CESifo WORKING PAPERS

7616
2019

April 2019

Compliance effects of risk-based tax audits

Knut Løyland, Oddbjørn Raaum, Gaute Torsvik, Arnstein Øvrum

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7616*
Category 1: Public Finance

Compliance effects of risk-based tax audits

Abstract

Tax administrations use machine learning to predict risk scores as a basis for selecting individual taxpayers for audit. Audits detect noncompliance immediately, but may also alter future filing behavior. This analysis is the first to estimate compliance effects of audits among high-risk wage earners. We exploit a sharp audit assignment discontinuity in Norway based on individual tax payers risk score. Additional data from a random audit allow us to estimate how the audit effect vary across the risk score distribution. We show that the current risk score audit threshold is set far above the one that maximizes net public revenue.

JEL-Codes: D040, H260, H830.

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

Knut Løyland
Norwegian Tax Administration
Oslo / Norway
Knut.Loyland@skatteetaten.no

Oddbjørn Raaum
Frisch Centre and OFS
Oslo / Norway
oddbjorn.raaum@frisch.uio.no

Gaute Torsvik
University of Oslo and OFS
Oslo / Norway
gaute.torsvik@econ.uio.no

Arnstein Øvrum
Norwegian Tax Administration
Oslo / Norway
Arnstein.Ovrum@skatteetaten.no

29 March 2019
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.

1 Introduction

Audits are a cornerstone of tax enforcement policy. They detect and correct noncompliance immediately, but may also change taxpayers' future filing. The behavioral consequences may extend beyond those taxpayers actually audited, through either networks or the general perception of enforcement effectiveness. While the immediate tax revenue effects of audits are mechanical and easily measured, the behavioral effects are not directly observable and must be estimated within a counterfactual framework. This paper estimates the behavioral effects of audits among high-risk wage earners, using two instances where similar taxpayers were randomly selected for audit.

It is worth noting that we address random assignment to audits among high-risk filers, that is, among taxpayers exposed to real-world tax audits. Tax administrations occasionally draw audits randomly from the entire universe of taxpayers to learn the scope and scale of tax evasion. Several studies use this data to estimate the audit effects on future tax compliance (Kleven et al., 2011; Kleven, 2014; Gemmell and Ratto, 2012; Advani et al., 2017; DeBacker et al., 2018a,b). The policy relevance of population average compliance effects of audits is, however, questionable. First, the audited taxpayers are typically informed that they have been selected for a random check, and this may affect their behavioral responses (Slemrod, 2018)[1]. Second, tax administrations increasingly use machine learning to predict risk scores for noncompliance and to select individual taxpayers for audit. It is then unclear whether a population-wide average compliance effect is informative about the behavioral effects of taxpayers selected for standard operational audits.

Evidence from operational risk-based audits provide data for the relevant population, but there internal validity is an issue because compliance effects are difficult to identify. The problem is that 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 postaudit transitory component could mistakenly be interpreted as a behavioral response (Ashenfelter, 1978).[2] Identification of behavioral responses to audits must rely on an element of randomness in the assignment of audits.

This means that the literature asserting the compliance effects of tax audits face a quandary. To inform practical enforcement policy, research should provide estimates of the behavioral effects of risk-based audits, but estimates based on the data from such audits are

[1] This is 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".

typically biased. Random audits for their part can identify the causal effect on compliance, but only for the population included, which often extends beyond those taxpayers that are potential targets for audit. This paper resolves this quandary by using random assignment to targeted, risk-based audits. We can therefore estimate behavioral responses to audits that are both causal *and* informative for actual tax enforcement policy.

In 2013, the Norwegian Tax Administration (NTA) singled out a group of taxpayers with relatively high self-reported tax deductions. From this population of about 350,000 taxpayers, the NTA audited a random sample of 15,000 taxpayers. This was a low-cost office-based audit, similar to what is commonly labelled a correspondence audit (Hodge et al. 2015). We use this 2013 audit experiment to estimate the average compliance effect of the 2014 and 2015 audits. Fortunately for our research purpose, one objective of this audit was to build a machine learning model using an extensive list of taxpayer characteristics to predict noncompliance. In 2014, the tax administration used the model calibrated on the 2013 audit to risk score all taxpayers with high self-reported deductions. They then audited all taxpayers with a risk score above some critical value. Our second identification strategy uses a regression discontinuity method to estimate the local average effect of audits on future compliance for high-risk taxpayers. We estimate two average compliance effects of the audit, (i) the average effect among those who are audited and (ii) the average effect among the audited who had (some of) their self-reported tax deductions rejected. From a policy point of view, the average audit effect is relevant. But in order to understand why the audit effect varies with the risk-score of the audited, we also estimate the average compliance effect among those who had their self-reported deductions adjusted because of the audit.

We provide several important contributions to the literature on tax enforcement policy. First, our analysis identifies the behavioral effects of audits among wage earners. It is often said 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). It is therefore important to measure how audits affect future compliance among those high-risk tax filers that are also wage earners. Second, we can check the responses of the spouses of the audited taxpayers. This is important because diminished compliance for spouses could counteract the benefits of increased compliance for those audited. However, we identify no evidence for such an effect.

