Check for updates

Accepted: 27 September 2023

DOI: 10.1111/ecca.12499

ORIGINAL ARTICLE

Economica LSE

# Evaluating compliance gains of expanding tax enforcement

**Knut Løyland**[1] | **Oddbjørn Raaum**[2] | **Gaute Torsvik**[3] | **Arnstein Øvrum**[1]

[1]Norwegian Tax Administration

[2]Ragnar Frisch Centre for Economic Research and NFS

[3]University of Oslo and NFS

**Correspondence**
Oddbjørn Raaum, Ragnar Frisch Centre for Economic Research, Oslo 0349, Norway
Email: oddbjorn.raaum@frisch.uio.no

**Abstract**

This paper demonstrates how tax administrations can evaluate future compliance gains from risk-based tax enforcement that audits all taxpayers above a risk threshold. Expanding tax enforcement in this setting means reducing the audit threshold. The compliance gains from such an expansion consist of a mechanical audit correction effect and a behavioural effect that reflects changes in self-reporting in the subsequent years. We estimate this behavioural effect in a regression discontinuity analysis with the risk score as the forcing variable. We find that taxpayers at the margin had a significant reduction in self-reported deductions in the next years' tax filing. The behavioural effect over a three-year post-audit period is estimated to be of a magnitude similar to that of the direct adjustment of the audit. This compliance effect does not change when we include the reporting of the spouse. We find that the risk score threshold that maximizes net public revenue from the audits is considerably below current practice.

## 1 | INTRODUCTION

Operational tax audits are typically risk-based, targeted towards filers predicted to be non-compliant. Although the most obvious and immediate impact of audits is the detection and correction of non-compliance on the spot, audits may also change subsequent tax filing behaviour of audited (specific deterrence) and non-audited (general deterrence) taxpayers. In this paper, we focus on the specific deterrence effects of tax audits.

To optimize tax audits, it is essential to take into account the long-term behavioural responses to such tax enforcement policies. However, obtaining this information can be difficult. In contrast to the immediate disclosure and revenue effects of tax audits, effects on subsequent tax filing cannot be measured directly. Estimates of behavioural effects must be based on a counterfactual, but a suitable control group can be hard to define since audit selection will often be based on criteria that are unobservable for outside analysts.

We use data from an operational audit to estimate the response of high-risk filers to being audited. The traditional way of targeting audits is to flag taxpayers with suspicious features of specific items in their tax files. Today, it is becoming more common for tax administrations to use advanced predictive analytics to risk score taxpayers (OECD 2019). These methods typically use a large array of individual characteristics, including past filing behaviour, to predict the likelihood of non-compliance with tax rules. Our data come from Norway, where the tax administration risk scores all personal taxpayers who self-report substantial income tax deductions. All taxpayers with a risk score above a threshold are subsequently audited. With this assignment procedure, audits are as good as random around the threshold, and this provides a powerful strategy to estimate how being audited causally affects future compliance. Several years of post-audit observations enable us to estimate how the compliance effect of being audited evolves over time.

The behavioural effect that we identify with this regression discontinuity (RD) framework is 'local', that is, valid for taxpayers around the threshold. The fact that RD design estimates a local average treatment effect can be a limitation when treatment effects are heterogeneous. In our case, however, the local estimate around the threshold contains exactly the information that tax administrations need in order to decide whether they should adjust the risk threshold for audit assignment.

Our analysis shows that being audited leads to a considerable drop in self-reported deductions in the years following the audit. The effect declines over time, but is significant also in the second year after the audit. Summing over three post-audit years, the average drop in self-reported deductions is about 2600 USD, or 23,476 Norwegian kroner (NOK), which is of the same order of magnitude as the direct adjustment of the audit. We also investigated the effects of multiple audits and spillover effects on the spouse of the audited taxpayer. The evidence suggests that the effect of being audited more than once is negligible, and we find no indication of spillover effects on the tax filing of the spouse of the audited taxpayer.

We contribute to the literature on long-term compliance effects of tax audits; see Andreoni *et al.* (1998) and Slemrod (2019) for a general discussion and an overview of this literature. Our study demonstrates how tax administrations can evaluate future compliance gains from risk-based tax enforcement in a setting where taxpayers above a pre-specified risk threshold are audited. Future effects of audits can then be identified by using an RD analysis with the risk score as the forcing variable. Population-wide randomized audits have been used to identify effects on future compliance (Kleven *et al.* 2011; Gemmell and Ratto 2012; DeBacker *et al.* 2018; Advani *et al.* 2023), but it is less expensive and more practical to use data from operational audits if audit assignments are based on risk scores. The operational RD also provides the effect for the marginally audited taxpayers.

To our knowledge, we are the first to combine data from operational audits of high-risk filers with a credible empirical strategy to identify behavioural effects of being audited.[1] A strand of this literature uses traditional flag audits and matching techniques to find a suitable control group for the audited (Beer *et al.* 2020; Mazzolini *et al.* 2022). However, it is difficult to find a credible comparison group since taxpayers are typically selected for such audits based on suspicious filing patterns, often reflecting a transitory drop in taxable income. For audited taxpayers, taxable income will therefore typically be mean-reverting, and it is difficult to isolate this mechanical effect from the behavioural effect of audits (Ashenfelter 1978; Heckman and Smith 1999).

In the final section of the paper, we go beyond the current threshold and ask how far down on the risk score the tax administration should go in order to maximize the tax revenue. We use data from a random audit to locate the risk score where the marginal tax income from an extra audit, taking into account both short- and long-term compliance effects, equalizes the marginal audit costs. We find that the risk score threshold that maximizes net public revenue from audits is considerably below current practice. This analysis contributes to the literature on optimal tax enforcement (Keen and Slemrod 2017).

Closing the gap between the taxes that people owe and the taxes that they actually pay can be an efficient and equitable way to raise government revenue (Sarin and Summers 2020). Since more tax enforcement resources can reduce this gap, our methods and findings are of considerable policy relevance and should have practical interest for tax administrations. Although predictive risk score analysis combined with threshold assignment to audits has become increasingly common in modern tax enforcement (OECD 2019), we have not seen any study using such data to estimate the behavioural effects of operational audits. Such strategies are called for by, for example, OECD (2021), and our study illustrates how operational audit data can be used to estimate the long-term effects of audits. The external validity of our study is strengthened by the similarity between the audits that we study and correspondence audits in other countries. The substantial positive long-term compliance effects of operational audits that we identify should be of relevance and interest to other tax administrations.

The structure of the paper is as follows. Section 1 reviews key features of the Norwegian tax system, including tax audits, the risk score model based on random audit data, and the risk-based threshold audit. In Section 2, we present the data. Section 3 explains the empirical strategy used to estimate compliance effects. In Section 4, compliance effects are discussed and main results are presented. Extensive evidence of robustness is provided in Section 5. Section 6 discusses the optimal audit capacity, and Section 7 concludes.

## 2 | INSTITUTIONAL SETTING

*Taxes, tax filing and audits in Norway*

Norway had about 5.2 million inhabitants in 2015, of which 79% were liable to pay taxes and file a tax return. The administration and enforcement of personal taxation in Norway are divided between the Norwegian Tax Administration (NTA) in assessing personal taxes based on gross and taxable income, and the municipalities that are responsible for direct tax collection.[2] Norwegian income tax differentiates between income from work ($Y$) and capital ($I$). For wage earners, taxes also depend on deductions ($D$). Liable taxes ($T$) are given by $T = t(Y, I, D)$, with $0.22 < \partial T/\partial Y < 0.47$ and $\partial T/\partial I = -\partial T/\partial D = 0.22$, such that the marginal tax on wage income is higher than for interest and other capital income.

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 depicts the timeline of tax returns for personal taxpayers. Information on relevant items in year $t$ is filed during April and May in year $t + 1$. For wage earners, employers report taxable income and withhold taxes to the NTA. For the self-employed, there is no third-party reporting of labour income. Many itemized tax income deductions, such as donations to charitable organizations, interest paid on bank loans, realized losses in share values, etc., are reported by third parties, but some are self-reported by the taxpayer.

Based on third-party information, the NTA prepares and distributes tax returns 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 30 April (for the self-employed, the deadline is 31 May). The difference between the *total* income or deductions in the final tax return and those in the *pre-filled* version is what we label as *self-reported* in this analysis. For example, a taxpayer who has pre-filled income tax deductions of 60,000 NOK in terms of bank loan interest payments, but a total claim of 90,000 NOK in deductions, has self-reported deductions of 30,000 NOK.

Tax audits are carried out from May to December following the income year. Our study concerns audits of itemized income tax deductions. Taxpayers are selected for such audits in two different ways, based on either computer-generated flags that depend on some specific feature of deduction items, or the risk profiles of individual taxpayers. In the latter case, taxpayers are risk scored, and audits targeted towards those at the high end of the score. In our data, the risk score is generated by a machine learning model, which is described in the next subsection.

**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 | Pre-filled tax returns distributed | NTA | Income by source, deductions, gross wealth, debt |
| April–May | Check, correct and self-report if relevant | Taxpayers | Acceptance of pre-filled or self-reported income and deductions items |
| May–December | Checks (standard, automatic) Audit | NTA | |
| | | | Documentation |
| | Final assessment | NTA | Taxable income and wealth, sanctions |

*Risk score threshold audits*

In 2013, the NTA implemented a random audit to collect data to build a model to predict non-compliance. The NTA singled out a population of approximately 310,000 taxpayers who filed self-reported deductions above a threshold of $Z$ NOK on one or two items from a list of 29 specified expenses. The target population constitutes 7% of the whole population of personal taxpayers of 4.4 million individuals. The most common self-reported income tax deduction items are interest on debt, personal work-related expenses on stays away from home, including foreigners who commute to Norway to work, childcare deductions, and deficit from letting out property. From the target population of 310,000 taxpayers, around 15,000 taxpayers were selected randomly for a thorough check of the claimed deductions, and 17% were found to be non-compliant.

