Compliance effects of risk-based tax audits*

Knut Løyland, Oddbjørn Raaum, Gaute Torsvik and Arnstein Øvrum

29 August 2019

Abstract

Tax audits detect and correct noncompliance on the spot, but can also change compliance in future filing behavior. In contrast to the immediate impact on tax revenues, the audit effects on future filing behavior are not directly observable and must be estimated within a counterfactual framework. This paper uses data from an experiment with random assignment to estimate the effects of audits on future self-reported deductions among wage earners. Modern tax administrations use machine learning methods to guide selection of individual taxpayers in risk-based audits. We study how the future filing response to an audit varies with the risk score of the taxpayer. This estimate speaks directly to the design of optimal risk based audits that typically sets a risk threshold and audit all taxpayers with a risk score above that value. When we account for effects on future filing behaviour, we show that the risk score audit threshold applied by the Norwegian tax administration is set far above the threshold that maximizes net public revenue.

JEL-codes: D04, H26, H83

Keywords: tax audits, tax revenue, tax reporting decisions, income tax, machine learning, risk profiling.

[*] This paper is part of the Oslo Fiscal Studies (OFS) research program financed by the Norwegian Research Council. The authors gratefully acknowledge the assistance of Majken Thorsager and Øystein Olsen in preparing the 2013 audit data and the analytical scoring for 2013 and 2014. The authors would also like to thank Edwin Leuven, Tarjei Havnes, Andreas Kotsadam, Attila Lindner and participants in the OFS workshop "Empirical Analyses of Tax Compliance" and Skatteforum 2018 for their useful comments and suggestions. Affiliations: Løyland and Øvrum (Norwegian Tax Administration), Raaum (Frisch Centre and OFS) and Torsvik (University of Oslo and OFS).

1 Introduction

Audits are a cornerstone of tax enforcement policy. They detect and correct noncompliance on the spot, but may also change taxpayers' future filing behavior. This compliance effect of tax audits is important for the design of optimal enforcement policy (Keen and Slemrod, 2017). A-priori it is not clear how an audit of this years tax return will affect filing behavior in subsequent years. Taxpayers violate the tax rules for different reasons. Some want to do it right, but deviate because they do not know the rules. Others deliberately underreport taxable income to evade taxes. Confused individuals will comply after they learn the correct rules from the audit. The future responses to audits among tax evaders are, however, ambiguous and depend on how the audit affects their perceived probability of subsequent audits, and their assessment of how likely it is that evasion is detected if they are audited (Kleven et al., 2011; Gemmell and Ratto, 2012).

Empirically, it is challenging to identify future responses to audits. While the immediate tax revenue from audit outcomes is observable, the impact on future filing behavior is not and must therefore be estimated within a counterfactual framework. Regular operational tax audits often select units for reasons only known or observed by the tax authorities, raising fundamental identification problems. To overcome these, we use data from a field experiment with random audit assignment. In 2013, the Norwegian Tax Administration (NTA) targeted a group of taxpayers with relatively high self-reported income tax deductions. From a population of about 310,000 taxpayers, the NTA audited a random sample of 15,000. These were standard low-cost office-based audits, commonly labelled correspondence audits (Hodge et al. (2015)). The objective of the experiment was to build a machine learning model, using an extensive list of individual taxpayer characteristics to risk score their noncompliance.

We use these experimental data to study how an audit affects filing behavior in subsequent years. With four years of post audit data we estimate the persistence of the behavioral responses. In addition to the average treatment (audit) effect, we also estimate heterogeneity in the audit response across the entire range of risk scores of this population. We estimate the effect for the household, accounting for any impact on future tax filings of the spouse.

The random audit uncovered that around 20% of the wage earners with self-reported deductions claimed fictitious or unwarranted expenses. The audit also had strong and long-lasting effects on compliance in subsequent years. In aggregate, the tax revenues arising from future compliance effects exceeded the immediate tax revenue disclosed by the audit. Important for tax enforcement policies, we find that the effect on subsequent compliance behavior increases with the risk score of the audited taxpayer.

We also provide evidence from an operational audit in 2014. The NTA used the model to risk score taxpayers and audit those with a risk score above a threshold. We use a regression discontinuity design to estimate future responses to this risk score targeted audit. While the random audit data enable us to trace out behavioral responses across the entire risk score, the subsequent risk based audit provides precise estimates of behavioral responses to audits around the risk threshold. The operational audit also had significant and lasting effects on self-reported deductions. The evidence from the two audits is highly consistent.

There is a growing literature using data from experiments where tax administrations draw audits randomly from the entire universe of taxpayers to learn the scope and scale of tax evasion and to estimate deterrence effects of audits, some recent examples are (DeBacker et al., 2018; Gemmell and Ratto, 2012; Advani et al., 2017; Kleven et al., 2011). The policy relevance of the population wide average compliance effects of random audits is, however, questionable. First, in these audits taxpayers are often informed that they have been selected for a random check, and this may affect their behavioral responses (Slemrod, 2018)[1]. Second, it is unclear whether a population-wide average compliance effect is informative about how taxpayers that are selected for standard operational risk-based audits respond to being audited. On the other hand, data from operational risk-based audits have low internal validity, since the selection of taxpayers into audit is typically triggered by suspicious filing patterns that often reflect transitory shocks (negative for income and positive for deductions). The future reporting of the audited taxpayers is therefore mean-reverting, and without proper empirical design, the post audit transitory component could mistakenly be interpreted as a behavioral response (Ashenfelter, 1978).[2]

By using data with random assignment to targeted, risk-based audits, our paper solves the quandary that the estimates of audit responses are either causal but not policy relevant, or relevant but not causal – a problem Slemrod (2018) describes a methodological catch-22 in the estimation of tax enforcement effects. We identify behavioral responses to audits that are both causal *and* informative for the design of actual tax enforcement policy. Tax administrations increasingly use big data and predictive analytics to predict noncompliance risks among taxpayers as a basis for targeted audits towards high-risk filers. Our data enables us to compare the compliance of audited and unaudited taxpayers along the whole range of risk scores and thereby estimate how the total revenue effect of audits depends on the risk score of the audited. This information enables the tax authorities to calculate the elasticity of tax revenue with respect to enforcement input, which is a key parameter for designing

[1] This is, however, not the case for the study conducted by Kleven et al. (2011).

[2] Mazzolini et al. (2017) employ observational data on tax audits to estimate compliance effects. They then use either a matching and/or a difference in differences estimator to construct the counterfactual to audit. Heckman and Smith (1999) demonstrate that these estimators can yield biased estimates of treatment effects if the selection into treatment is reminiscent of "Ashenfelter's dip".

optimal audit policy (Keen and Slemrod, 2017; Kreiner, 2010). Comparing the tax revenue effects at different risk scores with the unit costs of correspondence audits, we find that the threshold for the operational risk based audits is set far above the risk-score that maximizes the total (short- and long-term) net revenue effect of the audit.

Our study also highlights a considerable scope for noncompliance among wage earners. It is often claimed that third-party reporting and tax withholding makes it almost impossible for this group of taxpayers to underreport income (Kleven et al., 2011; Kleven, 2014). In most countries, however, employees have considerable leeway when it comes to claiming expenses that can be deducted from gross earnings to reduce net taxable income, (Fack and Landais, 2016; Gillitzer and Skov, 2018). Compliance studies that ignore wage earners are likely to miss an important element of effective tax enforcement policies.

Finally, we contribute to the literature by estimating how audits affect the filing behavior of the spouse. 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 audit. Spouses may also update their knowledge about tax rules or enforcement probabilities when the other spouse is audited. It is therefore important check if increased compliance for the audited is counteracted by the filing response of the spouse and we do estimate the effects for the total household, accounting for any impact on future tax filings of the spouse.

