adjustment*** by the NTA
	Programmed audit routines (flags)	NTA	**Non-treatment (A=L=0)**	Approval or audit-adjustment by the NTA
October-December	Final assessment	NTA		***Final total deductions***, taxable income and wealth

with our estimate of the future compliance effects of the examined audits.These are audits that are automatically generated if a tax payer display some filing behaviour.

3 Data structure

3.1 Tax filing and treatment timeline

Given our research question, it is important to have a clear understanding of the sequence of actions and the information exchange between the NTA and taxpayers. Table 1 details the timeline of tax returns for employees. As shown, the filing of tax returns occurs during April and May following the end of the income calendar year. Employers report taxable income to the NTA and they withhold the stipulated amount of taxes workers must pay. Other sources of individual income (such as capital income) are reported by third parties (including financial institutions). Some of the itemized tax income deductions (including donations to charitable organizations) are also reported by third parties (such as the receiving organization). Based on the third-party information, tax returns are pre-filled and distributed by the NTA to taxpayers at the beginning of April. Taxpayers can then make corrections to their tax returns and self-report income and/or deductions until April 30. Over the next months (actually up to three years under the current tax law), the taxpayer can self-adjusted their tax items.

Tax audits are carried out during the May–December period the following income year. Today there are two main types of tax audits for taxpayers. The first is the business as usual system that uses computer generated flags depending on some specific features of the tax return. The second type of tax audit is also targeted, but is randomly selected from the sample defined by the predictive machine learning models that produce taxpayer-specific risk scores. In both cases, the audits may not approve some deductions and make an audit-adjustment.

3.2 Population and samples

The sample consists of high-risk tax payers that score above a threshold of risk scores. The risk scores are based on a predictive machine learning model that produce taxpayer-specific risk scores using a large set of individual characteristics, including the tax return and taxpayer history.

The gross sample consists of 14 902 taxpayers (Table 2). To analyze the short run treatment effects (and future compliance effects), we need data from both the initially submitted tax return for income year 2017, and any self-adjustments by the taxpayer or audit adjustments by the NTA after this initial submission. Data on the initial submission of tax returns for the income year 2017 were extracted from the data warehouse of the NTA as of May 17 2018. At this date, tax return data were missing for 826 taxpayers because they had not yet submitted their tax return (mainly self-employed taxpayers). Tax return data were also missing or incomplete for another 1 108 taxpayers at this date. For these taxpayers, the main reason for missing data were that they had submitted their tax return on paper (non-electronically). In such cases, tax return data are entered manually into the tax systems by the NTA, and this was done after the extraction date (May 17 2018). Due to the overwriting of previous versions of tax returns in the data warehouse, initial tax return data on these 1 934 taxpayers are not available. Our net sample for the analyses of short run treatment effects (and future compliance effects) therefore consists of 12 968 taxpayers.

Filing data on taxpayers contain outliers stemming from different sources of measurement error and/or extreme random numbers. Previous studies have used different strategies to deal with these problems. While DeBacker et al. [2015] winsorize at 90 percent, Kleven et al. [2011] trim income changes (post treatment) at $-200{,}000$ and $+200{,}000$ kroner "to get rid of extreme observations that make estimates imprecise" and Advani et al. [2017] "trim the top 1 per cent to avoid outliers having an undue impact on the results". We are following the practice of trimming, and exclude tax payers with values above the 99th percentile for one or more of the four variables; pre-filled deductions, pre-treatment claimed deductions, post-treatment claimed deductions og post-treatment final deductions.

3.3 Pre-treatment balance

We test for balance on a set of pre-treatment values of relevant variables. The results are seen in Table 3. We test whether there is a difference between the Audit group and the control group (Non-treat) and then between the Letter group and the control group. First we test each variable separately and then we conduct an F-test of whether the variables jointly can predict treatment status. The F-tests are passed for both treatments and the individual variables seem balanced across groups. In the F-test we code eventual missing observations as zero and include a dummy variable for missing status in order not to lose observations.

Table 2. Gross and net samples, attrition and sample exclusion across groups

Criteria	Observations				Comments
	Audit	Letter	Non- treatment	All	
Total sample	4 964	4 945	4 993	14 902	
Missing data due to late subs	285	285	256	826	Mostly self-employed
Missing other reasons	376	366	366	1108	Manual submission, delayed handling by NTA
Trimming	151	164	193	409	Taxpayers with deductions (4 items) above 99th percentile excluded
Final sample	4 151	4 130	4 178	12 459	For short run effects analysis and sample for compliance effects

Table 3. Balance across treatments: Means, standard deviation and test.