These audits were carried out as ordinary operational audits of itemized income tax deductions. The audits are standard low-cost office-based audits, commonly labelled 'correspondence audits' (Hodge *et al.* 2016). The audit checked 'suspicious' itemized tax deductions and asked for taxpayer documentation if necessary. The auditors did not check other items in the tax report. For example, they did not check income reporting. Taxpayers would be notified if the auditor found irregularities in the claimed deductions. Hence every taxpayer who had their deductions adjusted by the NTA knew that they had been audited. Those who were audited without disclosed irregularities would not know about the audit, unless they were asked for additional documentation.[3]

When the taxpayer is found to be non-compliant, the NTA shall consider issuing a pecuniary penalty on top of the audit adjustment. The penalty is 10-40% (normally 20%) of the tax advantage that is detected and corrected during the audit. In this study, we are unable to separate the compliance effects of the audit adjustment itself from any additional penalties.[4]

The NTA used the random audit data to train and test a gradient boosting machine learning algorithm to predict a binary classifier of compliance/non-compliance. When training and building the machine learning model, random audits drawn from the target population were divided into a training (8168 taxpayers), validation (2972 taxpayers) and test (3717 taxpayers) dataset.

The binary classifier was the outcome in the final model that used 29 individual characteristics, including current and past tax filings, as predictors. After running the model on the training data and tuning hyperparameters on the validation data, the model was confronted with test data and the performance was good.

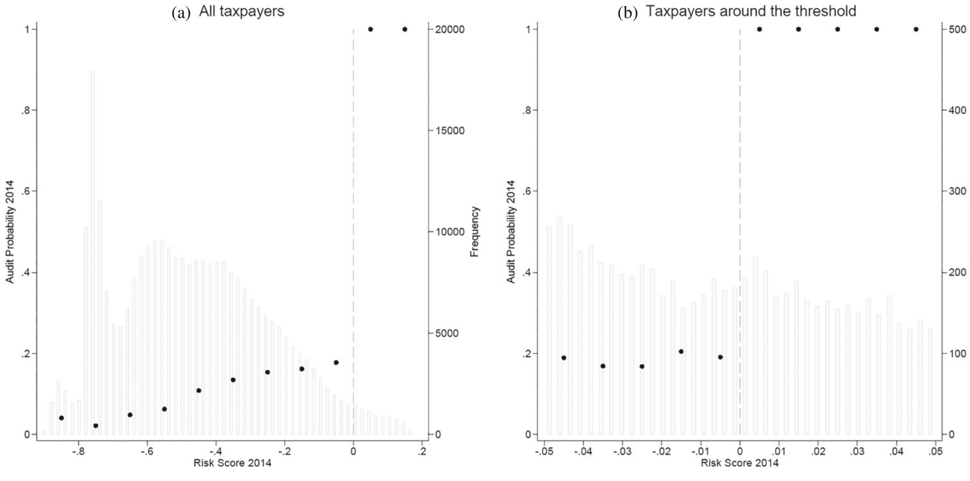


**FIGURE 1** Audit probability and risk score distribution. *Notes*: (a) The risk score and the audit probability across the entire risk score. (b) Zoom in on our estimation sample. The NTA selected taxpayers for audits in three different waves for the 2014 audit. In the first two, conducted in May and June, the thresholds were basically identical, with risk score 0.8. In the July audit, only 800 taxpayers were selected for an audit, and the risk score threshold was set at a higher level (0.92). Our analysis is based on data from all three waves, and we have normalized the thresholds to 0.

Since 2013, this prediction model has been used to select taxpayers for audits. Each year, the model assigns a risk score to every taxpayer in the target population. All taxpayers with a risk score above a threshold are selected for audit. The risk score threshold for audits has varied over the years, mainly due to variation in the budget allocated to this type of audit. In the first year when the machine learning model was employed, for the income year 2014, resources allowing for about 8000 audits were allocated from the management of the tax returns. In this year, 71% of the audited taxpayers were found to be non-compliant. Thus the share of non-compliance was 54 percentage points higher with the risk score audits in 2014, than with the random audits in 2013.

## 3 | DATA DESCRIPTION AND SAMPLE RESTRICTIONS

Our main analysis is based on data from the 2014 audit because it provides the longest post-audit period to estimate the persistence of compliance effects. Taxpayers submitted information for 2014 to the NTA over a three-month period (May to July 2015). During this period, the NTA selected individuals for the risk score audit in three waves, 8000 taxpayers in total. The risk score threshold is determined by the resources allocated to audits each year; a larger audit budget implies that more taxpayers in the target population can be audited, and this means a lower risk score threshold. Thus there is no room for the auditors to influence the cut-off level. In the first two waves, conducted in May and June, the threshold was basically identical, 0.82. In the July wave, the risk score threshold was set at a higher level (0.92). Our analysis is based on data from all three audits.[5]

Figure 1(a) displays the distribution of the 2014 risk score in the entire group of taxpayers with self-reported deduction above $Z$ NOK, together with the probability of audit (indicated by dots). The threshold audit focused on a minority as less than 3% of the taxpayers with self-reported deduction above $Z$ NOK were selected. Figure 1 also reveals that a small fraction of taxpayers below the threshold are audited in flag-based audits. This fraction increases in the risk score. Just below the threshold, one in five taxpayers were audited based on single-item flags. The important observation for our identification strategy is that the audit probability makes a distinct leap

TABLE 2 Pre-audit Characteristics—2014 Threshold Audit

| | Target population | | All taxpayers | |
|---|---|---|---|---|
| | Mean | S.D. | Mean | S.D. |
| Variables | (1) | (2) | (3) | (4) |
| *Income and deductions* | | | | |
| Pre-filled deductions | 131,006 | 66,748 | 102,205 | 64,119 |
| Self-reported deductions | 54,033 | 50,880 | 4745 | 28,221 |
| Taxable income | 432,576 | 369,926 | 286,710 | 283,841 |
| *Individual characteristics* | | | | |
| Age | 41 | 12 | 47 | 19 |
| Female | 0.33 | | 0.49 | |
| Immigrant | 0.16 | | 0.06 | |
| Married | 0.44 | | 0.42 | |
| Risk score | 0.11 | 0.32 | | |
| Observations | 286,488 | | 4,377,723 | |

*Notes:* Taxable income is from the pre-audit year 2013, and deductions are for 2014, before the audit was conducted.

to unity at the threshold. In Figure 1(b), we zoom in on the bandwidth (±0.05) used in the estimations.

The sample includes taxpayers between 17 and 70 years of age at the time of audit. To avoid estimates driven by outliers, all variables in NOK (income and deduction variables) are trimmed such that observations above the 99th percentile are dropped. Columns (3) and (4) of Table 2 provide background information on the whole population of about 4.4 million Norwegian personal taxpayers. Columns (1) and (2) provide the corresponding information for the target population. All monetary values in the text, tables and figures are measured in 2015 prices.

## 4 | EMPIRICAL STRATEGY

We use the jump in audit probability at the risk score threshold in a fuzzy RD regression to estimate the compliance effects of being audited. The estimator combines the two equations

$$y_{i,t_0+k} = g(rs_i) + \pi_1 1(rs_{i,t_0} \geq \overline{rs}) + \beta_1 y_{i,t_0} + u_{i,t_0+k}, \qquad (1)$$

$$Audit_{i,t_0} = f(rs_i) + \pi_2 1(rs_{i,t_0} \geq \overline{rs}) + \beta_2 y_{i,t_0} + \varepsilon_{i,t_0}, \qquad (2)$$

where $y_{i,t_0+k}$ is the tax filing variable of interest measured $k$ years after the audit. The risk score $rs$ is the forcing variable, and $g$ and $f$ are unknown functions that are continuous at $\overline{rs}$. The only pre-audit variable that we control for in the outcome regression is self-reported deductions filed just before the audit ($y_{i,t_0}$). The effect of the audit on future tax filing behaviour (compliance) is found by dividing the effect of the risk score threshold on future tax filing behaviour ($\pi_1$) by the effect of the risk threshold on the probability of being audited ($\pi_2$). This is a standard IV scaling of the reduced-form effect of a random treatment. Our strategy is based on the assumption that the compliance effect is the same for threshold and flag audits. This is justified since the taxpayers were not informed about the audit selection mechanism, and the same protocol was used whether the taxpayer was selected by the risk score or single items in a flag audit. All estimates are based on the robust RD estimator developed by Calonico *et al.* (2014).

The critical identifying assumption for our analysis is that taxpayers just above and below the audit threshold would have had the same expected future tax filing behaviour had it not been for the audit. This assumption could be violated if taxpayers were able to manipulate their own risk score or if there are other reasons for discontinuities in taxpayer characteristics around the threshold. In our case, there are strong reasons to expect balance across the threshold. Most taxpayers are not aware that they are risk scored, and even if they knew, it would be impossible for them to affect their audit status by manipulating their own risk score, which is based on a complex prediction model that loads on a large set of individual characteristics. Moreover, the exact threshold chosen by the NTA for each audit is determined by the budget allocated to this type of audit, and does not relate to expectations about taxpayers' future filing behaviour. As expected by these arguments, predetermined taxpayer characteristics are balanced around the threshold. Figure 2 reveals no discontinuities at the audit threshold for characteristics such as gender, age, marriage, immigrant background and risk score. The pre-audit deductions and income are also balanced around the threshold, as shown by Figure 3.