The structure of the paper is as follows. Section 2 reviews key features of the Norwegian tax system, including tax audits and the random audit data. Section 3 gives details on the empirical strategy used to estimate the compliance effects of the random audit. Section 4 presents the main results based on the random audit. Section 5 uses data from the operational audit to estimate the effect of audit locally around the risk threshold. Section 6 discusses optimal audits and Section 7 concludes.

2 Institutional setting and data

2.1 Taxes, tax filing and audits in Norway

In 2015, Norway had about 5.2 million inhabitants, of which 79% were liable to pay taxes and file a tax return. The administration and enforcement of personal taxation in Norway is divided between the NTA in assessing personal taxes based on gross and taxable income, and the responsibility of the municipalities in the collection of taxes. Norwegian income tax differentiates between income from work (Y) and capital (I). For wage earners, taxes also depend on deductions (D), with the taxes liable (T) being given by $T = t(Y, I, D)$, where

, $0.23 < \frac{\partial T}{\partial Y} < 0.47$, $\frac{\partial T}{\partial I} = -\frac{\partial T}{\partial D} = 0.23$ such that the marginal tax on wage income is higher than for interest and other capital income. Married couples in Norway are typically taxed as individuals.

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 time line 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 it withholds 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 prefilled and distributed by the NTA to taxpayers at the beginning of April. Wage earners can then make corrections to their tax returns and submit self-reported items (income and/or deductions) until April 30. The difference between the *total* (income or deductions) in the final tax return and those in the *prefilled* version is what we label as *self-reported* in this analysis.

Table 1. Stylized time line of employee tax returns for tax year t

Period Year $t+1$	**Action**	**Actors**	**Outcomes**
January–February	Third-party reporting	Employers and financial institutions	Income, interests, wealth
March	Prefilled 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 prefilled or self-reported income and deductions items
May–Dec	Checks (standard, automatic) Audit	Programmed audit routines (flags) Risk-score above threshold Selected taxpayers	Approval or audit adjustment Approval or audit adjustment Documentation
	Final assessment	NTA	Taxable income and wealth, sanctions

Tax audits are carried out during the May–December period following the income year.

Today there are two main types of tax audits for wage earners. The first uses computer generated flags that depend on some specific features of the tax return. The second type of tax audit is also targeted, but based on predictive machine learning models that produce taxpayer-specific risk scores. The model that provides the risk scores from a large set a individual characteristics, including the tax return and taxpayer history, is estimated from the random audit data we study in the paper. It is always the taxpayers with the highest risk scores that are selected for a tax audit, but the exact number of audits that are carried out depend on budget allocations, and may differ between years.

The audits will result in higher taxable income if the self-reported items are not accepted by the NTA. In the case of misreporting, taxes owed are paid with interest. In addition, a fine can be imposed if the misreporting is considered as deliberate cheating. It is important to note that many taxpayers are audited without being contacted by the NTA. They are unaware of the audit and should therefore not be affected. This occurs when the NTA has sufficient information to approve the self-reported items without further correspondence with the taxpayer.

2.2 Random Audit Data

The NTA provides our data on audits and tax returns. In 2013, the NTA singled out a population of about 310,000 taxpayers claiming self-reported deductions above an (undisclosed) threshold of X Norwegian kroner (NOK) on one or two items from a list of 29 specified expense deduction items.[3] The NTA selected a stratified random sample of 15,000 individuals from this population for an audit of all deduction items for which the taxpayer had self-reported above X NOK. See the Appendix for details on the different deduction items, the stratification process, and corresponding weights used in our estimations.

We make a few sample selection restrictions. First, we exclude self-employed taxpayers and focus on deductions among wage earners and transfer recipients (e.g., the retired, unemployed). Second, we exclude taxpayers below the age of 17 and above the age of 70 in the year of the tax audit. Finally, we exclude a small number of taxpayer-year observations because of suspected data errors. Due to outliers, variables in NOK are winsorized at the 99 % level. Variables in NOK that include negative values are also winsorized at the 1 % level.

With random assignment, we expect those who are audited and those who are not, to be equal across all observable and unobservable pre-audit characteristics. Table 2 provides the pre-audit characteristics of the audited and unaudited taxpayers in the 2013 population. As

[3] For reasons of confidentiality, we are not permitted to disclose the exact amount that triggered this audit. Taxpayers who self-reported this amount of deductions on three or more deduction items were automatically audited (flagged).

expected, the means are similar for the two groups. Owing to the large sample of unaudited tax payers, age and gender are individually significant predictors of audit, but the point estimates are negligible (see Appendix).

Table 2. Balance by treatment in the audit. 2013.

	Unaudited		**Audited**	
	Mean	Std. Dev.	Mean	Std. Dev.
Total deductions	172 009	76 586	171 839	38 934
Third-party reported (prefilled) deductions	121 900	62,580	122 281	62 179
Self-reported deductions	49 114	39 709	48 596	38 934
Total income	563 091	374 175	561 899	361 054
Third-party reported (prefilled) income	553 961	358 357	552 676	346 141
Age	40.3	11.7	39.9	11.7
Female	0.334		0.342	
Immigrant	0.173		0.164	
Married	0.448		0.458	
Risk score	0.381	0.225	0.383	0.228
Observations	252 537		12 011	

Note: The split between audited and unaudited taxpayers is based on the audit of a random sample of taxpayers claiming more than X NOK on one or two of 29 pre-filled deduction items (the 2013 population). Alongside this audit, the tax administration conducted other audits based on flags raised because of deviant tax filing behavior in 2013. These other audits are running in the background and are independent of the random audit we consider, and will therefore not bias our estimate of the future compliance effects of the examined audits.

The main purpose of the 2013 random audit was to collect data to estimate individual risk scores of the taxpayers. Machine learning techniques were used to predict non-compliance (i.e. illegitimate self-reported tax deductions) by means of the outcome of the audit in combinations with about 50 individual characteristics, including current and historic tax filing. The individual non-compliance probabilities is labelled risk score, calculated for the whole sample. As shown in Table 2, the average risk score is similar for the audited and unaudited taxpayers. When we estimate the probability of audit, the risk score coefficient is negligible and far from significant (see the Appendix for details). Thus, the machine-learning model used by the tax authorities to indicate non-compliance cannot predict the audit status of the taxpayer.

Alongside with our random audit, the tax authorities performed regular audits based on suspicious filing behavior (flags). Hence, even in the control group of our study who consists of individuals not exposed to the random audit, some are audited and get their tax

Table 3. Audit Outcomes

	Unaudited		**Audited**	
	Mean	Std. Dev.	Mean	Std. Dev.
Adjustment from audit (random or flag)	3,731	21,146	8,262	27,696
Fraction with adjustments	0.056		0.204	
Adjustment in NOK among adjusted	55,726	72,064	37,860	51,571
Observations	252, 537		12,011	

files adjusted by the tax authorities. It is important to note that these flag based audits are business as usual and are independent of the random audit we study. The outcome for audited and non-audited taxpayers in the audit year is described in Table 3. Both the fraction who obtained an adjustment of their tax filings, and the average adjustment is considerably higher among taxpayers who were randomly selected for audit. On the other hand, conditioning on getting an adjustment, the amount adjusted is higher among those who are audited due to flags (and not by random selection).

The challenge faced by studies that estimate audits effects on future filing compliance without random assignment, can be illustrated by the typical time profile of self-reported deductions among individuals at risk of being audited. In the audit year, the average self-reported deductions of non-audited tax payers were 49,091 NOK, these deductions dropped to 36,267 and 29,615 NOK in the two subsequent years. This mean reversion in a population of taxpayers with "risky" filing behavior is illustrated in the Appendix which displays the time profile of the self-reported deductions of non-audited tax payers (in the 2013 audit). Mean reversion makes it difficult to identify causal effects of targeted audits without some kind of experimental design, simply because the units selected for audit will tend to have outcome dynamics that can easily be interpreted as responses to the audit. In our study of random audit, however, both the treated (those audited) and the controls (not assigned to the audit) have the same mean-reverting process in the absence of treatment.