Third, we find that the compliance effect of audits increases sharply with the taxpayers' risk score. However, when we estimate the behavioral effect of actual self-reported deductions adjusted by audit, the evidence suggests that the behavioral effect of the adjustment is

independent of the risk score. Fourth, our evidence on the association between risk score and behavior is informative for designing optimal audit strategies. Tax administrations increasingly use big data and predictive analytics to estimate noncompliance risks among taxpayers to target 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 revenue effect depends on the risk score of the audited. In this regard, the elasticity of tax revenue with respect to enforcement input is a key parameter for designing optimal audit policy and should be compared to the marginal enforcement cost (Keen and Slemrod, 2017; Kreiner, 2010).

To estimate this parameter, we need solid evidence on both the short- and long-term revenue effects of audits. Our estimates suggest that behavioral effects are important. By averaging over two years of postaudit filing, we find that the increase in tax revenues from increased compliance is on a par with the amount disclosed directly by the audit. To determine the effect on public budgets, marginal revenues must be compared with audit costs. A central policy question is then how far down the risk score tax authorities should audit (OECD, 2018), and we find that an expansion of correspondence audits among wage earners in Norway would be expected to generate public revenues exceeding the additional costs incurred.

The value of credible empirical estimates of the long-term revenue effects of audits is high because the existing theory does not give much guidance. Even the sign of the compliance effect is contested from a theoretical viewpoint. Further, among those who had their deductions corrected (not approved) by the audit, there is presumably a mix of mistakes and deliberate attempts at tax evasion. We expect audits and any adjustments to improve tax compliance for taxpayers that mistakenly claimed excessive deductions because they now have the correct information. But for those taxpayers claiming excessive deductions to evade income taxes, the canonical model of tax evasion predicts that the effect of an audit today on future compliance depends on how this event affects the perceived probability of being audited in the future (Allingham and Sandmo, 1972). Although it would be natural to assume that those who were caught as evaders will adjust this probability upwards, it could also go the other way.

One mechanism supporting this is the so-called bomb-crater effect (gamblers' fallacy) whereby compliance drops immediately after a taxpayer is audited because the audited taxpayer perceives it as unlikely they will be audited again (Mittone, 2006). This hypothesis has been tested both in the field and in the lab. The evidence from lab experiments emulating tax audits is mixed. Kastlunger et al. (2009) find evidence of a bomb-crater effect among students. Choo et al. (2016) argue that taxpayers and students do not behave in the

same manner in the lab, and that there is no evidence of a bomb-crater effect among actual taxpayers. The studies that use data from random of individual taxpayers audits typically find large positive compliance effects, which are hard to reconcile with the idea that audits today lowers the perceived audit probability for the future (Advani et al., 2017; DeBacker et al., 2018b). A study of firm audits, on the other hand, find patterns that are consistent with the bomb-crater effect of audits (DeBacker et al., 2015).

Elsewhere, Gemmell and Ratto (2012) propose a theoretical model, partly based on an extension by Snow and Warren Jr (2007) of Allingham and Sandmo (1972) to a two-period model where taxpayers update their perceived audit probability in period 2 based on their audit experience in period 1. Gemmell and Ratto (2012) distinguish between audit probability on the one hand and the probability of detection on the other, and drop the assumption that all evaded taxes are detected through audits. They argue that the effect of an audit on the expected audit probability and expected detection rate in period 2 is then conditional on the outcome of the audit in period 1. In brief, the results suggest that taxpayers who evade taxes and are detected through audits improve future compliance because they adjust their perceived probability of being audited and detected upwards. Conversely, taxpayers who evade taxes, but are not detected through audits, seek to evade paying taxes even more in the future because their perceived probability of being audited and detected is reduced. Splitting a sample of randomly audited taxpayers into the detection of evasion (noncompliant or compliant), Gemmell and Ratto (2012) find that noncompliant taxpayers, as detected by the tax authorities in period 1, are more compliant in the following period while the opposite holds for compliant taxpayers.

The structure of the remainder of the paper is as follows. Section 2 reviews some key features of the Norwegian tax system, including tax audits and the data used. Section 3 details the empirical strategy we use to estimate the compliance effects of audits. Section 4 presents the main results and Section 5 discusses how the effects revealed vary over the compliance risk distribution. Section 6 discusses the notion of 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. For wage earners, there are two main tax audit categories, which are both correspondence audits. The first uses flags that are computer generated depending 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. These risk scores depend on aspects of both the tax return and the taxpayer, with taxpayers with the highest risk scores then selected for a tax audit. The number of audits 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, that is, 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 Data

The NTA provides our data on audits and tax returns. We analyze two samples of audited taxpayers, one from income year 2013 and one from income year 2014. In 2013, the NTA singled out a population of about 300,000 taxpayers that claimed 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] We denote this as the 2013 population. 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.

The purpose of the 2013 random audit was to use the data to estimate risk scores of the taxpayers by a machine learning model linking around 50 individual characteristics to noncompliance in the self-reported deductions. Both current and historic tax filing behaviors were used to predict compliance. Since income year 2014, the NTA has used this model to select individuals for a risk score-based audit. In 2014, around 6,500 taxpayers in the population of taxpayers with self-reported deductions above X, having a risk-score above a certain threshold, were audited. [4]

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 (2013 or 2014). 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. Table 2 provides descriptive statistics for the 2013 and 2014 populations of taxpayers with high self-reported deductions on 1 or 2 of the 29 items.