With annual tax audits, some taxpayers will be audited multiple times over the years. This complicates the estimation of behavioural responses. The future behaviour of a taxpayer who experienced an audit in a given year may depend on both previous and future audits. With long-lasting audit effects, one concern is that the 2014 audit is unbalanced with respect to previous audits. Approximately 20% of the taxpayers close to the risk score threshold were selected for the same type of audit in the previous year, as shown in Figure 3. Reassuringly, there is no discontinuity at the 2014 risk threshold.

A visual inspection of the distribution of taxpayers around the threshold in Figure 1 indicates a bump just above the audit threshold. The McCrary test also rejects continuity of the risk score density at the threshold.[6] However, for the reasons explained above, it is unlikely that this elevation in the distribution is caused by taxpayers manipulating their own risk score. Furthermore, if manipulation were possible, then we should expect bunching below and not above the threshold; why would taxpayers, if they could, manipulate themselves into an audit? Thus we find it highly unlikely that the (small) bunching just above the audit threshold leads to any selection bias in our estimate of audit effects.

## 5 | COMPLIANCE

This section presents our main results of how audits influenced post-audit tax compliance. Before we present the behavioural effects of tax audits, it is useful to consider the direct audit adjustment as a benchmark. The estimated effect of being audited on the probability of an adjustment of self-reported deductions is 0.47 (see Table 3). The average adjustment in self-reported deductions is −24,546 NOK (≈ 2700 USD). Thus the average claimed adjustment was 52,225 NOK (= 24,546/0.47).

Although these adjustments reflect the reporting prior to the audit, Table 4 reports the behavioural effects of an audit on future self-reported deductions. We report the effect for each of the three post-audit years, as well as the aggregated effect over all three post-audit years. We can see that the audit had a substantial impact on future filing behaviour. In the first year after the audit (column (2), panel A), the drop in self-reported deductions caused by the audit is estimated to be 14,345 NOK (or about 1590 USD). This corresponds to 19% of the self-reported deductions claimed by taxpayers with a risk score just below the threshold, or 58% of the average audit adjustment. Although the future compliance effect tapers off rather quickly, we do find a negative and significant effect on self-reported deductions in the second year. In the third post-audit year, the estimate is also negative, but not statistically significant. Adding up the first three post-audit years, we find that the behavioural effect of the audit on self-reported deductions is 23,476 NOK, that is, of the order of the same magnitude as the direct audit adjustment reported in Table 3.[7]

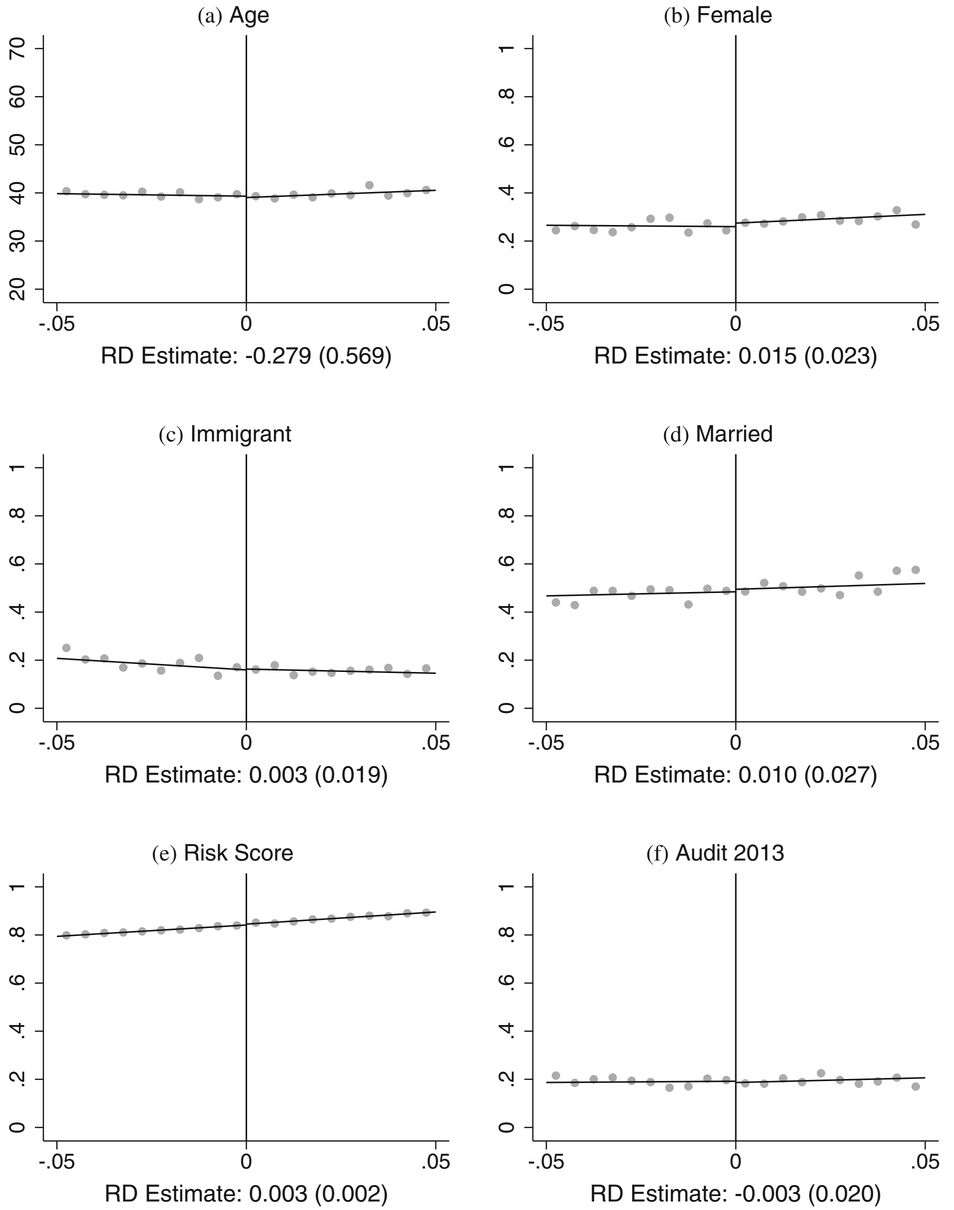


**FIGURE 2** Pre-audit balance around the audit threshold—demographics and risk score. *Notes*: Estimates are based on the robust RD estimator, with fixed bandwidth 0.05 around the risk score threshold (normalized at zero). The local linear estimators are specified using triangular kernel functions. The sample sizes are affected by trimming the one percentile tails of the annual distributions. Age in years; female, immigrant, married, and previous audit in 2013 are shares.

The effects of the audit on self-reported deductions are illustrated in Figure 4, where we display graphs for the three years after the audit. These graphs reveal a distinct reduction in self-reported deductions around the threshold, but the drop is smaller and more blurred over time. Compared to Advani *et al.* (2023), our post-audit compliance effects are greater but more short term. We find compliance effects that are on a par with the direct audit effect. In Advani *et al.* (2023), compliance effects linger for five post-audit years, but in terms of taxable income, the aggregate behavioural response constitutes only around 60% of the direct audit effect.