3 Empirical strategy

To assess how the random audit influenced future tax filing, we estimate the following equation

$$y_{i,t_0+k} = \beta_{t_0+k} Audit_{i,t_0} + \gamma X_{i,t_0} + \varepsilon_{i,t_0+k}, \tag{1}$$

where y_{i,t_0+k} is a self-reported item for taxpayer i in year $t_0 + k$, and t_0 is the year of the audit. Our main outcome is the aggregate self-reported deductions as these items defined the population from which the tax authorities randomly audited a subsample. $Audit_{i,t_0}$ is an indicator variable equal to one for audited taxpayers and X_{i,t_0} is a set of pretreatment controls that we may add to gain precision. With this specification, the β_{t_0+k} coefficients capture what we define as compliance effects. For each post-audit year, the coefficient is equal to the difference in average self-reported deductions between the audited and unaudited and captures the average treatment effect (ATE) of an audit within the population of taxpayers with high self-reported deductions. We use the term "compliance effects" to denote how an audits affects subsequent (post-audit) tax filings of those who are audited, not including the immediate outcome of the audit.

Since everyday tax enforcement must take a stand on who to select for audit, it is informative to see if and how the compliance effects of audit vary with the risk profile of the taxpayer. We have the risk scores of both audited and non-audited tax payers and can therefore compare future filings for both groups over the whole range of risk scores. To identify how the future effects of audits vary with risk score of the audited tax payer, we estimate a flexible functional form model with risk score dummies that we interact with an audit indicator.

The audit we study is an office based correspondence audit. For such audits, the vast majority of taxpayers who did not receive an adjustment of their tax returns will not be aware that they have been audited. When we divide the (average) compliance effect, β, by the fraction who actually had their deductions adjusted, we obtain an estimate of the average behavioral effect of obtaining an adjustment.[4]

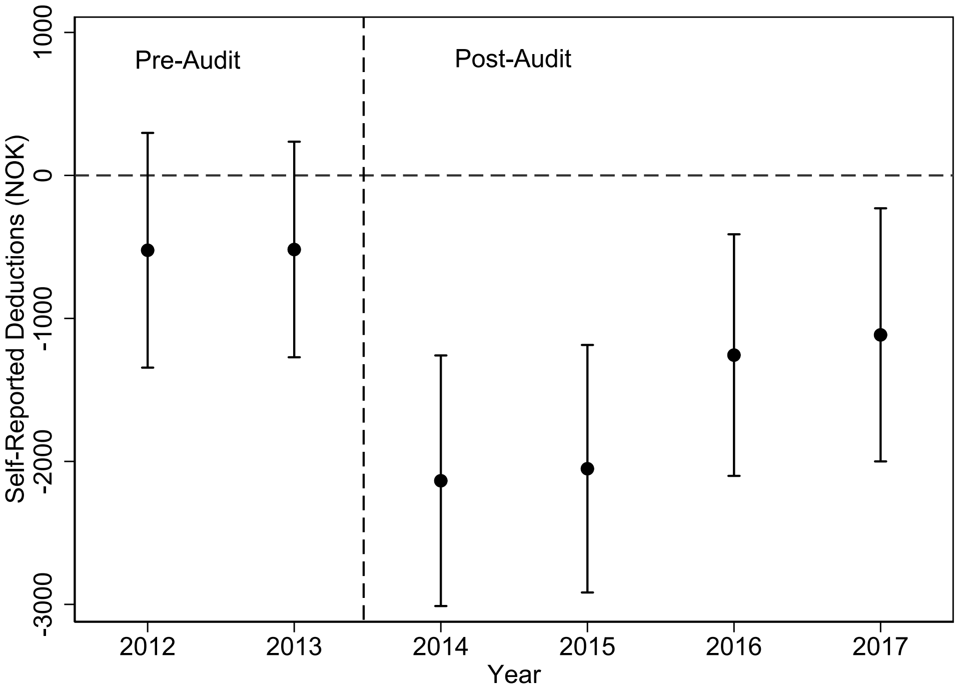
4 Results

4.1 The average compliance effect

Our population of taxpayers is audited because they self-report income tax deductions above X NOK. We therefore use self-reported deductions as our outcome when estimating the effects of audit. Figure 1 depicts the (average) compliance effects on self-reported deductions, up to four years after the audit.

[4] It is possible that some of those audited were asked for documentation and would then become aware that the tax authorities had looked into their tax files.

Figure 1. Compliance effects. Self-reported deductions.



.
Notes: The figure depicts the year by year average difference in self-reported deductions among those who where audited and those who were not audited.

The audit caused self-reported deductions to decline by more than 2,000 NOK in each of first two years after the audit. The compliance effect is reduced over time, but remains significant at just above 1,000 NOK annually for the next two years. When we aggregate the self-reported deductions over the four post-audit years, the audit caused a decline in self-reported deductions of 6,707 NOK (st.err. 1,563). This is a substantial amount, it is higher than the average adjustment made by the audit. Hence, in terms of recovered tax revenues generated by the audit, the indirect effects on future filing behavior is at least at par with the revenue that comes directly from the outcome of the audit.

Spillover to spouse 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 audit. Spouses may also update their knowledge about tax rules or audit probabilities when their partner has been audited. Both mechanisms make it important to include the filing behavior of the spouse in the estimation of future compliance effects of audits. In our analysis we keep the sample unchanged and simply add the outcome for the spouse for the audited taxpayer that was married or had a cohabitant when audited.

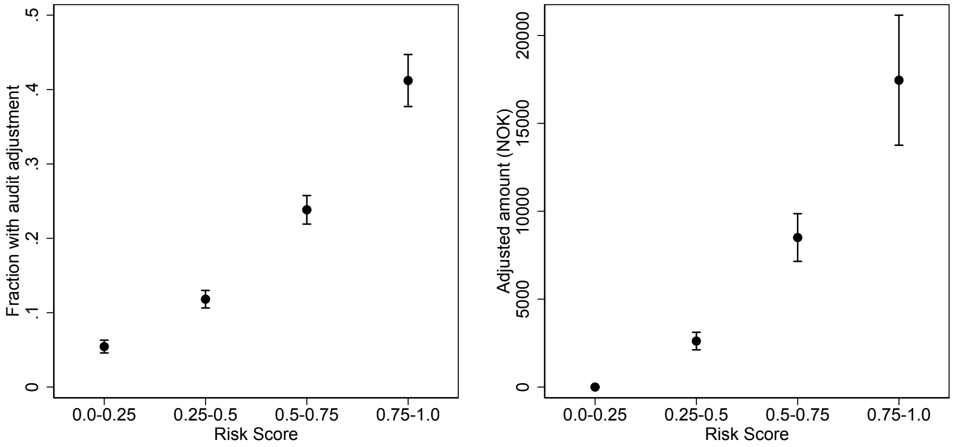
Total self-reported deductions will naturally increase when we use self-reported household deductions as our outcome, but given random audits, the average contribution to the deductions from spouses should be identical for the treatment and the control groups, unless the audit generates a reallocation of items within the household or creates behavioral effects on the spouse. Reallocation would imply that we should observe audit effects on future filings that are smaller, closer to zero, when we estimate the audit effect on household data, than the result we got on individual data. Behavioral spillovers, on the other hand, contribute to stronger responses.

The evidence suggests that the behavioral spillovers dominate. If we aggregate self-reported deductions over the four post-audit years for the audited and spouse, the estimated drop in self-reported deductions is 7,786 NOK (std.err 1,600), which is a slightly larger effect than the one we find when only the audited tax payers are included.