[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).

[4] Comprising three audits, two with a cutoff of 0.82 and the third with a cutoff of 0.92.

Table 2. Summary statistics for the 2013 and 2014 populations.

Variables	**Random audit (2013)**		**Threshold Audit (2014)**	
	Mean	Std. dev.	Mean	Std. dev.
Total deductions	172 001	76 550	174 659	74 351
Third-party reported (prefilled) deductions	121 917	62 562	122 971	61 336
Self-reported deductions	49 091	39 674	50 784	40 272
Total income	563 038	373 589	567 180	367 283
Third-party reported (prefilled) income	553 903	357 812	559 834	354 795
Age	40.3	11.8	40.5	11.7
Female	0.334		0.336	
Immigrant	0.173		0.174	
Married	0.450		0.434	
Risk score	0.381	0.225	0.377	0.224
Observations	264 584		249 187	

3 Empirical strategy

3.1 Estimations

Our empirical analysis consists of different sections, estimating three set of parameters. First, we assess how the random 2013 audit influenced future compliance by estimating 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 variable is aggregate self-reported deductions as this is the variable that defines the population from which the tax authorities randomly audited a subsample. $Audit_{i,t_0}$ is an indicator variable equal to one for taxpayers audited in 2013 and X_{i,t_0} is a set of pretreatment controls that we may add to gain precision. With this specification, β captures the compliance effect. This 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 2013 population of taxpayers with high self-reported deductions.

The audit was a office based correspondence audit. This means that in the vast majority of cases, taxpayers that did not receive an adjustment were not aware that they had been audited.[5] If audit only affects those who actually receive their adjusted tax file (if the

[5]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.

exclusion restriction holds), we can divide β by the fraction adjusted, to obtain an estimate of the average behavioral effect of obtaining an adjustment of self-reported deductions.

Our second estimation exploits the threshold-based selection of taxpayers to audit. In 2014 the NTA used the risk scores from the machine-learning model to decide to audit every taxpayer with a risk score above some threshold value $\overline{rs}$. We estimate the following regression discontinuity (RD) model to identify the effect α

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

where rs is the forcing variable and every taxpayer with a value equal or above $\overline{rs}$ is treated (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 threshold $\overline{rs}$, α can be estimated in a RD model to identify the (local) average treatment effect or risk-based audits around the risk-score threshold.

Finally, we apply the risk score of the machine-learning model for the whole 2013 population. This enables us to estimate how the compliance effect of audits varies across the risk score of those audited. We can then use the RD estimate of the 2014 audit to calculate the revenue effects of audits around this threshold. However, if we wish to design optimal risk-based audits, we need to know how the compliance effects of audits vary over a range of risk scores. To provide this information, we estimate a simple linear model that basically augments (1) with an interaction term between the risk score and audit status:

$$y_{i,t_0+k} = \theta_{t_0+k} rs_{i,t_0} + \delta_{t_0+k}(rs_{i,t_0} * Audit_{i,t_0}) + \beta_{t_0+k} Audit_{i,t_0} + \gamma X_{i.t_0} + \varepsilon_{i,t_0+k} \tag{3}$$

Overfitting may be an issue because the risk score is estimated by means of data from the same 2013 audit. The machine learning model was estimated with the whole random audit sample split into training, validation and test data sets. The training set estimates the model, while the validation set is used to evaluate the model and provide feedback for the fine tuning of the model parameters. Even if the interaction model (3) then should be estimated on the test sample only, we also include the validation set to gain statistical power. However, we also estimate (3) using only the test set (see appendix 8.4). The results are similar, but here precision is compromised because of fewer observations.

3.2 Preaudit balance

With random assignment to risk-based audits, we expect that those who are audited and those who are not are equal across all observable and unobservable preaudit characteristics (for the

RD analysis, we expect similarity to hold around the risk-score threshold that separates the audited from the unaudited). Table 3 provides the preaudit characteristics of the audited and unaudited taxpayers in the 2013 population. As expected, the means are similar for the two groups.

Table 3. Balance by treatment in the 2013 audit.

	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 prefilled 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 key variable for a balance check is the risk score. As noted, the NTA used part of the 2013 sample to estimate a model on a large set of covariates to predict the likelihood that a taxpayer belonging to the 2013 population had deductions not approved and therefore adjusted. The mean risk score is very similar for the audited and unaudited taxpayers in this population. When we estimate the probability of audit, the risk score coefficient is far from significant, see the Appendix for details. Thus, the machine-learning model used by the tax authorities to detect risky behavior cannot predict the audit status of the taxpayer. This indicates that the randomization was successful. Owing to the large sample of unaudited tax payers, age and gender are significant predictors of audit, but the point estimates are negligible.