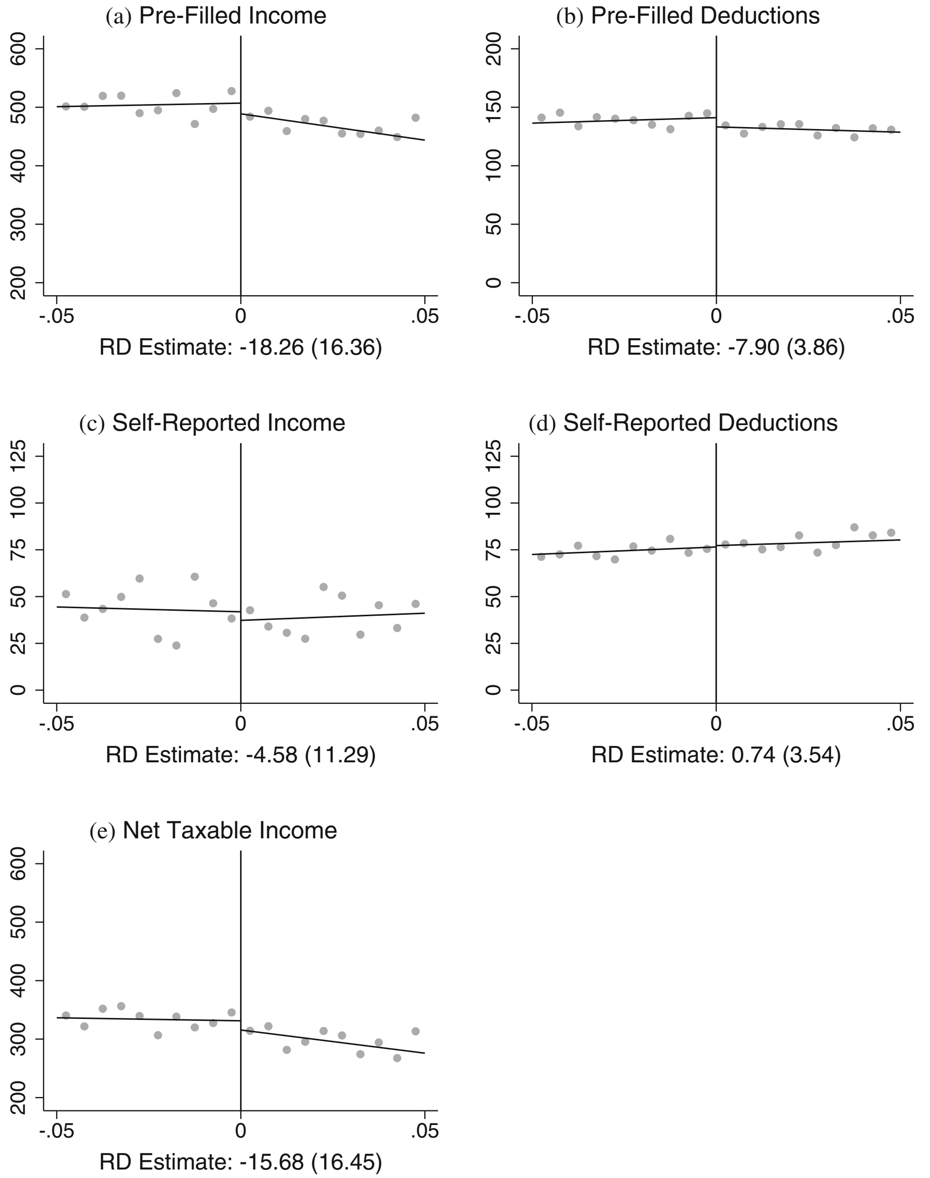


**FIGURE 3** Pre-audit balance around the audit threshold—deductions and income. *Notes*: Estimates are based on the robust RD estimator, with fixed bandwidth 0.05 around the risk score threshold (normalized at zero). The local linear estimators are specified using triangular kernel functions. The sample sizes are affected by trimming of the one-percentile tails of the yearly distributions. Income and deductions are in 1000 NOK (exchange rate NOK/USD is around 9).

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 an audit. Spouses may also update their knowledge about tax rules or audit probabilities when their partner has been audited. Both mechanisms make it potentially important to include the filing behaviour of the spouse in the estimation of future compliance effects of audits. In panels B and C of Table 4, we focus on married taxpayers and households. The compliance effects of married taxpayers are very similar to those of the full sample, but the effect estimates are less precise because of fewer observations.

**TABLE 3** Audit Adjustment by the NTA of Self-reported Deductions

| | Average adjustment | Probability of adjustment |
|---|---|---|
| Audit | −24,546*** | 0.470*** |
| | (2828) | (0.0219) |
| Observations | 6997 | 6997 |

*Notes:* Unconditional average adjustment in NOK. Estimates are based on the robust RD estimator. The bandwidth is 0.05 around the risk score threshold (normalized at zero). The local linear estimators are specified using triangular kernel functions. The sample sizes are affected by the trimming of the one-percentile tails of the yearly distributions. Standard errors in parentheses.
*, **, *** indicate significance levels $p < 0.05$, $p < 0.01$, $p < 0.001$, respectively.

**TABLE 4** Compliance Effects on Self-reported Deductions

| | Pre-audit | | Post-audit | | All post |
|---|---|---|---|---|---|
| | 2014 | 2015 | 2016 | 2017 | 2015–17 |
| | (1) | (2) | (3) | (4) | (5) |
| *A. Audit* | 741 | −14,345*** | −7299* | −2799 | −23,426** |
| | (3544) | (3513) | (3301) | (3567) | (7757) |
| Observations | 6997 | 6915 | 6842 | 6749 | 6960 |
| *B. Married taxpayers* | 5705 | −12,421*** | −6886 | −6333 | −25,646* |
| | (5976) | (5002) | (4962) | (4481) | (11,086) |
| Observations | 2557 | 2528 | 2526 | 2508 | 2548 |
| *C. Married households* | 1953 | −10,591* | −10,871* | −6136 | −26,136* |
| | (5778) | (5316) | (5235) | (4778) | (11,571) |
| Observations | 2390 | 2329 | 2330 | 2321 | 2383 |
| *D. Wage earners* | 2536 | −14,467*** | −7299* | −3723 | −24,497** |
| | (3605) | (3436) | (3300) | (3209) | (7803) |
| Observations | 6339 | 6277 | 6207 | 6127 | 6312 |
| *E. Self-employed* | −14,395 | −15,256 | −10,218 | 6106 | −16,162 |
| | (14,511) | (10,045) | (17,977) | (13,073) | (35,437) |
| Observations | 658 | 638 | 635 | 622 | 648 |

*Notes:* Estimates are based on the robust RD estimator. Control for pre-audit self-reported deductions as in equation (1). The bandwidth is 0.05 around the risk score threshold (normalized at zero). The local linear estimators are specified using triangular kernel functions. The sample sizes are affected by the trimming of the one-percentile tails of the yearly distributions. Standard errors in parentheses.
*, **, *** indicate significance levels $p < 0.05$, $p < 0.01$, $p < 0.001$, respectively.

Total self-reported deductions increase when the partner's items are included, but in the absence of audit spillovers to spouse, the audit effect on future compliance should be of the same magnitude as for individual filings. If the audited person reallocates some of the deductions to the spouse, then we expect lower compliance effects for households than for the audited person. On the other hand, with behavioural spillovers, we should expect a larger drop in self-reported deductions at the household level than for individual taxpayers. Since the effect estimates are very similar based on household deductions (panel C of Table 4), the evidence indicates minor reallocations and spillovers within the household.

It is often pointed out that non-compliance in terms of under-reporting is first and foremost a problem among the self-employed since they self-report their income to the tax authorities (Kleven *et al.* 2011). For itemized income tax deductions related to personal income, wage earners have, however, the same opportunity to evade taxes as the self-employed. An interesting question

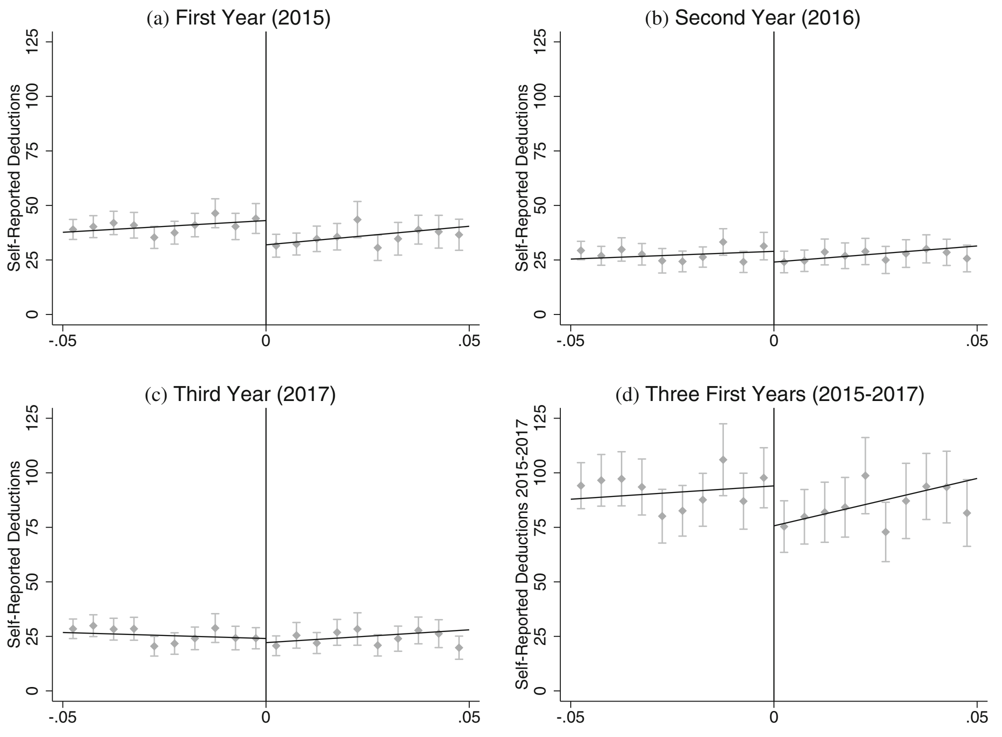


**FIGURE 4** Compliance effects—post-audit self-reported deductions. *Notes*: Plots of self-reported deductions (in 1000 NOK, exchange rate NOK/USD about 9) for each of the three post-audit years as well the accumulated deductions against the forcing variable (risk score). The bandwidth is 0.05 around the risk score threshold (normalized at zero). The local linear estimators are specified using triangular kernel functions, and ten bins are shown on both sides of the threshold. Error bars show the 95% confidence limits.

is if the self-employed are of a different stature than wage earners with respect to how they respond to tax enforcement policies. Since our data include personal tax files for both wage earners and self-employed, we can check if the compliance effects for the self-employed are different from the response of the wage earners. A comparison of panels D and E in Table 4 reveals that the point estimates for the self-employed and wage earners are about the same. Since the data contain relatively few self-employed, the compliance estimates for these taxpayers are imprecise.

## 6 | ROBUSTNESS AND EXTENSIONS

In this section, we assess the robustness of our results with respect to alternative placebo thresholds and empirical specifications. We also discuss effects on income and taxes, possible implications of multiple audits, and results from the 2015 audit.

*Placebo tests and specifications*
As a first robustness test, we estimate the model for alternative (placebo) thresholds, and the results are summarized in Figure 5. We find it reassuring that the only threshold for which we find a significant compliance effect is the one where the audit actually took place. For the placebos, the estimated effects are close to zero.