4.2 Heterogenous compliance effects by risk score

There are two main reasons why we expect the compliance effect to be increasing in the individual risk score. First, in correspondence audits the tax authorities only contact those for whom they find irregularities in the tax filing. Therefore, only those who did not have their self-reported deductions approved, will be aware that they have been audited. In Figure 2 (left panel), we display the fraction adjusted by risk score, splitting the population in four risk score brackets. While only 5 percent of the payers in the lowest risk bracket had their self-reported deductions adjusted, this happened to more than 40 percent of the tax-payers in the group with the highest risk scores. Second, Figure 2 also shows that the amount adjusted by the tax authorities increases in the risk score (right panel). For the taxpayers with high deductions, the potential for increased compliance (in NOK) is expected be larger. The patterns of Figure 2 simply reflect how the risk score is constructed; i.e. the magnitude of the self-reported deductions is one of the factors that predicts noncompliance.

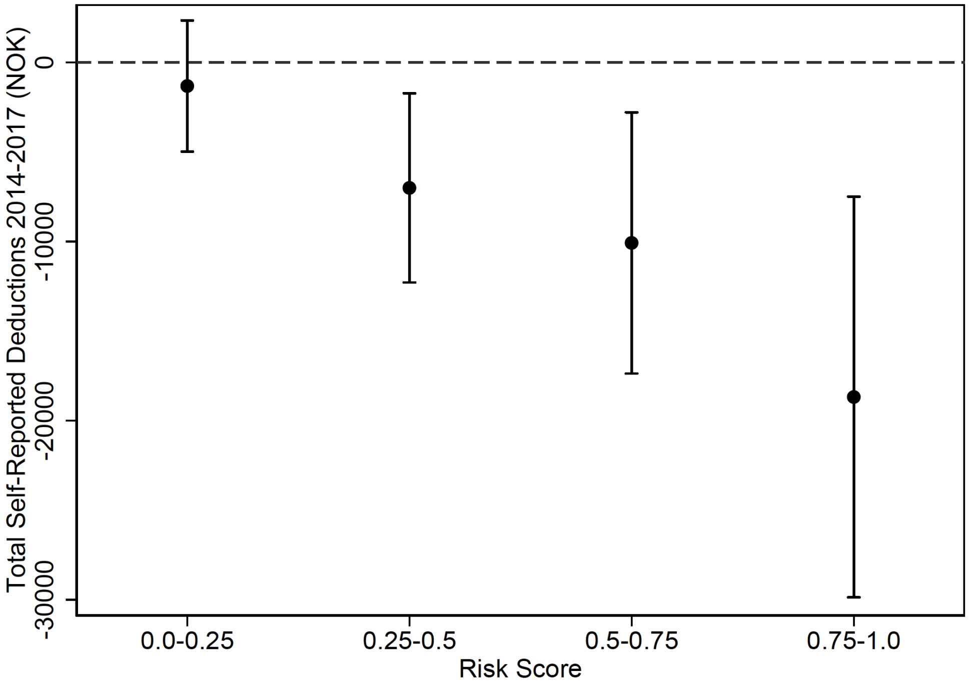
Figure 2. Audit outcomes by risk score



Note: The left panels shows, within four different risk score intervals, the fraction of audited tax payers that did not get their deductions approved. The right panel shows the average amount that was adjusted by NTA at different risk score levels.

The compliance effects differ by risk score as expected. Figure 3 depicts the aggregated four year drop in post-audit self-reported deductions for different risk score levels. When we split the sample based on four risk score intervals, the point estimates are substantially different. The compliance effect is close to zero in the lowest bracket and becomes statistically significant when the risk score exceeds 0.25.

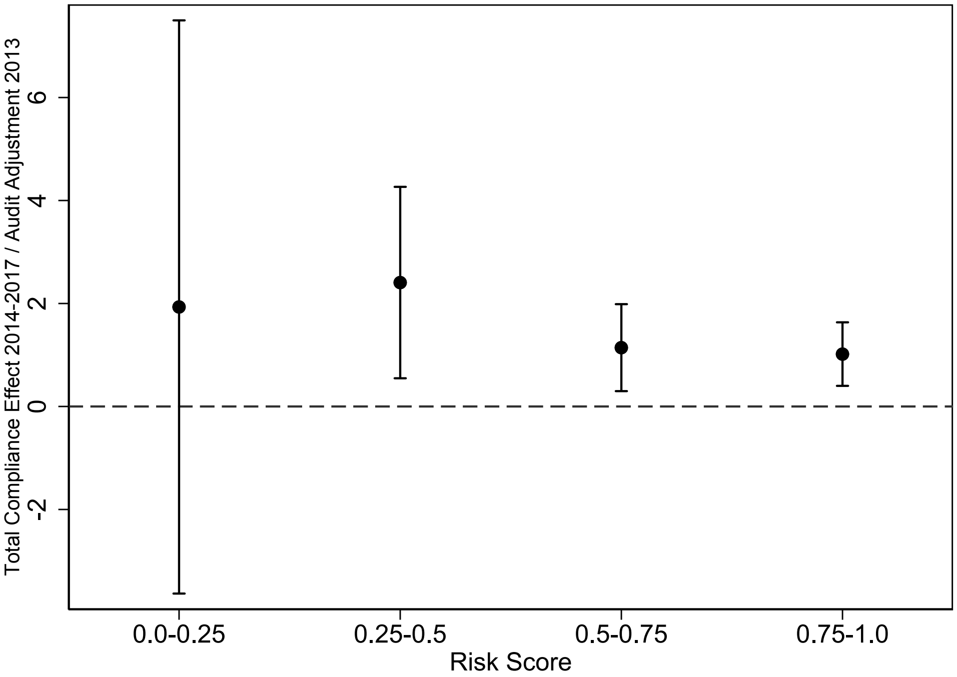
Figure 3. Compliance effects by risk-score. Self-reported deductions.



Note: This figure plots the effect of audit on subsequent self-reported deductions, aggregating over all post-audit years, within different risk-score intervals.

Figure 4 shows that if we consider the effect of adjustment (rather than the the effect of assigned audit) we find less heterogeneity across the risk score in taxpayer compliance. For every NOK that is adjusted by the auditor, the taxpayers in the two highest risk brackets reduced their future self-reported deductions, summing over four post-audit years, by a factor of about two. Even if the estimated compliance scaled by the outcome of the audit is larger in the second risk bracket, there are no significant differences in this “relative” future compliance across the risk distribution. This pattern may shed light on how different risk types respond to obtaining an adjustment of their tax files. For enforcement policies, however, the relevant outcome is the change in taxable income in absolute terms which corresponds to the effect of audit in Figure 1 (see also discussion in section 6 on optimal policies).

Figure 4. The effect of adjustment



5 Evidence from an operational threshold audit

External validity can be low in field experiments and the identified effects may not arise under normal institutions. Since 2015, the strategy of the NTA has been to select individuals for audits using the personal risk scores of that year (based on the machine-learning model estimated on the random 2013 sample, as explained above). For the tax year 2014, the NTA audited all taxpayers (6,500 individuals) with a risk score above a threshold value.[5] We use the following regression discontinuity (RD) model to identify audit effects (comparable to those from the random audit)

$$y_{i,t_0+k} = \alpha_{t_0+k} 1 \left\{ rs_{i,t_0} \geq \overline{rs} \right\} + f(rs_{i,t_0}) + \varepsilon_{i,t_0+k}, \tag{2}$$