The RD design requires that the taxpayers just above and below the threshold that splits the 2014 population into audited and unaudited taxpayers are similar. A critical issue for

this estimator is the manipulation of treatment status. If individuals can take actions to cross the treatment threshold, those on either side of the cutoff are no longer similar. This possibility is not a concern in our case because it was impossible for taxpayers to predict the audit threshold because (i) the tax administration selected the threshold risk score based on audit capacity, and (ii) the forcing variable (the noncompliance risk score) is a complicated composite of almost fifty individual characteristics, which makes it practically impossible for taxpayers to know and manipulate their own risk score. As expected, Figure 1 depicts that the running variable (the risk-score), is smooth across the audit threshold and therefore shows no signs of being manipulated.

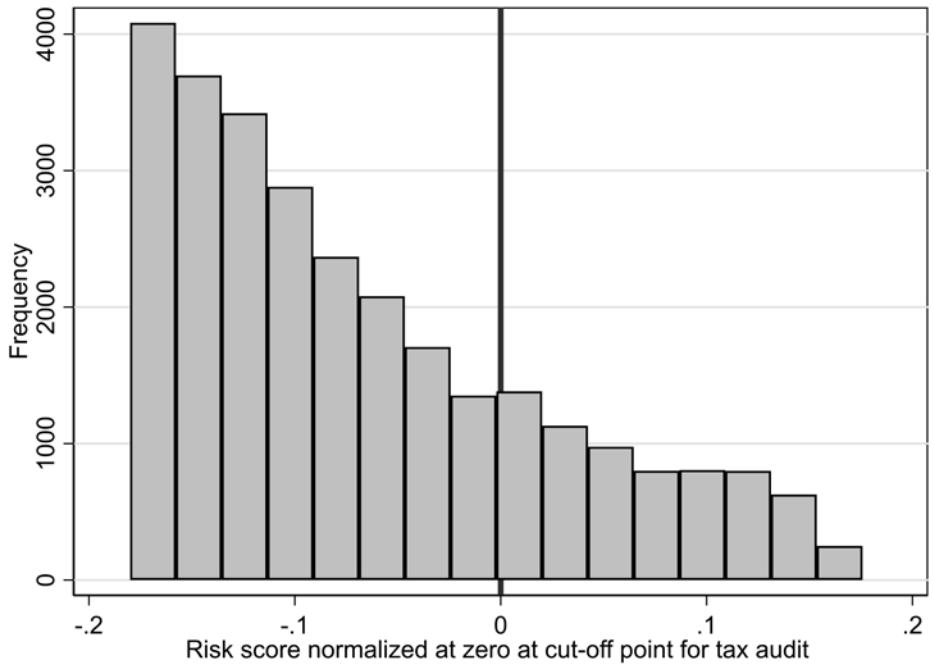


Figure 1. Density check of the running variable

Note: Figure 1 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 in this figure is limited to taxpayers with a normalized risk score $\geq$ -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).

Figure 2 depicts no discontinuity in any predetermined covariates of the taxpayers. It is 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. There are also no discernible discontinuities around the audit threshold in the values of any of the predetermined variables, including self-reported deductions and income, which clearly supports the assumption that the taxpayers just above and below the threshold are

similar.

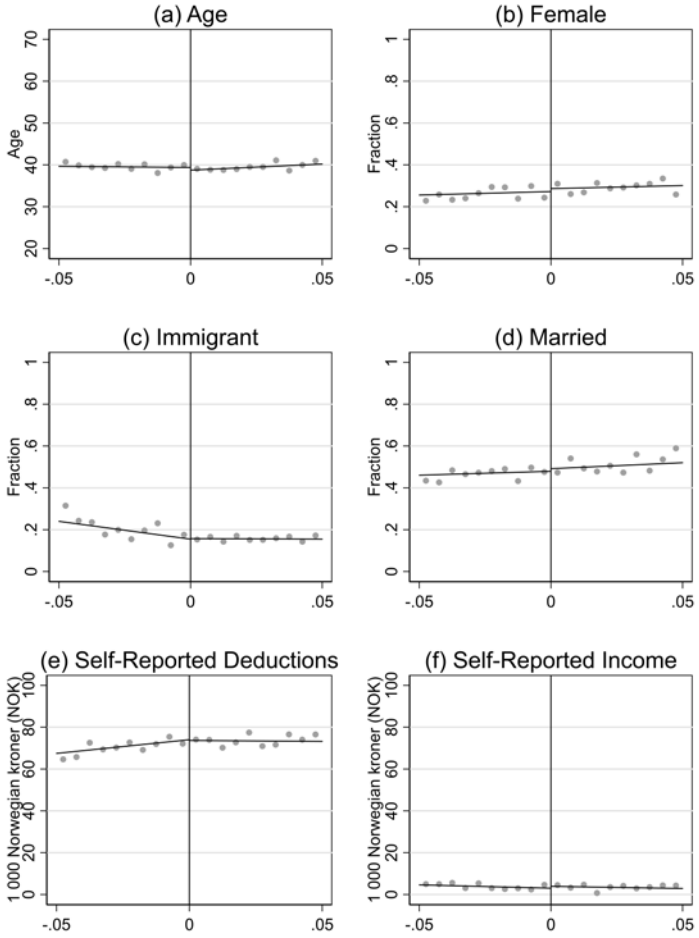


Figure 2. RD plots of predetermined taxpayer characteristics

Note: Panels (a)–(d) provide RD plots for the age and proportion of female, immigrant and married taxpayers in 2014 (the tax audit year) and panel e and f show self-reported deductions and income around the threshold prior to the audit. The bandwidth is 0.05 around the cutoff point for tax audit, which is the risk score 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. The RD estimates for the predetermined characteristics are zero and negligible, see the Appendix for details.