In our main specification, we control for pre-audit values of self-reported deductions. In Table A1 of the Appendix, we present the estimates without pre-audit controls (panel A). The

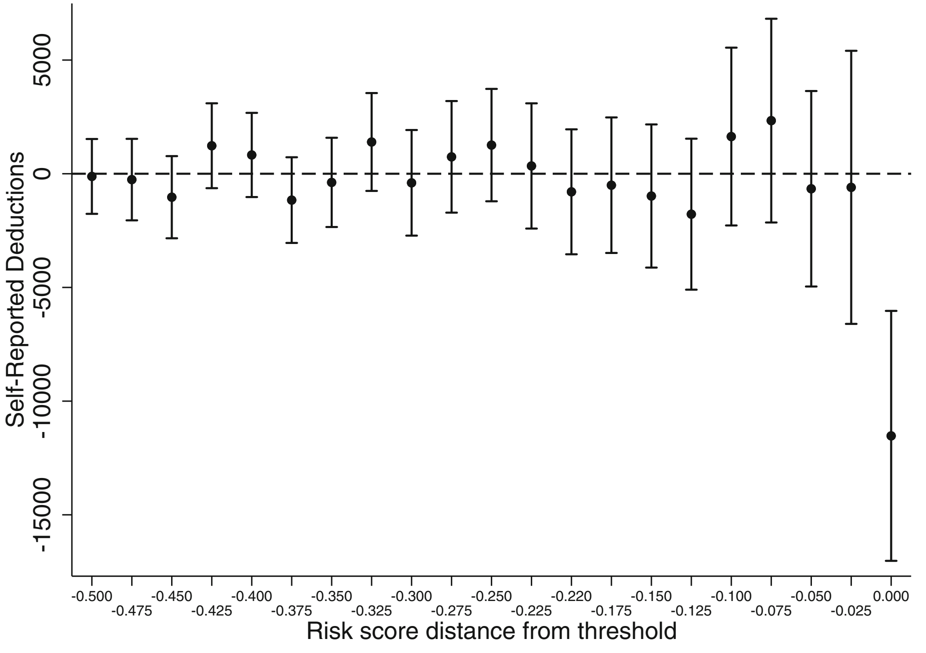


**FIGURE 5** Placebo tests—first post-audit year compliance effects at different placebo thresholds. *Notes*: RD robust estimates of the effect on self-reported deductions (NOK) in the first post-audit year. Separate estimates for different placebo thresholds. Error bars show the 95% confidence limits.

results are very similar. Another robustness check is to ignore flag-based audits below the threshold. This corresponds to a situation where all the differences in outcomes can be attributed to the threshold audit. The effects are estimated with a sharp RD, and the results are presented in panel B of Table A1. As expected, the effects are slightly smaller but not qualitatively different from our main specification. The final check in Table A1 (panel C) addresses the low-risk threshold in wave three and shows that the results are very similar if we drop the individuals at both sides of the threshold in wave three.

As a final set of specification checks, we report estimates with different bandwidths, polynomial orders and kernel functions. As seen in Figures A1 and A2 in the Appendix, non-linearity of the risk score does not matter, and the estimates are robust with respect to the choice of bandwidth and kernel function.

*Income and tax effects*

Taxpayers who got their self-reported deductions adjusted in the audit may respond not only by reducing self-reported deductions but also by adjusting their income and/or how they report it. In Table 5, we first check whether the audit affected subsequent gross income (panel A). There are no signs of effects on total income before deduction. When we run the regression with self-reported net income as the outcome, we obtain point estimates that are of the same order of magnitude as for self-reported deductions (with opposite sign); see panel B of Table 5. This indicates that taxpayers' behavioural response to being audited was confined to the compliance effects that we have seen in self-reported deductions. However, note that the income effects are not precisely estimated. The reason is that transitory components are more important for income than for deductions. The standard deviation of pre-audit self-reported net income is more than three times larger than the standard deviations for self-reported deductions. For taxable net income, which includes third party and self-reported items, the standard deviation is even greater. Given the large

**TABLE 5** Effects on Income and Taxes

| | Pre-audit | | Post-audit | | All post |
|---|---|---|---|---|---|
| | 2014 | 2015 | 2016 | 2017 | 2015–17 |
| *A. Gross income* | −22,838 | −8649 | −1647 | 942 | 5480 |
| | (17,854) | (11,769) | (14,351) | (16,043) | (38,174) |
| Observations | 6997 | 6915 | 6842 | 6749 | 6960 |
| *B. Self-reported net income* | −5322 | 19,465* | 15,458 | 4629 | 37,423 |
| | (11,760) | (9182) | (10,045) | (11,352) | (25,776) |
| Observations | 6997 | 6915 | 6842 | 6749 | 6960 |
| *C. Taxable net income* | −15,678 | 6245 | 5371 | 6120 | 26,160 |
| | (16,456) | (11,530) | (14,049) | (15,374) | (36,146) |
| Observations | 6997 | 6915 | 6842 | 6749 | 6960 |
| *D. Taxes paid* | 10,033** | 4943 | 2872 | 3487 | 13,565 |
| | (3476) | (4087) | (4522) | (5127) | (12,621) |
| Observations | 6997 | 6915 | 6842 | 6749 | 6960 |

*Notes:* Gross income is the sum of the reported and self-reported income of the third party before deductions. Self-reported net income is self-reported income minus self-reported deductions. Taxable net income is pre-filled income plus self-reported income minus pre-filled deductions minus self-reported deductions. Taxes paid include annual income and wealth tax. Standard errors in parentheses.
*, **, *** indicate significance levels $p < 0.05$, $p < 0.01$, $p < 0.001$, respectively.

variability of income, we do not have statistical power to detect effects in self-reported income or taxable income. For taxes paid, there are additional items, such as wealth, that contribute to variation and reduced power.

### *Multiple audits*

In light of the long-lasting effects of being audited, one might be concerned that the estimated effects reflect past audit experiences. We find, however, that the pre-treatment audit exposure (in 2013) is balanced around the 2014 threshold, as shown in Figure 2. Even with a balance in previous audits around the 2014 audit threshold, multiple audits may affect the interpretation of our estimates. Since the audit leads to a drop in self-reported earnings over the next few years, one might expect that a lower exposure to a future audit will limit the persistence of compliance effects. The timing of the tax filing implies that a 2015 audit could not impact the first-year compliance effect of the 2014 audit, simply because the self-reporting on 2015 items happened before any audit. In Figure 6, we see an indication of a decrease in the probability of audit for 2015 around the 2014 threshold. But the impact on the future audit is tiny and not statistically significant (see the Figure 6 Notes for details). Therefore, our estimated audit effects on tax filing behaviour for 2016 and 2017 are unlikely to be affected by subsequent audits. Finally, consecutive audits could be complements or substitutes. If mainly the first audit matters, then the results in Table 4 would underestimate the compliance effect of being audited for the first time in 2014, since a fraction of the control group had a first time audit in 2013. We can assess to what extent our main estimates are influenced by multiple audits by estimating the effects of the 2014 audit conditional on the 2013 audit status. The results in Table 6 show that the compliance effects are slightly higher for those without audit in 2013, compared to the entire sample (Table 4). This suggests that the compliance effect of an additional audit is lower than the effect of the first audit.

### *The 2015 threshold audit*

The NTA followed the same audit threshold procedure in 2015. However, compared to the 2014 audit, the NTA expanded the capacity and doubled the number of audits. This expansion implied

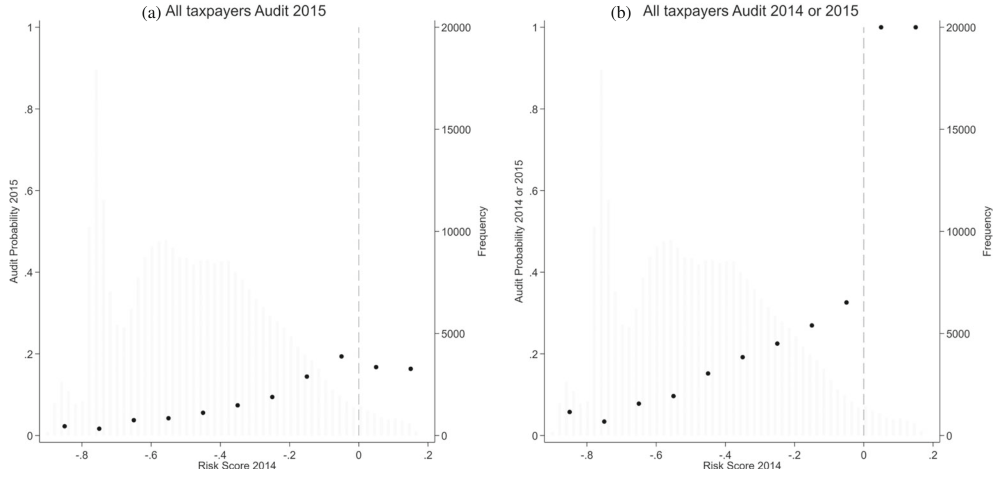


**FIGURE 6** Subsequent audit (2015) by risk score (2014). *Notes*: (a) and (b) depict the 2014 risk score distribution (as in Figure 1) and the audit probability in the subsequent year (2015), with (b) zooming in on our estimation sample. The RD robust estimate of the effect of audit in 2014 on the probability of audit in 2015 is −0.0313, with 95% confidence interval (−0.069, 0.006).