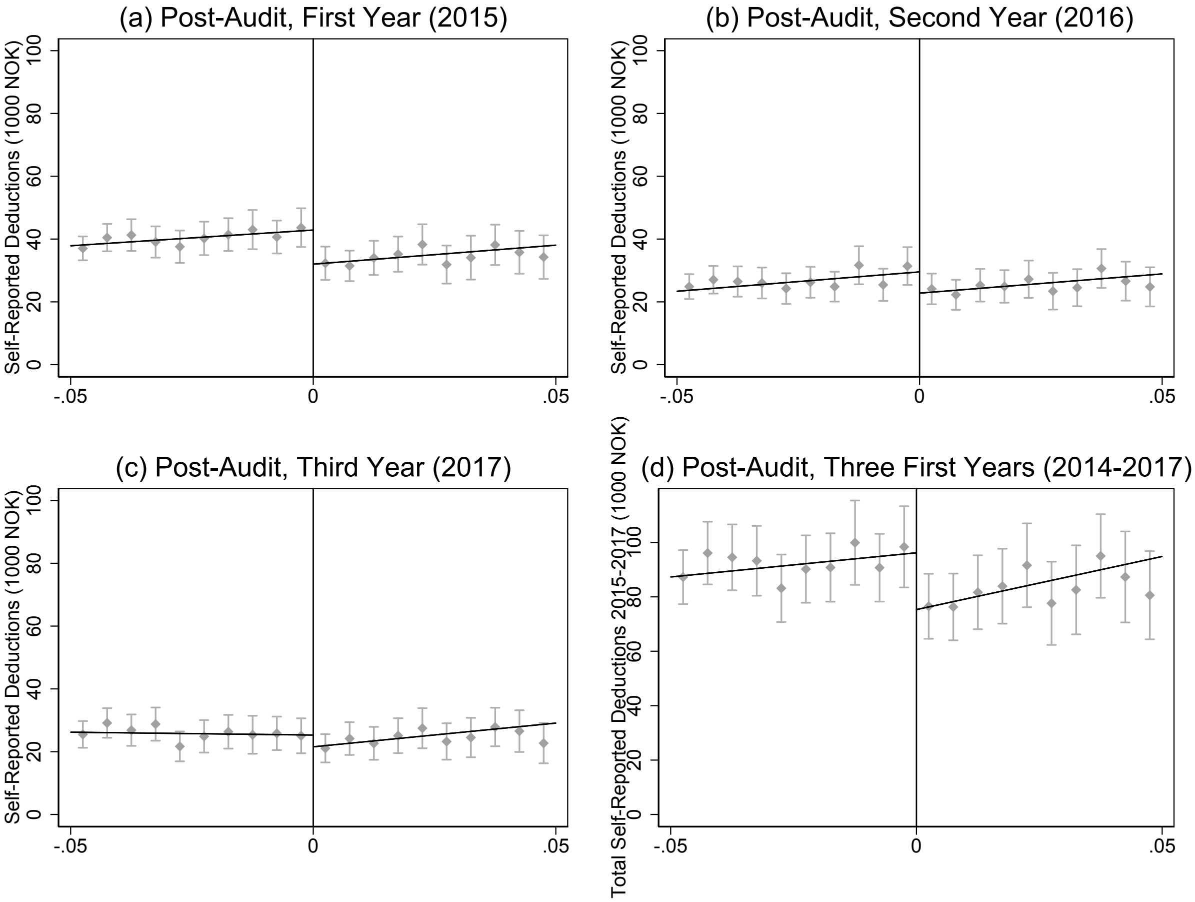
where rs is the forcing variable and every taxpayer with a value equal or above threshold $\overline{rs}$ was audited. The function $f()$ is assumed to be a smooth function. The variable ε_i captures individual unobserved compliance factors. With no discontinuity in the distribution of the error term around the risk score threshold $\overline{rs}$, α can be estimated in a RD model to identify the (local) average treatment effect of an audit.

The RD design requires that the taxpayers are similar, just above and below the threshold. This assumption can be questioned if the individual taxpayer has information and opportunity to manipulate her/his treatment status. If individuals can take actions to cross the threshold, those on either side of the cutoff are no longer similar. This possibility is not a concern in our case. Taxpayers cannot predict the audit risk score threshold because (i) the tax administration selected the threshold based on audit capacity, and (ii) the forcing variable (the risk score) is a complicated composite of almost fifty individual characteristics, which makes it practically impossible for taxpayers to even know their own risk score. In the Appendix, we report different checks that tax- payer characteristics are equal around the threshold. It is also reassuring that the predetermined value of our main outcome variable (self-reported deductions in 2014) moves smoothly across the threshold, and the same is true for self-reported income. The risk score distribution is smooth around the threshold. All in all, the predetermined variables clearly support the assumption that the taxpayers just above and below the threshold are similar.

The compliance effects are displayed by the RD plots in Figure 5 and the RD robust estimates are given in Table 4. Note first that self-reported deductions increases (slightly) with the risk score. In the audit year (2014), taxpayers at either side of the threshold claim the same magnitude of self-reported deductions (Table 4). Turning to the post-audit years, there is a sharp discontinuity in self-reported deductions at the audit threshold, indicating a strong compliance effect of the audit. Even if the compliance effect is enduring, it seems to decline over time and the third year effect is not statistically significant (Table 4).

[5] We combine three waves of the audit; two with a cutoff of 0.82 and the third with a cutoff of 0.92.

Figure 5. Compliance effects of the threshold audit. RD Plots for Self-reported Deductions.



Note: The figures plot self-reported deductions in three post-audit years as well the accumulated deductions against the forcing variable (risk score). The bandwidth is 0.05 around the risk score cutoff point for tax audit, normalized at zero at the cutoff point. The local linear estimators are specified using triangular kernel functions, and ten bins are shown on both sides of the cutoff point.

Table 4. Compliance effects. RD estimates for Self-reported deductions.

	Probability of positive deductions	Deductions in NOK
Audit year (2014)	- 0.0009 (0.003)	- 370 (2 996)
Post audit, first year (2015)	- 0.1157 (0.0265)	-10,858 (2,749)
Post audit, second year (2016)	- 0.0786 (0.0280)	- 6,768 (2,627)
Post audit, third year (2017)	- 0.0573 (0.0286)	- 3,740 (2,575)
All post years		- 20,876 (6,584)

The compliance effects of the operational threshold audit are substantial larger than for the random audit. This is easily explained by the different level of risk scores in the two audits. The relevant comparison would be the highest bracket in the random audit, where the estimated compliance effect over the four years is close to 18,000 NOK, not far from the three-year effect of about 21,000 NOK in the threshold audit. Thus, the evidence from the two audits is very consistent.

6 Optimal risk-based audits

For tax authorities, tax revenue net of enforcement costs is a core criteria for deciding the scale and scope of tax audits (OECD, 2006). If we evaluate audits from a social welfare perspective, we also need to consider private costs (Keen and Slemrod, 2017) and the fact that public revenue may have a shadow price greater than unity. The net social value of audits can then be written as

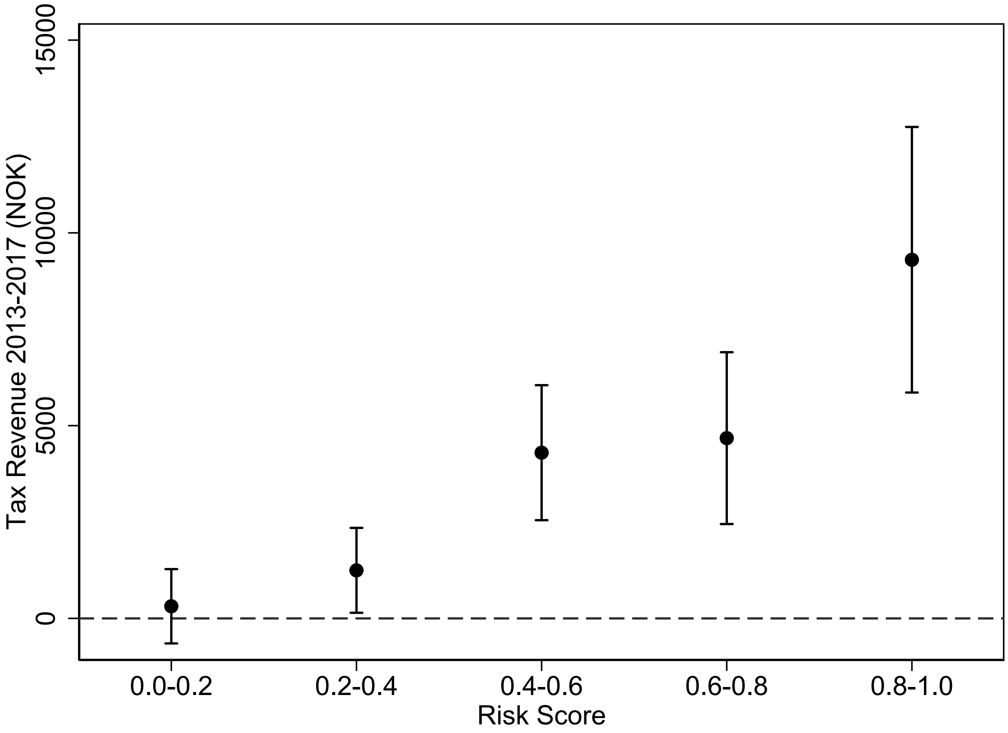
$$\phi(\triangle TaxRevenue - \triangle Administrative\,Costs) - \omega\triangle PrivateCosts$$