4 Compliance effects of audits

The population of taxpayers that we consider are audited because they self-report income tax deductions above X. We therefore use self-reported deductions as our main outcome when estimating the effects of audit. Before turning to the compliance effects, it is worth noting the strong decline in self-reported deductions among taxpayers in the control group. In 2013 (the audit year), the average self-reported deduction was 49,091 NOK (see Table 2).

This fell to 36,267 NOK in 2014 and further to 29,615 NOK in 2015. The decline in self-reported deductions over time for taxpayers claiming high self-reported deductions but who are, by chance, not audited, illustrates the importance of mean reversion in the population of taxpayers with "risky" filing behavior. It also shows how difficult it can be to identify causal effects of targeted audits without some kind of experimental design. In our study, both the treatment group (those audited) and the control group (the unaudited) will have the same mean-reverting process, and we can therefore identify the causal effect of being audited.

Table 4 presents the effects of the 2013 and 2014 audits on subsequent self-reported deductions. Given our data contain tax-filing data for 2015, we can estimate the impact of the 2013 audit on the filings in 2014 as well as in 2015.

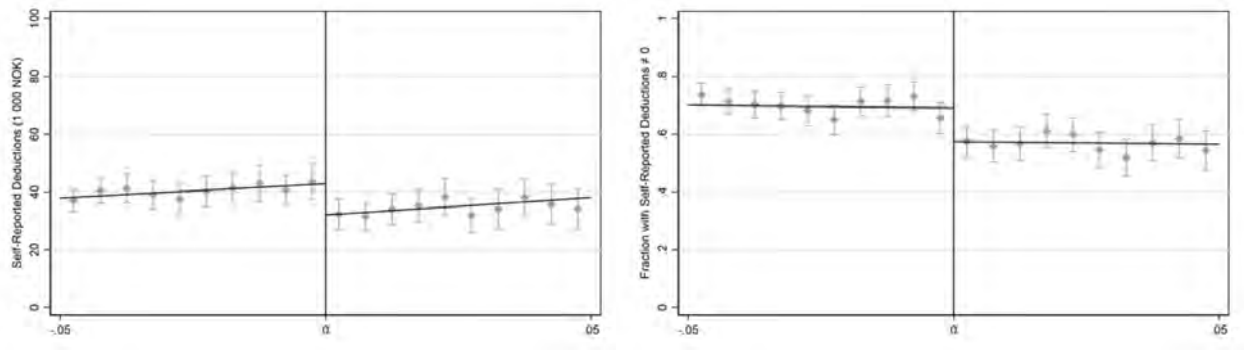
Table 4. The compliance effect of audit on self-reported deductions

	Random audit (2013)		**Threshold audit (2014)**
	2014	**2015**	**2015**
Self-reported deductions	-2 135***	-2 051***	-10 858***
	(447)	(441)	(2749)
Self-reported deductions > 0	-0.018***	-0.014***	-0.116***
	(0.005)	(0.005)	(0.027)
Household self-reported deductions	-2 635***	-2 516***	
	(469)	(462)	
Observations	258 219	252 039	249 187

Note: Figures in NOK, robust standard errors in parentheses. The RD estimates for the threshold audit are robust with respect to bandwidth and nonlinearity, see the Appendix for details.

As shown, the 2013 audit caused self-reported deductions to fall by more than 2,000 NOK in the subsequent year. The compliance effect also endures as the impact on the 2015 deductions is comparable to that of the year before. The random audit also reduced the percentage of taxpayers claiming self-reported deductions by approximately 1.5 percentage points.

The RD plots for the risk score-based audit are depicted in Figure 3. As expected, self-reported deductions increase with the risk score, and we can clearly see the effect of the audit on self-reported deductions given the sharp decline at the risk score threshold. There is a clear drop in both self-reported deductions (in NOK) and in the percentage of taxpayers claiming self-reported deductions.



Note: The left-hand side panel plots self-reported deductions against the forcing variable (risk score) and the right-hand side panel illustrates the percentage of taxpayers self-reporting deductions (those who claim deductions above those that are prefilled). The bandwidth is 0.05 around the cutoff point for tax audit, which is the risk score 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.

Figure 3. RD Plots for Self-reported Deductions

The effect estimate of the threshold audit in column 3 of Table 3 shows that the fall in the amount of self-reported deductions, as well as the reduction in the share of taxpayers choosing to adjust the prefilled deductions, are much higher for the threshold audit than for the random audit. This reflects that more taxpayers had their reported deductions adjusted in the threshold audit, and we discuss this in more detail in the next section.

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 control probabilities when the other spouse is audited. Both mechanisms call for a study of compliance effects that include the filing behavior of the spouse. We keep the sample unchanged and simply add the outcome for the spouse where the audited taxpayer is married or has a cohabitant.