Characteristic	Audit	Letter	Non-treatment	Audit vs Non-treat	Letter vs Non-treat
	mean (std dev)			t-test p-value	
Women	0.27	0.28	0.27	0.819	0.374
Age	40.1	39.9	40.3	0.657	0.472
	(18.6)	(10.7)	(32.4)		
Married	0.27	0.26	0.26	0.612	0.807
Norwegian citizen	0.62	0.52	0.61	0.595	0.476
Self employed	0.04	0.04	0.04	0.770	0.897
Risk score 2017	0.697	0.694	0.694	0.078	0.972
	(0.079)	(0.0797)	(0.079)		
Final total deductions 2015	140 636	140 460	140 742	0.949	0.863
	(70 582)	(69 416)	(71 413)		
Final total deductions 2016	144 398	144 324	144 269	0.929	0.969
	(62 833)	(62 420)	(63 125)		
Pre-treatment deduction 2017	195 893	195 649	196 674	0.555	0.437
	(60 878)	(60 282)	(59 961)		
All variables				F-test	F-test
				F(9, 8316) = 0.48	F(9, 8294) = 0.26
				p-value = 0.88	p-value = 0.98

Table 4. Short run treatment effects. Deductions tax year 2017.

	Audit	Letter	Non-treat	Audit	Letter
		Means		Regression coeff vs Non-treatment (p-values)	
A. Pre-filled deductions	129 270	127 569	128 310	960 (0.376)	-741 (0.489)
B. Self-reported deductions	66 623	68 080	68 365	-1 741 (0.063)	-284 (0.765)
C. Claimed deductions ((A+B))	195 893	195 649	196 674	-781 (0.555)	- 1 025 (0.437)
D1. Share with self-adjusted*	0.014	0.110	0.005		0.105 (0.000)
D2. Self-adjustment (NOK)*	-135	3 084	474		-3 584 (0.000)
E1. Share audit adjusted*	0.653	0.062	0.061	0.592 (0.000)	
E2. Audit adjustment (NOK)*				- 29 538 (0.000)	
F. Final total deductions (C+D2+E2)*	161 687	189 079	192 474	- 30 159 (0.000))	-2 503 (0.000)
Sample sizes	4 151	4 130	4 178		

Note: * Total pre-treatment deductions as control in the regression. p-values in parentheses.

4 Short Run Treatment Effects

With short run treatment effects we mean effects in the year of the treatment. The taxpayers have already filed their report when they get either of the two treatments; a Letter or an Audit. Both treatments can alter the final tax files for that year. In the case of the Letter this happens if the recipient reopens the tax report that he or she has filed and corrects self-reported tax deductions. In the case of the Audit this happens if the auditor discovers non-compliance and adjusts the files.

We consider the short term effects on different outcomes specified in Table 4. The short run effects on these outcomes are average treatment effects, equal to differences in means between the treatment, either Letter or Audit, and the control group. To potentially gain precision we also report regression coefficients conditional on pre-treatment deductions.

Table 4 depicts that the share with self-adjustment is 11 percent for the letter group and close to zero for the non-treated. Conditional on pre-treatment deductions, the self-adjustment is about 3 584 NOK larger in the letter group. Obviously, there are tax payers who have over-reported deductions who do not respond to the letter, but are adjusted by the NTA in case of an audit. Presumably, few taxpayers will correct or withdraw deductions that are actually correct according to the tax rules.

Turning to the Audit, we first observe that the share who obtain audit adjusted taxes is 65.3 percent for those who are exposed to this audit, while it is only 6% for those who are in the control group (6% is also the number for the Letter group, indicating that in this population around 6% would obtain an ordinary flag audit, and hence that the effect of the audit we study is to increase the fraction who are audit adjusted by 59 percentage points). The audit group get an audit adjustment of -29,538 NOK as compared to the control group when controlling for pre-treatment deductions. In total we see that the effect of the audit is larger than the effect of the letter on the final total deductions.

Table 5. Outcome and controls. Coding and definitions.