where ϕ is the marginal value of public funds and ω is the social welfare cost of taking a unit from a noncompliant taxpayer. Thus, tax revenue minus administrative enforcement costs is a good approximation of the social gain if the (marginal) welfare weights of the noncompliant are low. The private costs are income lost from higher taxes, but also include concealment costs ((Kreiner, 2010; Keen and Slemrod, 2017; Slemrod, 2018), filing effort (Meiselman, 2018) or the moral costs of cheating.

Even with a positive weight on noncompliant taxpayers, a key statistic for optimal tax audits is the tax revenue changes given a marginal expansion of the audit capacity. If audits are random, the revenue effect of a small expansion in audit capacity is by construction constant, and the optimal number of audits is determined by the convexity of the administrative enforcements costs when $\omega = 0$. In practice, tax authorities do *not* randomly select taxpayers for audits in everyday enforcement activities. When marginal audit costs depend only on the number of audits (i.e., taxpayers are perfect substitutes in the audit cost function), the optimal policy rule is separable as long as $\omega = 0$. We can then find the optimal audit capacity as follows. First, rank taxpayers by their net taxable income response to an audit. Given marginal tax rates, this provides the marginal audit tax revenue curve. In our case, evidence in Section 5 suggests that this curve is an increasing function of the risk score. Second, choose the risk score threshold where the marginal net revenue equals the marginal administrative costs, assuming convex audit capacity costs.

In the Norwegian tax system, the ordinary income tax is income net of deductions with a marginal tax rate of 23%. Based on the observed immediate audit outcome and the estimated

Figure 6. Predicted tax revenue from an audit, by risk score.



Note: Tax revenue is the sum of adjusted deductions by the audit and the compliance effects over the four subsequent, multiplied by the marginal tax rate.

compliance we can predict the tax revenue effect of this type of audits. Figure 6 depicts the predicted aggregate tax revenue generated by the audit at different risk scores. The tax revenue generated by an audit increases in the risk score, since both effects (i.e. the immediate outcome and the future reductions in self-reported deductions) are increasing in the risk score.

To estimate how an expansion of this type of office-based audit affects public budgets, we have to compare the marginal revenues with the marginal audit costs. According to the tax authority, the estimated unit cost of conducting a correspondence audit (called "SKD 3002") is just above 1,330 NOK.[6] Comparing costs and revenues, our estimates suggest that audits of tax payers with a risk score above 0.2 would raise net public revenues. Hence, if audit capacity were based on simple cost–benefit analysis, the tax administration should expand this type of audit considerably relative to their current policy. Even if the exact threshold risk score of the 2014 audit is not public, it is above 0.75.

Of course, it is only under special circumstances that the redistributive element of audits (redistributing from the noncompliant to compliers via public budgets) justifies a zero weight on the income taken from the noncompliant (Slemrod and Yitzhaki, 1987). In most cases, it seems reasonable to assign a positive welfare weight to noncompliant taxpayers, especially when noncompliance results from ignorance or confusion because the taxpayer lacks the resources to fully understand the complicated tax system. The more weight the social planner puts on the loss of income for noncompliants, the fewer audits should be carried out. Anyhow, the net revenue effect and how it varies with risk scores is, of course, still a key parameter for deciding optimal audits.

There are also other non-distributive arguments for expanding tax enforcement policies beyond the breakeven point for cost and revenues. Audits may induce concealment costs as well as the moral costs of cheating. In addition, noncompliance breaks the principle of horizontal equity and upholding the principle that individuals with similar income and assets should pay the same tax provides a separate argument for more extensive auditing. In this paper, we focus on the individual preventive effects of audits, but there are also general deterrence effects of audits, as well as potential network effects stretching beyond the spouse of the audited taxpayer that should influence the optimal audit capacity.

[6] Wage cost per day is 4,028 NOK = annual wage costs/working days in a year = 725,000 NOK/180. The norm is three audits per day.

7 Conclusion

The effect of tax audits on future compliance is a key parameter for the design of optimal tax enforcement policy. A novel feature of our study is that we identify the behavioral responses of taxpayers exposed to real-life operational tax audits. The evidence is based on a random audit as well as audits based on risk scores from machine learning methods. While studies of compliance using random audit data typically estimate the effect on the average taxpayer, we study data from random as well as operational threshold audits among wage earners claiming substantial self-reported income deductions. In particular, we estimate heterogeneous compliance effects across tax payers with different risk scores.

The audits revealed a substantial proportion of irregular self-reported deductions among the population of taxpayers. Our main finding is that the audit also affected future behavior and improved compliance among tax payers. Self-reported deductions were significantly lower among the tax payers subject to audit and this effect were still present after four years. Actually, the accumulated drop in self-reported deductions during post-audit years exceeded the amount that was disclosed by the audit. This compliance effect is shown to be increasing in the risk score, mainly reflecting that high risk tax payers are more like to have their self-reported deductions adjusted by the audit. Most of the low-risk tax payers had their filings accepted and did not know that they were selected for audit.

The external validity of our study is high as it estimates the effects of audits that are common and implemented as a part of normal activities of tax authorities in rich countries. The evidence from the operational audit based on a risk score threshold is highly consistent with the estimate from the random audit, indicating external validity of the field experiment. Several of the insights are therefore relevant for the design of tax enforcement policy. For instance, we show that efforts of tax authorities to enhance compliance, or researchers trying to understand it, should not be limited to the self-employed. Even when income of most wage earners are reported by a third party to the tax authorities, and therefore leaves little room for noncompliance, wage earners also have considerable leeway to reduce their taxes through self-reported deductions. Our study also demonstrates that the methods used by modern tax administrations for risk profiling taxpayers can be used to identify policy-relevant behavioral responses to tax audits. Another important policy lesson is that if we include the compliance effects of audits, an expansion of audits in Norway targeted at wage earners with “high” self-reported deductions will generate tax revenues well above the costs of any additional audits.

Finally, the compliance effect of audits in this study can be given alternative explanations. It may arise from a reduction in intentional misreporting (evasion), but also from improved knowledge about the tax rules. This distinction is important for policymakers because the relative weight assigned to the tax revenue that is recouped through audits likely depends on whether noncompliant taxpayers are indeed tax evaders or merely confused about the complex tax rules.

References

Advani, A., W. Elming, J. Shaw, et al. (2017). The dynamic effects of tax audits. Technical report, Institute for Fiscal Studies.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, 47–57.

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.

Fack, G. and C. Landais (2016). The effect of tax enforcement on tax elasticities: Evidence from charitable contributions in france. *Journal of Public Economics 133*, 23–40.

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

Gillitzer, C. and P. E. Skov (2018). The use of third-party information reporting for tax deductions: evidence and implications from charitable deductions in denmark. *Oxford Economic Papers 70*(3), 892–916.

Heckman, J. J. and J. A. Smith (1999). The pre-programme earnings dip and the determinants of participation in a social programme. implications for simple programme evaluation strategies. *The Economic Journal 109*(457), 313–348.