**TABLE 6** Compliance Effects on Self-reported Deductions–first Time Audit

| | Pre-audit | | Post-audit | | All post |
|---|---|---|---|---|---|
| | 2014 | 2015 | 2016 | 2017 | 2015–17 |
| Audit, without 2013 audit | 3088 | −18,635*** | −10,368** | −2148 | −30,511** |
| | (3097) | (3886) | (3475) | (3316) | (8290) |
| Observations | 5449 | 5189 | 5303 | 5238 | 5425 |

*Notes:* These estimates are based on the robust RD estimator. We control for pre-audit self-reported deductions. The bandwidth is 0.05 around the risk score cut-off point for tax audit, normalized at zero. The local linear estimators are specified using triangular kernel functions. Standard errors in parentheses.
*, **, *** indicate significance levels $p < 0.05$, $p < 0.01$, $p < 0.001$, respectively.

a distinct jump in the audit probability at a lower threshold (see Figure A3 of the Appendix). With a lower audit threshold, we expect to find a lower compliance effect of the audit, since both the probability of non-compliance and the self-reported deductions of taxpayers increase in the risk score. As expected, Table 7 shows a smaller first-year compliance effect for self-reported deductions compared to the 2014 audit, and the second-year effect is not statistically significant.

# 7 | OPTIMAL AUDIT THRESHOLD

For tax administrations, tax revenue net of enforcement costs is a natural criterion to decide the scale and scope of tax audits (OECD 2006). Evaluated from a social welfare perspective, however, tax revenues from audits are to a large extent transfers from private individuals to the government. The net social value of the marginal audit can therefore be written as

$$\phi(\Delta\text{Tax revenue} - \Delta\text{Administrative costs}) - \omega\ \Delta\text{Private costs},$$

where $\phi$ is the marginal value of public funds, and $\omega$ is the social welfare cost of taking one unit of income from a non-compliant taxpayer. In addition to reducing private after-tax income,

**TABLE 7** Compliance Effects of the 2015 Audit

| | Pre-audit | Post-audit | | All post |
|---|---|---|---|---|
| | 2015 | 2016 | 2017 | 2016–17 |
| *A. Audit adjustment by NTA* | −25,797*** | | | |
| | (1198) | | | |
| *B. Self-reported deductions* | 2119 | −6283*** | −3516 | −9979** |
| | (2183) | (1852) | (1950) | (3335) |
| Observations | 15,011 | 14,764 | 17,465 | 14,861 |

*Notes:* Panel A reports the average adjustment in self-reported deductions as a result of the audit, with the standard error of the estimate in parentheses. These estimates in panel B are based on the robust RD estimator, controlling for pre-audit deductions. The bandwidth is 0.05 around the risk score cut-off point for the tax audit, normalized at zero. The local linear estimators are specified using triangular kernel functions.

*, **, *** indicate significance levels $p < 0.05$, $p < 0.01$, $p < 0.001$, respectively.

reducing future tax evasion can also reduce private concealment costs and moral costs to the taxpayer (Kreiner 2010; Keen and Slemrod 2017; Meiselman 2018; Slemrod 2019).

In the calculations below, we simplify and assume that $\omega = 0$. More specifically, we ask how far down the risk threshold the tax administration must go to hit the audit threshold that maximizes its net revenues. A positive weight on audited taxpayers ($\omega > 0$) would imply a higher optimal audit threshold. One way to interpret our calculations is that it derives a lower bound for the optimal audit risk score threshold. Another caveat is that we do not consider the general deterrence effect of audits, which should also be accounted for in a complete cost–benefit calculation of tax enforcement. The general deterrence effects are likely to depend on the overall level and visibility of tax enforcement policies. Here, we consider a more narrow question, namely the costs and benefits of expanding one specific type of tax audit, taking into account both the direct audit effect and the specific deterrence effect.

With random audits, the extra tax income generated by a small expansion of audits equals the average treatment effect and is therefore independent of the initial number of audits. This is not the case for threshold audits targeted towards high-risk filers. When audit selection is based on the taxpayer's risk score, an expansion of audits implies a lower risk score at the threshold, and since taxpayers with a lower risk score have fewer misfilings, the expected marginal tax revenue declines in the audit level.

We can use the estimates from Section 5 to calculate the net tax revenue generated at the 2014 audit threshold. The audit reduced deductions by 24,546 NOK in the year of the audit, and by an additional 23,476 NOK in the three post-audit years. Multiplying the total reduction in tax deductions by the relevant tax rate, 22%, we find that the audit increased taxable income by 10,565 NOK. The unit cost of this type of audit is 1625 NOK.[8] Hence, based on our estimates, a small expansion of the 2014 audit would generate a net tax income of 8940 NOK. This is a substantial surplus, and to maximize tax revenue, the tax authorities should lower the risk score threshold and audit more taxpayers. How far down the risk score should they go?

Since RD estimates provide only local information at the threshold, our analysis so far cannot answer this question. To locate the risk score where the marginal tax revenue of an audit equals the unit audit costs, we need to know how the tax income generated by the marginal audit vary over the entire risk score distribution. To obtain this information, we use data from the 2013 sample of about 15,000 randomly audited taxpayers; this is the sample that was used to build the machine learning model in the first place. With these data, we can obtain estimates of the detection effect and the compliance effect over the whole range of risk scores in the target population.

In Figure 7(a), we display, by risk score, the fraction who had their self-reported deductions adjusted upon audit. The population is divided into four risk score brackets. Although only 5% of the taxpayers in the lowest risk bracket had their self-reported deductions adjusted, more than

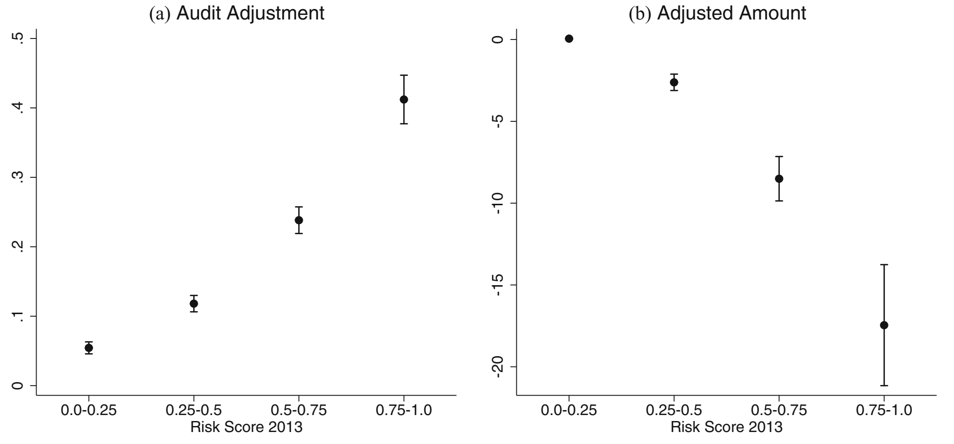


**FIGURE 7** Random audit (2013)—audit outcomes by risk score. *Notes*: (a) For four different risk score intervals, the fraction of audited taxpayers that had their deductions adjusted (i.e. not approved). (b) The average amount (in 1000 NOK) that was adjusted by the NTA at different risk score levels. Error bars show the 95% confidence limits.

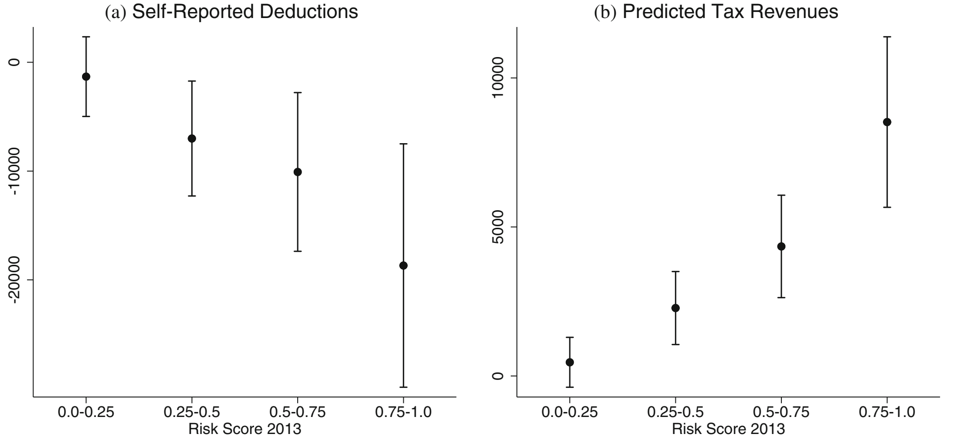


**FIGURE 8** Random audit (2013)—post-audit self-reported deductions and predicted tax revenues by risk score. *Notes*: (a) The estimated accumulated four-year reductions in post-audit self-reported deductions. (b) The predicted tax revenue defined as the total reduction in deductions from the audit adjustment and subsequent post-audit compliance effects, multiplied by the marginal tax rate (0.22). Error bars show the 95% confidence limits.

40% of the taxpayers in the group with the highest risk scores got an adjustment. Figure 7(b) shows that the average amount adjusted by the tax authorities also increases the taxpayer risk score.

Turning to the estimated behavioural effects, Figure 8(a) depicts the aggregated four-year post-audit drop in self-reported deductions for different risk score intervals. The compliance effect is close to zero in the lowest bracket, but turns statistically significant when the risk score exceeds 0.25. If we sum the direct adjustment effect and the four-year post-audit behavioural effects, and multiply by 0.22, then we get an estimated tax revenue effect by risk score interval as shown in Figure 8(b). The tax revenue generated by an audit increases in the risk score, since both effects (i.e. the direct adjustment and the future reductions in self-reported deductions) are increasing in the risk score.