Variable type	Content	Definition/description
Main outcome	Claimed deductions 2018 (t_0+1)	Pre-filled + self-reported (pre-audit year t_0+1). Trimmed
Secondary outcomes	Household Claimed deductions 2018 (t_0+1)	Aggregate Claimed deductions in a household
Controls	Claimed deductions 2017 (t_0)	Pre-treatment year t_0 (Trimmed, from Table 4)
	Age	Years since birth by 31 Dec
	Immigrant	Resident in Norway, foreign born with foreign born parents
	Temporary Labour Migrant	Non-resident, with D-number, citizenship
	Marital status	Dummy = 1 if married/cohabitat
	Risk score 2017	A continuous variable in $[0,1]$
Groups	Self employed	Dummy = 1 if self employed
	Female	Dummy = 1 if registered as woman

5 The effect on future tax compliance

5.1 Coding outcomes and controls

In this section we present the variables to be used in the analysis of future compliance. We start with the main outcome variables and continue with the covariates and the variables used to study heterogeneous effects. Our main outcome is the claimed deductions for tax year 2018 (in NOK). We cannot use final total deductions as an outcome (as we did for short run effects) since the field experiment was implemented in such a way that the group that received a letter in 2018 were audited in July-October 2019. Self-reported deductions were filed before this "extra audit" and are not affected by the 2019 audit, since the taxpayer had no information about the forthcoming audit.

We view the claimed deductions as a good outcome measure as it is highly influenced by behaviour in terms of self-reported deductions. In addition, the measure is equal to the final total deductions (and thereby net taxable income) in case of no audits. Our secondary outcome is the partner's claimed deductions for the tax year 2018 (in NOK).

The control variables and the codings are described in Table 5. These control variables will also be used to test for balance where we create a Table as the one in Table 4. The success of the randomization will be judged by the F-tests in the regressions where we include all the variables together and see whether they predict the treatments together. The control variables will also be used in exploratory analyses of heterogeneous treatment effects (see below). Whenever a control variable is included in a regression together with other variables we will code eventual missing observations as zero and include a dummy variable for missing status in order not to lose observations.

5.2 Main empirical specification and hypothesis

We estimate the following regression using ordinary least squares:

$$Y_{t_0+1,i} = a + bLetter_i + cAudit_i + u_i \tag{1}$$

$Y_{t_0+1,i}$ is the claimed deductions for tax year 2018 (in NOK) and we use robust standard errors. The main specification is a regression using individuals still present in Norway (unless there is non-random attrition, see below) without any controls included, if there is no imbalance across the groups.

If there is imbalance across the groups, the main specification will be one with the full set of control variables.

Our main hypotheses are

- (i) $b < 0$
- (ii) $c < 0$
- (iii) $b < c$

Taxpayers with a spouse will typically not make filing decisions in isolation. Some deductions are household specific and can potentially be transferred from one spouse to the other as a response to the Letter and Audit treatment. Spouses may also update their knowledge about tax rules or audit probabilities when their partner has been subjected to Audit or Letter. Both mechanisms make it important to include the filing behavior of the spouse in the estimation of future compliance effects of these treatments. We therefore also run the same regressions for the spouses.

We will also see if precision can be improved by adding the control variables described above. In particular, we will investigate if we can improve precision in the estimates by picking optimal controls from the total list of controls using a double debiased LASSO procedure [Belloni et al., 2014].

5.3 Test of mechanisms

In order to get a better understanding of the mechanism behind potential behavioural responses to the interventions we conduct a survey to assess how the interventions influenced taxpayers perceived probability of being audited in the future. In particular, we fielded a phone survey where we tried to contact all individuals in the main sample. We were not able to get the phone numbers for everyone and not everyone responded. Our main question of interest is Subjective Detection Risk (SDR) and is as follows:

"What do you think the probability is that the tax authorities will control your reported taxes in 2019?". The answer categories are

- 1) Not likely at all (0 percent).
- 2) Very unlikely (1-20 percent).
- 3) Quite unlikely (21-40 percent).
- 4) Neither likely nor unlikely (41-60 percent).
- 5) Quite likely (61-80 percent).
- 6) Very likely (81-99 percent).
- 7) Certainly (100 percent).