Hodge, R. H., A. H. Plumley, K. Richison, G. Yismaw, N. Misek, M. Olson, and H. S. Wijesinghe (2015). Estimating marginal revenue/cost curves for correspondence audits. In *IRS Research Bulletin, Presented at the 2015 Internal Revenue Service—Tax Policy Center Research Conference.*

Keen, M. and J. Slemrod (2017). Optimal tax administration. *Journal of Public Economics 152*, 133–142.

Kleven, H. J. (2014). How can scandinavians tax so much? *Journal of Economic Perspectives 28*(4), 77–98.

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.

Kreiner, C. T. (2010). Optimal tax enforcement. Unpublished working paper. Center for Economic Behavior and Inequality, University of Copenhagen.

Mazzolini, G., L. Pagani, and A. Santoro (2017). The deterrence effect of real-world operational tax audits.

Meiselman, B. S. (2018). Ghostbusting in detroit: Evidence on nonfilers from a controlled field experiment. *Journal of Public Economics 158*, 180–193.

OECD (2006). Strengthening tax audit capabilities: General principles and approaches. Technical report.

Slemrod, J. (2018). Tax compliance and enforcement. Technical report, National Bureau of Economic Research.

Slemrod, J. and S. Yitzhaki (1987). The optimal size of a tax collection agency. *Scandinavian Journal of Economics 89*(2), 183–92.

8 Appendix

8.1 Stratification of random audits

There are three main categories of deduction items in the Norwegian tax return: Item category 3.2 - Deductions from income from employment etc.; Item category 3.3 - Capital expenses and other deductions; and Item category 3.5 - Special allowances. The tax audits considered in this paper cover 29 specific deduction items from within these three main deduction categories. These 29 deduction items represented the starting point for the stratified sampling of random audits in 2013. Table 10 provides the stratum number, the deduction item and the number of taxpayers from the audited and unaudited groups in the 2013 population, net of the sample restrictions described in Section 2.2. Six of the 29 deduction items were not evident in either the audited or unaudited groups in 2013, and therefore not listed in Table 10. All estimations for the 2013 population are weighted according to the stratum sizes of the audited and unaudited groups in Table 10.

Table 5. Samples by deduction item, 2013 random audit

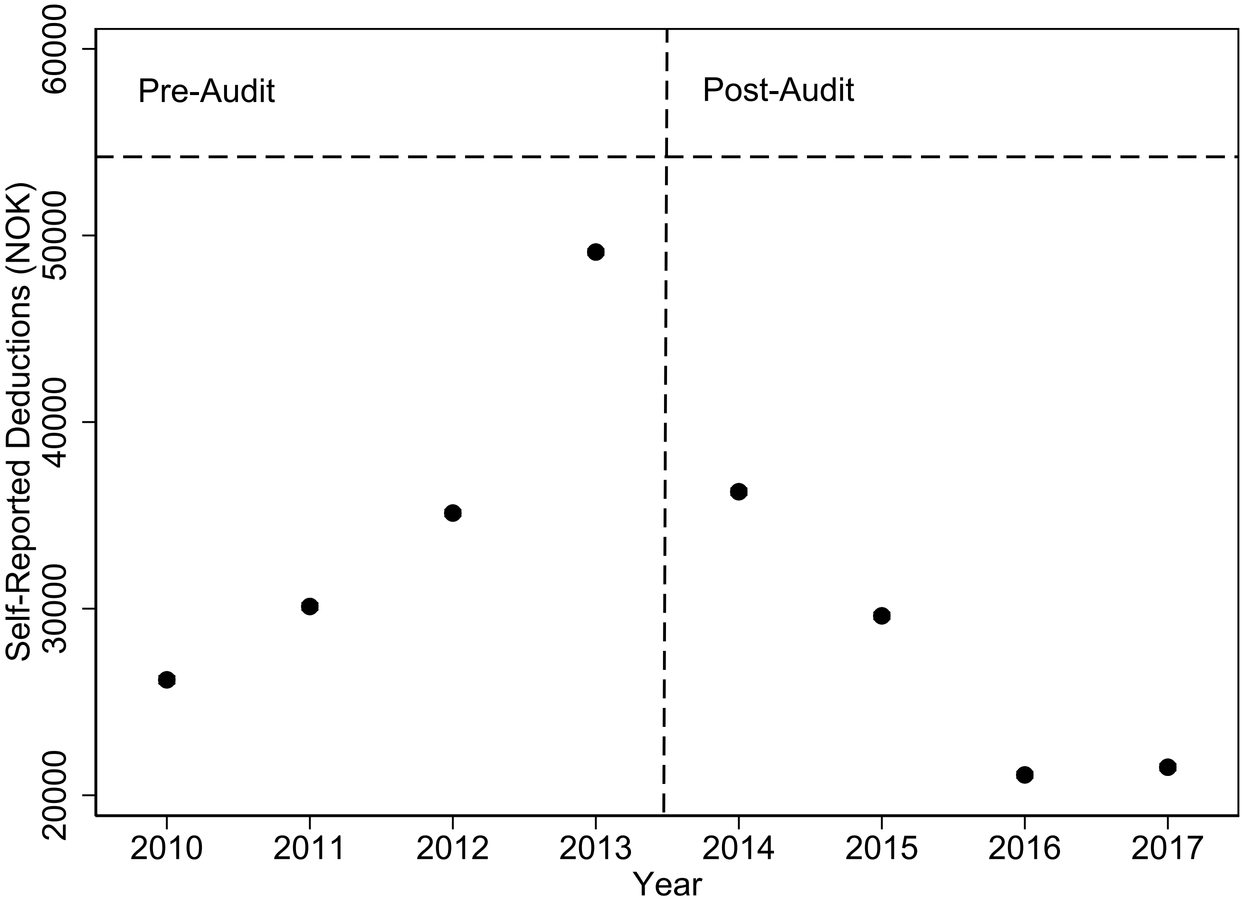
Deduction item	**Item number**	**Audited group**	**Unaudited group**	**Total**
Special allowance - large sickness expenses	3.5.1 / 3.5.4	327	4 864	5 191
Special allowances	3.5.1 / 3.5.3	11	884	895
Interest on debt	3.3.1	3 217	63 834	67 051
Expenses for food and accommodation, work-related stays away from home	3.2.7	1 649	52 708	54 357
Expenses seamen	3.2.7	328	754	1 082
Deduction for travel between the home and work	3.2.8 / 3.2.9	2 740	57 498	60 238
Maintenance payments	3.3.3	240	444	684
Standard deduction for foreign employees	3.3.7	893	38 713	39 606
Other deductions	3.3.7	480	6 864	7 344
Childcare deduction	3.2.10	929	14 855	15 784
Deficit on letting of real property outside business activities	3.3.12	345	5 011	5 356
Deficit unit link	3.3.7	22	11	33
Income deduction from profit and loss accounts	3.3.7	34	12	46
Allowance for minor impairment of earning capacity	3.5.3	47	38	85
Benefits derived from surrendered property	3.3.3	46	86	132
Interest on debt abroad	3.3.2	369	5 510	5 879
Deduction of positive balance	3.3.7	139	96	235
Annual fees for VPS account, safe rental, etc.	3.3.7	39	30	69
Donations to scientific research and vocational training	3.3.7	5	15	20
Deficit on letting of real property outside business activities, from spouse	3.3.12	3	75	78
Donations to voluntary organizations and religious and belief-based communities	3.3.7	6	36	42
Deficit on real property abroad	3.3.12	4	98	102
Deficits carried forward from previous years	3.3.11	133	96	229
Remaining six strata/deduction items with zero taxpayers in either the audited or unaudited groups		5	5	10
Observations		12 011	252 537	264 548