Total deductions will increase when we consider the household, but the contribution from spouses should be identical for both the treatment and the control groups, unless there is a reallocation of items within the household or behavioral effects on the spouse. Reallocation would imply that we should observe effects closer to zero than we do when we consider individual filing data, while behavioral spillovers would contribute to stronger responses. The evidence in (Table 4) suggests that the behavioral spillovers dominate. The estimated drop in self-reported deductions caused by audits strengthens when we include contributions from the spouse.

5 Behavioral responses across different risk scores

The average drop in self-reported deductions caused by the threshold audit is almost five times the fall caused by the random audit. These estimates come from two different populations. The threshold audit is the LATE for taxpayers with a particularly high risk score, while the random audit is the ATE for the population of taxpayers with self-reported deductions above X. Both audits are correspondence audits. It is therefore primarily those taxpayers who had their self-reported deductions adjusted by the auditor who will be aware that they are being audited. The compliance effects reported in Section 4 are average audit effects, but are driven by taxpayers receiving an adjustment that may change behavior and they are more frequent given the higher risk score. Table 5 confirms this where we compare the outcomes of the two audits. As shown, the percentage of taxpayers who have their self-reported deductions adjusted is much higher for the threshold audit than for the random audit. Note that a small share of taxpayers who are not controlled in the audit we study, that is, those who are in the control group in the random audit and below the audit threshold in the 2014 audit, also experience an adjustment in their tax filings (of 5.5 and 12 percentage points, respectively). These taxpayers are selected in other flag-initiated audits (see Section 2.1). While the random audit increased the adjustment probability by 14.9 percentage points, the threshold audit raised it by 51.0 percentage points. In terms of the adjusted amount, both audits increased taxable income considerably.

Table 5. Adjustment rates and amounts, by audit status.

	Random audit (2013)			**Threshold audit (2014)**		
	Audit	Control	Audit effect	Audit	Control	Audit effect
Adjustment rate	0.204	0.055	0.149	0.630	0.120	0.510
Adjustment amount	8 262	3 731	4 531	39 521	12 180	27 341
Observations	12 011	252 537		3 533	2 705	

Note: The figures for the 2014 audit are the average adjustment rate and amount, averaging over those taxpayers with risk scores 0.05 below (the control group) and above (the treatment group) of the audit threshold.

As behavioral responses are restricted to those with an adjustment, we can decompose the average compliance effects of the audit (Table 4), into the probability of obtaining an adjustment times the average behavioral effect among those who had their self-reported deductions adjusted. Thus, the behavioral effect of *adjustment* equals the compliance effect of audits scaled by the inverse of the adjustment rate. This corresponds to an instrumental

variable approach where adjustment is the treatment instrumented by the random audit (Kleven et al., 2011). These estimates are in Table 6, where we can see that the effects of an adjustment are smaller for the random audit than the threshold audit, both in terms of the probability of claiming self-reported deductions (first row) and the amount of deductions (second row). The larger effect for the threshold audit reflects the positive association between the risk score and the adjusted amount. However, if we measure the behavioral effect of adjustment relative to the size of the adjustment as in Kleven et al. (2011), the two audits are remarkably similar (see the final row in Table 6). On average, taxpayers obtaining an audit adjustment in the random audit reduced their self-reported deductions in the following income year by 47% of the adjusted amount disclosed by the audit. In contrast, taxpayers with an audit adjustment in the threshold audit lowered their self-reported deductions by 45%.

Table 6. The behavioral effects of audit adjustment

	Random audit (2013)		**Threshold audit (2014)**
	2014	2015	2015
Self-reported deductions > 0	-0.123***	-0.091***	-0.237***
	(0.031)	(0.034)	(0.053)
Self-reported deductions	-14 354***	-13 721***	-22 212***
	(3 007)	(2 965)	(5 543)
Self-reported deductions as a	-0.472***	-0.448***	-0.453***
share of audit adjustments	(0.102)	(0.100)	0.127)
Observations	258 219	252 039	247 138

Note: Instrumental variable estimates with audit adjustment (0,1) as the treatment and treatment status as the random instrument. The IV estimates for the 2013 random audit are from two-stage least squares regression, while the IV estimates for the 2014 targeted audit are from a fuzzy RD approach. Robust standard errors in parentheses.

To provide further evidence on how the behavioral responses vary with the risk score, we use the predicted risk score from the random audit and estimate equation (3) in Section 3. As explained, to avoid overfitting we use the validation and test samples of the 2013 population (not the training sample). The results are in Table 7.

Table 7. Audit adjustment and compliance effects of audits by risk score

	Audit adjustment	**Self-reported deductions**	
	2013	2014	2015
Audit	3 604***	2 139*	557
	(576)	(1 111)	(1 022)
Risk score	-15 519***	23 870***	23 756***
	(317)	(404)	(388)
Audit*Risk score	-21 221***	-10 560***	-7 122***
	(2 196)	(2 911)	(2 693)
Constant	2 186***	27 176***	20 579***
	(90)	(155)	(147)
Observations	257 898	251 708	245 633

These are the estimates of the linear model in Equation (3). Robust standard errors in parentheses.

A higher risk score is associated with a greater probability of an adjustment in self-reported deductions, even for taxpayers who were not selected for the random audit. The reason is once again that independently of the random audit we examine, other flag-based checks were conducted which also led to the adjustment of deductions. A higher risk score is associated with more flags.