Comparing audit costs and net tax revenues, our estimates suggest that an audit of a taxpayer with a risk score in the 0.5–0.75 region would clearly raise net public revenues. Of course, it is only under special circumstances that the redistributive element of audits (redistributing from the non-compliant to the compliant via public budgets) justifies a zero weight on the income taken from the non-compliant (Slemrod and Yitzhaki 1987). In most cases, it is reasonable to assign a positive welfare weight to non-compliant taxpayers, especially when non-compliance arises from ignorance or confusion because the taxpayer lacks the resources—or does not exercise effort—to fully understand a complicated tax system. The more weight the social planner puts on the loss of income for non-compliant taxpayers, 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 other non-distributive arguments for expanding tax enforcement policies beyond the capacity where marginal revenue equals audit costs. Audits may induce concealment costs, as well as moral costs of cheating. In addition, non-compliance breaks the principle of horizontal equity where individuals with similar incomes and assets should pay the same taxes. 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 could influence the optimal audit capacity.

## 8 | CONCLUSION

Audits are a cornerstone in tax enforcement policies. Audits detect and correct non-compliance on the spot, but may also change the subsequent behaviour of the audited, either by providing new information to misinformed taxpayers, or by changing the perceived risk of future audit for those who deliberately under-report taxable income. The design of optimal tax enforcement depends critically on how taxpayers respond to being audited. In contrast to non-compliance disclosed directly by the audit, any impact on future filing behaviour must be estimated within a framework of counterfactual outcomes. To identify behavioural compliance effects, we use data from operational audits in which all taxpayers with an individual risk score above a critical value were audited. Such threshold audits are ideal for using the RD design to identify effects on future tax filings that are relevant for policy.

We identify substantial long-term compliance effects of being audited. Compared to their non-audited neighbours in the risk distribution, audited taxpayers reduced self-reported deductions by 19% in the first post-audit year. Even if the yearly effect declined over the next years, the behavioural effect over a three-year post-audit period is similar to the direct adjustment of the audit. We find no indications of spillover effects on spousal tax filing, and the (imprecise) estimates on net income suggest that pre-tax income is unaffected by the audit.

The evidence from several audits, including the random audit used to estimate a prediction model, shows that the reduction in deductions is increasing in the risk score. By comparing marginal tax revenue and audit costs, we show that an expansion of the audit capacity would raise public net revenue. If the (marginal) welfare weight assigned to a non-compliant taxpayer is low, then the policy implication is to lower the threshold and audit a larger fraction of the taxpayers, even if the average net revenue from the audited taxpayers falls.

Since profiling and machine learning strategies are used increasingly as enforcement tools in many countries, our application of the RD design and the empirical results that we present should be of general interest and extend beyond Norwegian tax enforcement policies.

### ACKNOWLEDGMENTS

This paper is part of the Norwegian Fiscal Studies research programme 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, Knut Røed, Fred Schroyen and participants in the OFS workshop 'Empirical analyses of tax compliance' and Skatteforum 2018 for their useful comments and suggestions.

Løyland and Øvrum are analysts in the Norwegian Tax Administration. As independent researchers, Raaum and Torsvik have no conflict of interest.

## ENDNOTES

[1] In a recent paper, Sánchez (2022) uses an RD design to estimate the effects of receiving a non-compliance notification.

[2] From 1 November 2020, the responsibility for direct tax collection was transferred to the NTA.

[3] We do not know who among those who were checked but found to be compliant actually became aware of the audit. But even if we had this information, it could not be used to identify compliance effects for this group, as it is endogenous and we have no appropriate control group with which to make comparisons.

[4] See (Hebous *et al.* 2023) for a closer discussion on audits and penalties in a Norwegian context.

[5] Only 883 taxpayers were selected for audit in the third wave. Our robustness check shows that the results are very similar if we drop wave three observations. The small number of audits and the fact that the taxpayers in the last wave are mainly self-employed explains why the risk score threshold in the third wave is much higher compared to the first two waves.

[6] Using a default bin size calculation, the bin size is 0.000828079, with discontinuity estimate (log difference in height) 0.19122, standard error 0.05181, and significance at the 1% level.

[7] As pointed out by a referee, the effects on total deductions, i.e. on the sum of pre-filled (third party reported) and self-reported deductions, would be lower than the effects on self-reported deductions if corrected taxpayers had legitimate deductions but were audited and corrected because they had not done the necessary paperwork to get the deductions registered and reported by a third party. To check for this, we run the same regressions with total deductions as the outcome variable. This gives a compliance effect that is slightly higher than for self-reported deductions.

[8] This is an estimate that the NTA gave us based on the average variable cost associated with this type of desk audit. It includes wages and social costs (e.g. payroll taxes, mandatory employer insurance and holiday pay), IT equipment costs, and office costs. In the calculations below, we assume that the unit cost is independent of the risk score. This is a simplification; it is more plausible that the time required to control a taxpayer increases in the taxpayer's risk score. If that is the case, then our estimate of the optimal risk score threshold, which is based on a unit cost independent of the risk score, provides an upper bound of the true revenue maximizing threshold.

## REFERENCES

Advani, A., Elming, W. and Shaw, J. (2023). The dynamic effects of tax audits. *Review of Economics and Statistics*, **105**(3), 545–61.

Andreoni, J., Erard, B. and Feinstein, J. (1998). Tax compliance. *Journal of Economic Literature*, **36**(2), 818–60.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *Review of Economics and Statistics*, **60**(1), 47–57.

Beer, S., Kasper, M., Kirchler, E. and Erard, B. (2020). Do audits deter or provoke future tax noncompliance? Evidence on self-employed taxpayers. *CESifo Economic Studies*, **66**(3), 248–64.

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, **82**(6), 2295–326.

DeBacker, J., Heim, B. T., Tran, A. and Yuskavage, A. (2018). Once bitten, twice shy? The lasting impact of enforcement on tax compliance. *Journal of Law and Economics*, **61**(1), 1–35.

Gemmell, N. and Ratto, M. (2012). Behavioral responses to taxpayer audits: evidence from random taxpayer inquiries. *National Tax Journal*, **65**(1), 33–57.

Hebous, S., Jia, Z., Løyland, K., Thoresen, T. O. and Øvrum, A. (2023). Do audits improve future tax compliance in the absence of penalties? Evidence from random audits in Norway. *Journal of Economic Behaviour and Organization*, **207**, 305–26.

Heckman, J. J. and Smith, J. A. (1999). The pre-programme earnings dip and the determinants of participation in a social programme. Implications for simple programme evaluation strategies. *Economic Journal*, **109**(457), 313–48.

Hodge, R. H., Plumley, A. H., Richison, K., Yismaw, G., Misek, N., Olson, M. and Wijesinghe, H. S. (2016). Estimating marginal revenue/cost curves for correspondence audits. IRS Research Bulletin, Publication 1500.

Keen, M. and Slemrod, J. (2017). Optimal tax administration. *Journal of Public Economics*, **152**, 133–42.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S. and Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica*, **79**(3), 651–92.

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

Mazzolini, G., Pagani, L. and Santoro, A. (2022). The deterrence effect of real-world operational tax audits on self-employed taxpayers: evidence from Italy. *International Tax and Public Finance*, **29**(4), 1014–46.

Meiselman, B. S. (2018). Ghostbusting in Detroit: evidence on nonfilers from a controlled field experiment. *Journal of Public Economics*, **158**, 180–93.
OECD (2006). *Strengthening Tax Audit Capabilities: General Principles and Approaches*. Paris: OECD Publishing.
——— (2019). *Tax Administration 2019: Comparative Information on OECD and Other Advanced and Emerging Economies*. Paris: OECD Publishing.
——— (2021). *Behavioural Insights for Better Tax Administration: A Brief Guide*. Paris: OECD Publishing.
Sánchez, G. E. (2022). Non-compliance notifications and taxpayer strategic behavior: evidence from Ecuador. *International Tax and Public Finance*, **29**(3), 627–66.
Sarin, N. and Summers, L. H. (2020). Understanding the revenue potential of tax compliance investment. NBER Working Paper no. 27571.
Slemrod, J. (2019). Tax compliance and enforcement. *Journal of Economic Literature*, **57**(4), 904–54.
——— and Yitzhaki, S. (1987). The optimal size of a tax collection agency. *Scandinavian Journal of Economics*, **89**(2), 183–92.

**How to cite this article:** Løyland, K., Raaum, O., Torsvik, G. and Øvrum, A. (2024). Evaluating compliance gains of expanding tax enforcement. *Economica*, **91**(361), 142–162. https://doi.org/10.1111/ecca.12499

## APPENDIX

**TABLE A1** Compliance Effects, Self-reported Deductions—Robustness

| | Pre-audit | | Post-audit | | All post |
|---|---|---|---|---|---|
| | 2014 | 2015 | 2016 | 2017 | 2015–17 |
| *A. No pre-audit control* | 741 | −13,989*** | −7120* | −2373 | −22,775** |
| | (3544) | (3715) | (3374) | (3212) | (8170) |
| Observations | 6997 | 6952 | 3860 | 6779 | 6891 |
| *B. Sharp RD* | 741 | −11,528*** | −5896* | −2262 | −18,803** |
| | (3544) | (2804) | (2659) | (2550) | (6004) |
| Observations | 6997 | 6952 | 3860 | 6779 | 6891 |
| *C. Sample waves 1 and 2* | 5168 | −15,966*** | −8600** | −4142 | −27,706*** |
| | (4128) | (3540) | (3206) | (3387) | (8219) |
| Observations | 5473 | 5331 | 5337 | 5409 | 5461 |

*Notes:* These estimates are based on the robust RD estimator. The bandwidth is 0.05 around the risk score cut-off point for tax audit, normalized at zero. The local linear estimators are specified using triangular kernel functions. Standard errors in parentheses.
*, **, *** indicate significance levels $p < 0.05$, $p < 0.01$, $p < 0.001$, respectively.