8.2 Formal test of random audit balance

Table 6. Probability of random audit

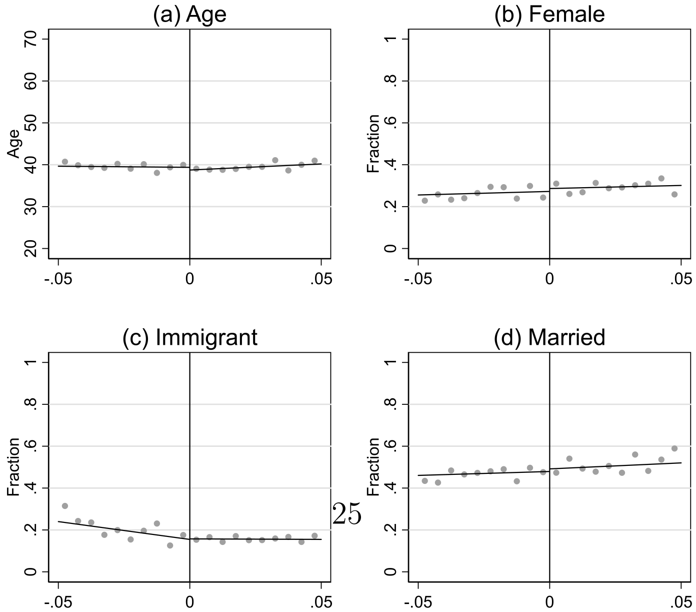
Pre-audit characteristics (2013)	Coeff. (Std.err) of bivariate linear probability models
Self-reported deductions	-518
	(385)
Self-reported deductions > 0	0.0003
	(0.0004)
Taxable income after taxpayer corrections	-1 192
	(3 467)
Prefiled taxable income	-1 284
	(3 338)
Age	-0.447
	(0.120)
Woman	0.008
	(0.005)
Labor migrant	-0.009
	(0.004)
Married	0.010
	(0.005)
Risk score	0.0015
	(0.0022)
Observations	264 548

Figure 7. Time profile of self-reported deductions. Non-audited (control) group.



8.3 Mean reversion

8.4 Operational Threshold Audit RD checks



As expected, Figure 9 depicts that the running variable (the risk-score), is evenly distributed across the audit threshold, showing no signs of being manipulated.

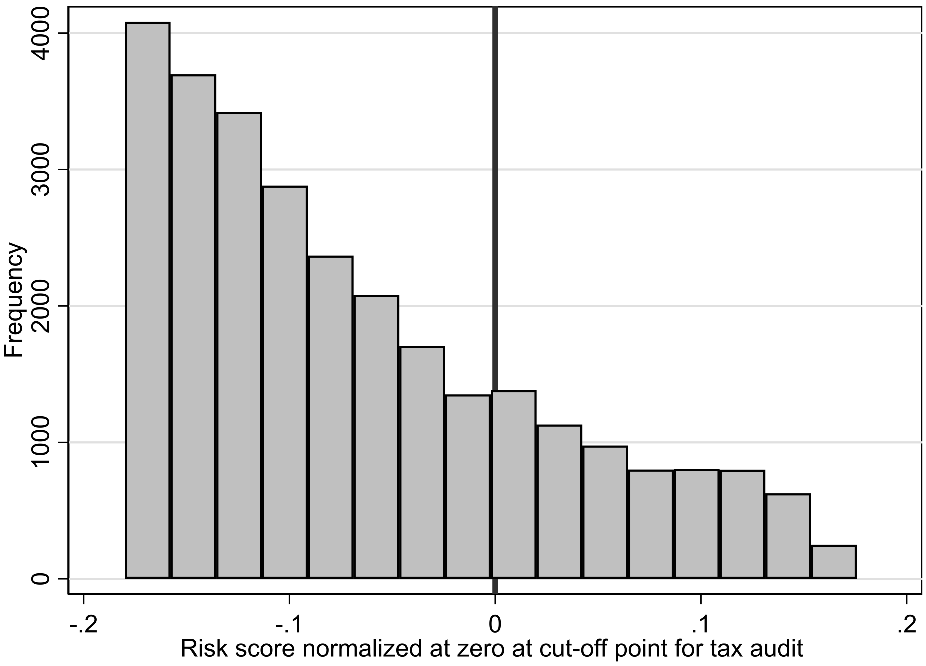
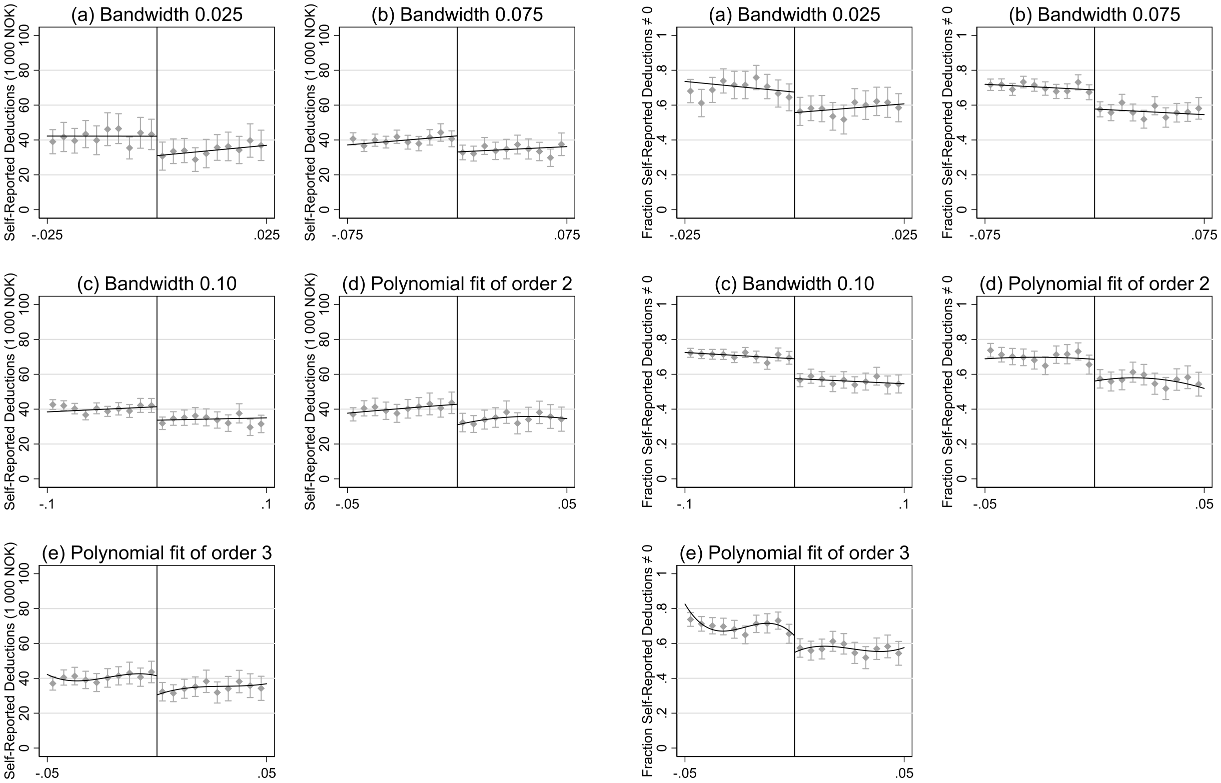


Figure 9. Density check of the running variable

Note: The figure plots the risk score distribution of taxpayers in the 2014 population. The risk score is normalized at zero at the cutoff point for tax audit. The distribution is limited to taxpayers with a normalized risk score ≥ -0.18, which is equivalent to an actual risk score between 0.64183 and 0.99759 (corresponding to the 88th percentile and the maximum score in the population, respectively). Each bin has a width of 0.025 (of the risk score).

Here we present graphical evidence of the robustness of the RD estimates in Section 5. Panel (a) in Figure 8 displays alternative bandwidths and polynomial orders of the proportion claiming self-reported deductions and Panel (b) provides the same for the amounts claimed. The pattern is the same across all specifications; that is, there is a discontinuity in self-reported deductions at the threshold.

Figure 10. Alternative bandwidths and polynomial orders.



(a) Fraction with self-reported deductions

(b) Self-reported deductions