Our main interest is in the interaction term between the audit indicator and the risk score, which is negative and highly significant. A higher risk score is associated with larger adjustments for those audited. Moving from risk score 0.5 to risk score 1, the audit led to an additional downward adjustment in self-reported deductions by more than 10,000 NOK. The compliance effects for the different risk score levels are reported in the second and third columns in Table 7. Consistent with our comparison of the random and threshold (high-risk) audit above, we find that the compliance effect increases with the risk score in terms of lowering future self-reported deductions.

Next, we use the estimates from Table 7 to predict audit adjustment and the increase in self-reported deductions at four levels of risk score. Table 8 provides the results. In the final two columns, we report the behavioral effect of adjustment and it turns out that this is fairly similar across the risk distribution.

Table 8. Adjustment, compliance and behavioral effects by risk score

	Audit adjustment	**Compliance effects of audit**		**Behavioral effects of adjustment**	
Risk score	2013	2014	2015	2014	2015
0.2	639	-626	-865	-15 650	-21 625
0.4	-4 887	-2 087	-2 291	-13 043	-14 318
0.6	-9 130	-4 187	-3 718	-11 630	- 10 328
0.8	-13 374	-6 309	-5 139	- 9 857	-8 029

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 $\omega = 0$. First, rank taxpayers by their net

taxable income response to an audit. This provides the marginal audit tax revenue curve. The evidence in Section 5 suggests that this curve can be expressed as an increasing function of the risk score. Second, choose some 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%. The predicted tax revenue of an audit at different risk scores is presented in Table 9 and follows directly from Table 7.

Table 9. Predicted tax revenue from an audit, by risk score.

	Audit adjustment	**Self-reported deductions**	**Marginal tax revenue**
Risk score	2013	2014-2015	2013-2015
0.2	- 639	-1 591	512
0.4	- 4 887	-4 378	2 131
0.6	- 9 130	-7 805	3 895
0.8	- 13 374	-11 448	5 709

The tax revenues from adjustment upon audit and the aggregate compliance effect (2014 + 2015) are comparable in size. With a longer horizon, the increased compliance would generate more revenues than disclosed by audit. 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 correspondenceaudit (called "SKD 3002") is just above 1,330 NOK.[6] Even if the exact threshold risk score of the 2014 audit is not public, the results of the targeted high-risk audit are comparable to that of the random audit. For the threshold audit, the adjustment is 13,943 NOK and the reduction in future self-reported deductions is 10,858 NOK, neither of which are significantly different from the predictions for a risk score of 0.82 in the random audit. 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. The threshold for the 2014 audit was above 0.82, while we can see from Table 9 that the breakeven risk score is somewhere between 0.4 and 0.2.

Of course, it is only under special circumstances that the redistributive element of audits (redistributing from the noncompliant to compliers) justifies a zero weight on the income taken from the former (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 under-

[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.

stand the complicated tax system. The more weight the social planner puts on the loss of income for noncompliance, the fewer audits should be carried out, but net revenue and how it varies with risk scores is, of course, still a key parameter for deciding optimal audits.

On the other hand, there may be other nondistributive 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.

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 based on risk scores from machine learning methods. While random audit studies typically estimate the effect on future compliance for the average taxpayer, our data contain random audits among wage earners claiming substantial self-reported income deductions. We estimate how risky taxpayers respond to audits and how this response varies with the risk score.

The audits revealed a substantial proportion of irregular self-reported deductions among the population of taxpayers. Our main finding is that the audit effect also extends to the future behavior of taxpayers, with those who did not have all their deductions approved reporting fewer deductions over the two years following the audit. In terms of tax revenue gained, the compliance effect over these two years is on a par with the direct disclosure effect, suggesting that in terms of audits effects and tax revenue, the gain from increased future compliance dominates in the long run.

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. Several of the insights are therefore relevant for the design of tax enforcement policy. For instance, we show that tax authorities attempting to enhance compliance, or researchers trying to understand it, should not limit their attention to self-employed taxpayers. It is true that the income of wage earners is third-party reported to tax authorities, and therefore leaves little room for noncompliance, but wage earners also have considerable leeway to reduce their taxes through self-reporting high income tax 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 targeted at wage earners with "high" self-reported deductions will generate tax revenues well above the costs of any additional audits.

The compliance effect of audits may also arise from a reduction in intentional misreporting (evasion) or 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.

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

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

Choo, C. L., M. A. Fonseca, and G. D. Myles (2016). Do students behave like real taxpayers in the lab? evidence from a real effort tax compliance experiment. *Journal of Economic Behavior & Organization 124*, 102–114.

DeBacker, J., B. T. Heim, A. Tran, and A. Yuskavage (2015). Legal enforcement and corporate behavior: An analysis of tax aggressiveness after an audit. *The Journal of Law and Economics 58*(2), 291–324.

DeBacker, J., B. T. Heim, A. Tran, and A. Yuskavage (2018a). The effects of irs audits on eitc claimants. *National Tax Journal 71*(3), 451–481.

DeBacker, J., B. T. Heim, A. Tran, and A. Yuskavage (2018b). 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.