We will retain the continuous coding of this variable (1-7) and call it SDR and run the following OLS regression:

$$SDR_i = e + dLetter_i + fAudit_i + u_i \tag{2}$$

We expect that $d \geq 0$. We cannot think of any reasons why a letter with this content should induce recipients to update their perceived probability of an audit downwards. For the effect of Audit (f) there is a literature arguing that being audited today may reduce the perceived probability being picked out in future audits, see Mittone et al. [2017]. Hence the sign of f is a-priori uncertain, but we find it more likely that $f > 0$.

It is unclear what variables we will be able to connect to the survey from the register data for reasons of anonymity. The main specification will in any case be one without any control variables but the tests of balance as well as the heterogeneity analyses may be restricted to the subset of the variables we are able to include.

6 Sample, attrition and trimming

We use the "Gross sample" from Table 2 as our point of departure in analyzing the effects on future tax compliance. There will be missing values and attrition from this sample due to trimming, deaths, and migration. In particular, there are a large number of migrants in our initial sample and many of them are likely to have left Norway. While the decision to stay may be affected by the treatment we do not view this as very likely. We will test whether there is a difference between the groups in the probability of being present in the tax register. The compliance effects will be tested for those present in the tax liability register of the NTA for 2019. If there is non-random selection into presence in the register we will also present estimates where we include everyone and code those not in the register as having zero deductions. We will further follow the trimming practice for restricting the gross sample as laid out in section 3.2.

With respect to the SDR mechanism, the sample will further be reduced by non-response to the survey. For that sample we will test attrition in the following way:

$$InSDRSample = g + hLetter_i + kAudit_i + u_i \tag{3}$$

Where *InSDRSample* is a dummy for being in the survey sample with valid information on question about detection probability in 5.3. If there are statistically significant coefficients for h and k (at the 5 percent level) we will follow Kling et al. [2007]'s correction. We will obtain lower bounds of the treatment effect by replacing missing observations in the treatment (control) arms by the corresponding arm's mean value minus (plus) 0.05, 0.10 and 0.20 standard deviations of the control group. Upper bounds of the treatment effects will be constructed in a symmetrical way.

7 Exploratory analyses

We will explore heterogeneity in the treatment effects by estimating equation (1) by groups. The groups we will create are:

- Self-employed

- Immigrant background; native, immigrant, foreign non-resident worker.
- Age-groups
- Gender
- Marital status
- Risk score

We will also use machine learning techniques to automate the search for heterogenous treatment effects. We will use all the control variables for this. There are many different types of machine learning algorithms and we have decided to use random causal forests (R package grf, Wager and Athey, 2017). 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.

It will also be interesting to examine which groups that are affected in their SDR by the treatment. We will also investigate if it seems to be the case that the effects of the different treatments run via SDR.

8 Power

We are testing two main hypotheses with the 12,495 individuals (see 2). In testing the b and c coefficients, we can think of them as tests with about 8,300 individuals. At the conventional level of significance of 0.05 and a power of 0.8, our sample size would allow for a minimum detectable effect of 0.06 standard deviations. We will also adjust the p-values for the fact that we are testing the impact on two outcomes. We follow the recommendations of Fink et al. [2014] and use a method developed by Benjamini and Hochberg [1995] and Benjamini et al. [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 two hypotheses and a significance level (a) of 0.05, the critical $p-$value would be 0.025 for the one with the lowest p-value (0.05* 1/2, which is the same as a Bonferroni correction. For the second hypothesis, the critical $p-$value is 0.05 (0.05*2/2). The minimum detectable effect for our variable with the lowest p-value after accounting for multiple hypothesis testing (p=0.025) is 0.07 standard deviations. We conclude that our experiment is very well powered.

9 Archive and data disclosure

The pre-analysis plan is archived before the data from the survey is received and before any analyses have been made on the compliance effects data for the year 2018 (see attached letters in the Appendix). We archive it at the registry for randomized controlled trials in economics held by The American Economic Association: https://www.socialscienceregistry.org/ on October 4 2019. We expect to get the data in by October 7 and we will then start analyzing data.

References

Arun Advani, William Elming, Jonathan Shaw, et al. The dynamic effects of tax audits. Technical report, Institute for Fiscal Studies, 2017.

Alexandre Belloni, Victor Chernozhukov, Lie Wang, et al. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014.

Yoav Benjamini and Yosef Hochberg. 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, 1995.

Yoav Benjamini, Daniel Yekutieli, et al. The control of the false discovery rate in multiple testing under dependency. *The annals of statistics*, 29(4):1165–1188, 2001.