**TABLE A2** Scaled Effects on Post-audit Self-reported Deductions

| | 2015 | 2016 | 2017 |
|---|---|---|---|
| Audit | −0.263* | −0.265*** | −0.125 |
| | (1.04) | (0.074) | (0.093) |

*Notes:* The dependent variable is self-reported deductions in each post-audit year, scaled by the pre-audit self-reported deductions in 2014. Estimates are based on the robust RD estimator. The bandwidth is 0.05 around the risk score cut-off point for tax audit, normalized at zero. The local linear estimators are specified using triangular kernel functions. The sample sizes are affected by the trimming of the one-percentiles of the yearly distributions. Standard errors in parentheses.
*, **, *** indicate significance levels $p < 0.05$, $p < 0.01$, $p < 0.001$, respectively.

**TABLE A3** Post-audit Pre-filled and Total Deductions—Sum All Post-audit Years

| | Pre-filled deductions (NOK) | Total deductions (NOK) |
|---|---|---|
| Audit | 141 | −30,078*** |
| | (8179) | (12,001) |
| Observations | 6950 | 6950 |

*Notes:* Estimates are based on the robust RD estimator. We control for pre-audit values of deductions. The bandwidth is 0.05 around the risk score threshold (normalized at zero). The local linear estimators are specified using triangular kernel functions. The sample sizes are affected by the trimming of the one-percentile tails of the yearly distributions. Standard errors in parentheses.
*, **, *** indicate significance levels $p < 0.05$, $p < 0.01$, $p < 0.001$, respectively.

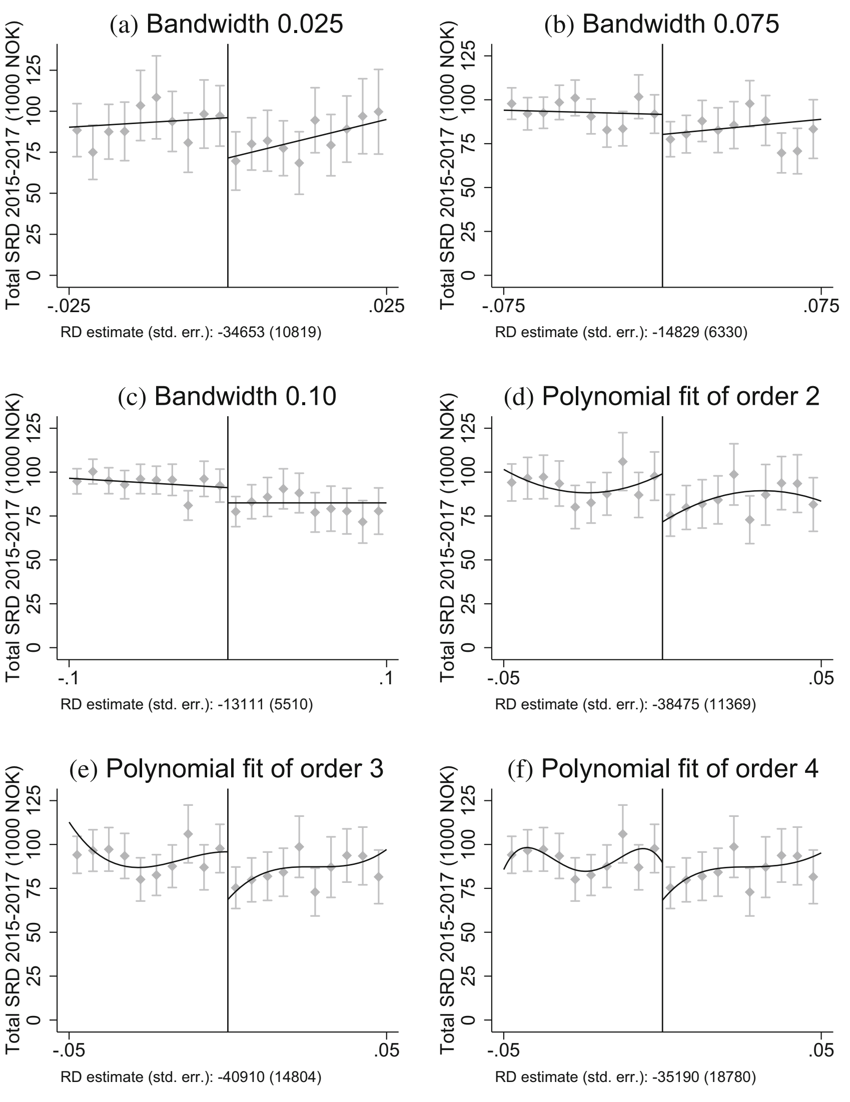


**FIGURE A1** Compliance effects for different bandwidths and polynomial orders. *Notes*: Regression discontinuity plots and estimates for accumulated self-reported deductions in 1000 NOK for the three first post-audit years against the forcing variable (risk score) for different bandwidths and polynomial orders, as robustness checks for the main specification in Figure 4. The local estimators are specified using triangular kernel functions, and ten bins are shown on both sides of the threshold. Error bars show the 95% confidence limits.

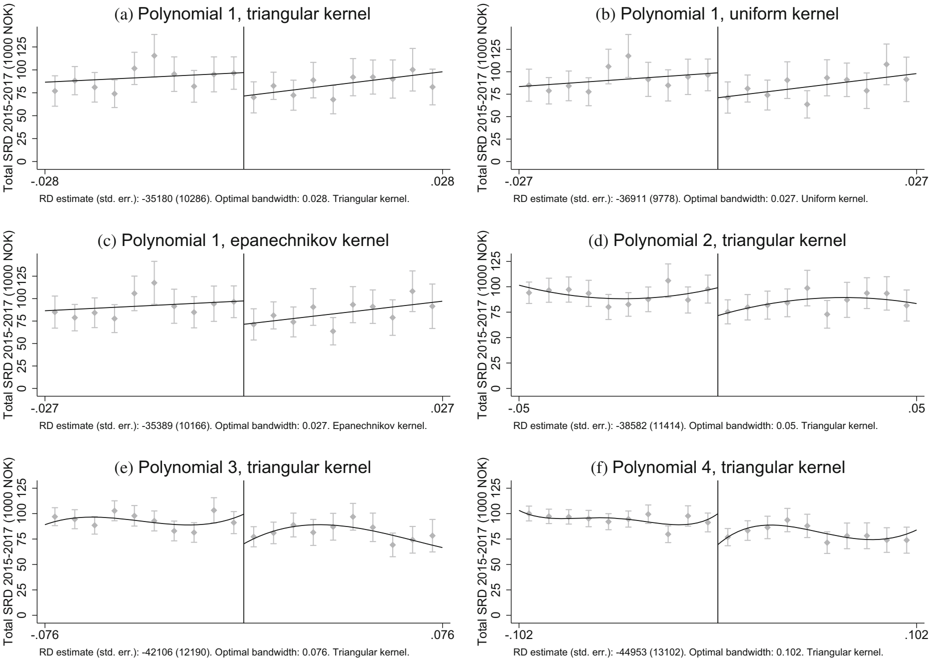


**FIGURE A2** Compliance effects for alternative kernels for optimal bandwidth. *Notes*: Regression discontinuity plots and estimates for accumulated self-reported deductions in 1000 NOK for the three first post-audit years against the forcing variable (risk score) for optimal bandwidths and different polynomial orders and kernel functions, as robustness checks for the main specification in Figure 4. Error bars show the 95 % confidence limits.

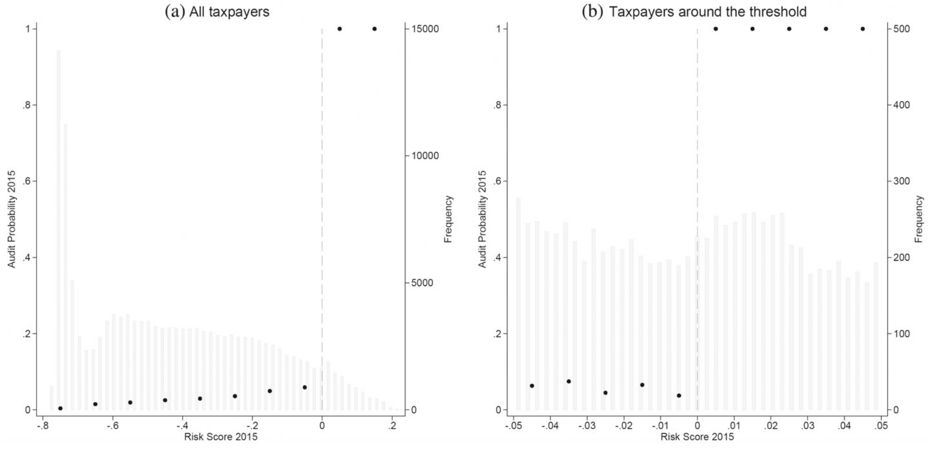


**FIGURE A3** Audit probability (2015) by risk score. *Notes*: In 2015, the NTA selected taxpayers for audits in three different waves. But the budget allocated to this audit was substantially larger, which means that the risk score threshold was set at a lower level than in 2014. Our analysis is based on data from all three waves, and we have normalized the threshold to 0. We can see that below the threshold, the probability of a flag audit of self-reported deductions increases in the risk score, and there is a distinct jump at the threshold.