Kastlunger, B., E. Kirchler, L. Mittone, and J. Pitters (2009). Sequences of audits, tax compliance, and taxpaying strategies. *Journal of Economic Psychology 30*(3), 405–418.

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.

Mittone, L. (2006). Dynamic behaviour in tax evasion: An experimental approach. *The Journal of Socio-Economics 35*(5), 813–835.

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

OECD (2018). *Tax Challenges Arising from Digitalisation a Interim Report 2018*.

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.

Snow, A. and R. S. Warren Jr (2007). Audit uncertainty, bayesian updating, and tax evasion. *Public Finance Review 35*(5), 555–571.

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 10. 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 11. Probability of random audit

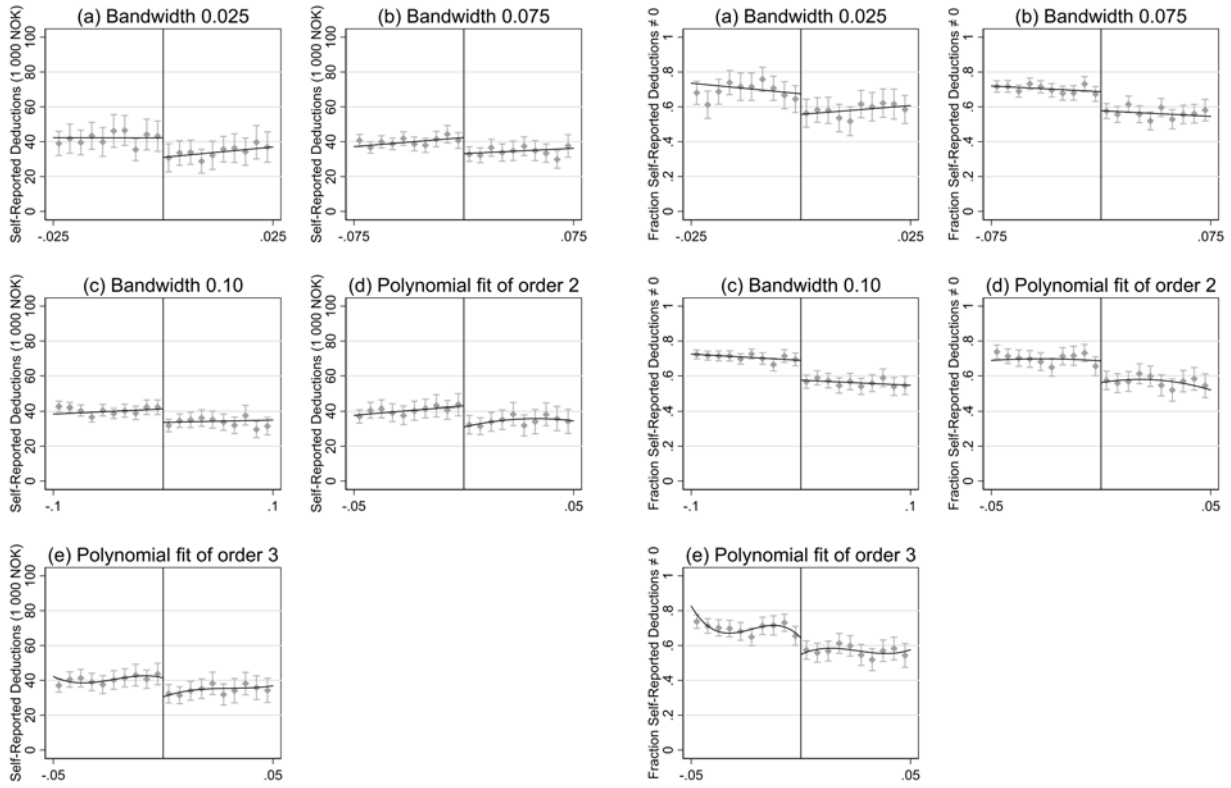
Preaudit characteristics (2013)	Coeff. of bivariate linear probability models
Self-reported deductions	-518
	(385)
Self-reported deductions > 0	0.0003
	(0.0004)
Taxable income after taxpayer corrections	-1 192
	(3 467)
Prefiled taxable income	-1 284
	(3 338)
Age	-0.447
	(0.120)
Woman	0.008
	(0.005)
Labor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

8.3 RD checks

Here we present graphical evidence of the robustness of the RD estimates in Section 4. Panel (a) in Figure 4 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 4. Alternative bandwidths and polynomial orders

(a) Fraction with self-reported deductions **(b)** Self-reported deductions



8.4 Results for the test sample

Table 12. Audit adjustment and compliance effects of audits by risk score. Test sample.

	Audit adjustment	**Self-reported deductions**	
	2013	2014	2015
Audit	4 015***	2 374*	522
	(935)	(1 608)	(1 515)
Risk score	-15 519***	23 870***	23 756***
	(317)	(404)	(388)
Audit*Risk score	-23 869***	-9 157**	-6 635
	(3 503)	(4 377)	(4 068)
Constant	2 186***	27 176***	20 579***
	(90)	(155)	(147)
Observations	254 910	248 783	242 760

These are estimates of the linear model in equation (3). Robust standard errors in parentheses.