Jason DeBacker, Bradley T Heim, Anh Tran, and Alexander Yuskavage. Once bitten, twice shy? the lasting impact of irs audits on individual tax reporting. *Journal of Financial Economics*, 117(1): 122–138, 2015.

Günther Fink, Margaret McConnell, and Sebastian Vollmer. Testing for heterogeneous treatment effects in experimental data: false discovery risks and correction procedures. *Journal of Development Effectiveness*, 6(1):44–57, 2014.

Norman Gemmell and Marisa Ratto. Behavioral responses to taxpayer audits: evidence from random taxpayer inquiries. *National Tax Journal*, 65(1):33, 2012.

Ronald H Hodge, Alan H Plumley, Kyle Richison, Getaneh Yismaw, Nicole Misek, Matt Olson, and H Sanith Wijesinghe. Estimating marginal revenue/cost curves for correspondence audits. In *IRS Research Bulletin, Presented at the 2015 Internal Revenue Service—Tax Policy Center Research Conference*, 2015.

Henrik Jacobsen Kleven, Martin B Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez. Unwilling or unable to cheat? evidence from a tax audit experiment in denmark. *Econometrica*, 79(3):651–692, 2011.

Jeffrey R Kling, Jeffrey B Liebman, and Lawrence F Katz. Experimental analysis of neighborhood effects. *Econometrica*, 75(1):83–119, 2007.

Knut Løyland, Oddbjorn Raaum, Gaute Torsvik, et al. Compliance effects of risk-based tax audits. 2019.

Luigi Mittone, Fabrizio Panebianco, and Alessandro Santoro. The bomb-crater effect of tax audits: Beyond the misperception of chance. *Journal of Economic Psychology*, 61:225–243, 2017.

The Norwegian
Tax Administration

Return address:
P.O.Box 9200 Grønland, N-0134 OSLO

Our date
24. Sept 2019

Your date

Inquiries to
Kavita Bhamra

+47 800 80 000
skatteetaten.no

Your reference

Telephone
47 97 63 6045

Org. nr:
974761076

Our reference
2018/456066

Postal address
P.O.Box 9200, Grønland
N-0134 OSLO

To whom it may concern

On exchanging survey data between the Norwegian Tax Administration and Opinion

Department of Security and Safety (DSS) in the Norwegian tax administration (NTA), is responsible for anonymization of taxpayer data collected by a private Norwegian survey institute, Opinion, on the behalf of Department of Analysis (DOA), NTA, and Oslo Fiscal Studies (OFS) at the University of Oslo. DSS is also responsible for the link key between the taxpayers social security number and a survey number constructed randomly for this particular survey. The responsibility held by DSS implies that we receive a sample of taxpayers from DOA, and conveys these for a survey by Opinion. When Opinion finish their survey, the collected data are sent back to DSS and are matched with other background variables before anonymization. The matched and anonymized data will be in the custody of DSS until January 15th 2020. After this date, the data will be made available for DOA, OFS and other interested parties.

Yours faithfully

Svein Mobakken
Director of security and safety
Security and safety department
Norwegian Tax Administration

The Norwegian
Tax Administration

Return address:
P.O.Box 9200 Grønland, N-0134 OSLO

Our date
25. Sept 2019

Your date

Inquiries to
Knut Løyland

+47 800 80 000
skatteetaten.no

Your reference

Telephone
47 92 28 0634

Org. nr:
974761076

Our reference
2018/456066

Postal address
P.O.Box 9200, Grønland
N-0134 OSLO

To whom it may concern

Self-declaration

Kotsadam, Løyland, Raaum, Torsvik and Øvrum (2019) have developed a pre-analysis plan for the analysis of a field experiment that was conducted by the Norwegian Tax Administration (NTA) in 2018 (for the tax year 2017).

The pre-analysis plan includes a section on estimating the future compliance effects of the two treatments (tax audits of taxpayers and letters to the taxpayers). This part of the analysis requires tax return data on the participants from the tax year 2018.

While Arnstein Øvrum and Knut Løyland have had access to part of the data as part of their daily work, we hereby confirm that the specific data related to this project have not yet been analyzed at the date of publishing the pre-analysis plan.

Yours faithfully

Marcus Zackrisson
Head of Department of Research and Analytics
Department of Research and Analytics
Norwegian Tax Administration