Using Intelligence from International Tax Cooperation to Improve Voluntary Tax Compliance: Evidence from a Swedish Field Experiment*

Nikolay Angelov† Per Johansson‡

December 18, 2020

Abstract

This study documents the results from a field experiment consisting of a low-cost intervention: a digitally disseminated, neutrally worded message sent to a group of Swedish taxpayers in March 2020, about six weeks before the final day to file income taxes for the income year of 2019. The message was sent to taxpayers who had foreign dividends during the income year of 2018 according to records from automatically exchanged financial information following OECD:s Common Recording Standard (CRS). The purpose of the study is to estimate the effect of the message on subsequent compliance, measured by two outcome variables: declared capital income and total tax paid for 2019. We use an efficient experimental design featuring stratification on two binary variables (previous compliance and gender) and then rerandomization within each of the four strata. We find no evidence of an average effect on compliance. In an exploratory heterogeneity analysis, we examine the data along two hypotheses: that the effect of the intervention should be larger among previously non-compliant compared to compliant taxpayers and larger among women compared to men. We find support for both conjectures for declared capital income. Finally, we provide numerical evidence on the magnitude of the efficiency gains from the experiment design based on real data.

*We are thankful for valuable comments and suggestions from Thomas Aronsson, Spencer Bastani, Per Engström, Daniel Hallberg, and Katarina Nordblom.

†The Swedish Tax Agency, Uppsala Center for Labour Studies (UCLS) and Uppsala Center for Fiscal Studies (UCFS), nikolay.angelov@skatteverket.se

‡Department of Statistics at Uppsala University, IFAU, and IZA, per.johansson@statistik.uu.se

1 Introduction

Globally, a considerable part of the wealth is undeclared. As noted in Zucman (2013), 6 percent of the global financial wealth of households is unrecorded. Also, offshore tax evasion is highly concentrated among the wealthiest thereby reinforcing inequality. According to the estimates provided in Alstadsæter et al. (2019), the top 0.01 percent evades 25 percent of its true tax liability through tax havens. Even a slight move down the wealth distribution implies a large drop in the evaded tax share. For instance, the estimated evaded tax share is 10 percent among those between the 99.95th and 99.99th wealth percentile and close to zero for those between the 90th and 95th percentile.

Because of the major consequences of cross-border tax evasion for lost government revenue and inequality, there has been an ongoing and, during the last decades, increasing international cooperation effort to combat such practices. The Organisation for Economic Co-operation and Development (OECD) has been paramount in coordinating these efforts. An initiative of particular importance is The Multilateral Convention on Mutual Administrative Assistance in Tax Matters which was initially co-signed in 1988 by the OECD and the Council of Europe with the explicit goal of combating international tax evasion and other forms of non-compliance among OECD member states. In 2010, the convention became open to non-member countries (OECD and Council of Europe, 2010) and currently, 141 countries participate.[1] An important feature of the convention is the automatic exchange of financial account information, guided by the so called Common Reporting Standard (CRS) which was approved by the OECD council in 2014.[2] Within the CRS, tax authorities, including The Swedish Tax Agency, obtain information from financial institutions in their own jurisdiction and automatically exchange that information with other jurisdictions on an annual basis. Although the CRS has some loopholes and it certainly cannot be expected to include information on all financial accounts, the exchanged information covers many countries and a vast amount of assets. In 2019, nearly 100 countries carried out automatic exchange of information in 2019, enabling their tax authorities to obtain data on 84 million financial accounts held offshore by their residents. This covered total assets of EUR 10 trillion which is twice as much as the number during 2018, the first year in which such automatic information exchange took place.[3]

The purpose of the present study is to estimate the effect of a low-cost intervention (a digitally disseminated message) on the subsequent compliance of taxpayers who, accord-

[1]Source: http://www.oecd.org/ctp/exchange-of-tax-information/convention-on-mutual-administrative-assistance-in-tax-matters.htm. The information was retrieved on September 30, 2020.

[2]Formally, Article 6 in The Convention on Mutual Administrative Assistance in Tax Matters, requires from the signing authorities to agree upon the scope of automatic information exchange. The Multilateral Competent Authority Agreement on Automatic Exchange of Financial Account Information (the "CRS MCAA") was developed to this end.

[3]Source: http://www.oecd.org/tax/international-community-continues-making-progress-against-offshore-tax-evasion.htm. The information was retrieved on September 20, 2020.

ing to CRS-data, had foreign dividends during the income year of 2018. We performed a field experiment consisting of a message sent to a group of taxpayers in March 2020, about six weeks before the final day to file income taxes for the income year of 2019. Since the tax agency receives CRS-information regarding a certain income year only after the taxpayers have filed their income tax returns and the final tax has been determined, the agency (with minor exceptions) cannot use the information for, e.g., risk-based audits of incoming tax declarations.[4] However, it can be expected that there is some correlation over time in foreign income flows at the individual level. This implies that knowing the foreign income of a taxpayer provides at least some information about next year's foreign income. We use a combination of foreign income records for the income year of 2018 together with records from the tax declaration for the same year to build a measure of non-compliance, which is used in the experimental design as well as for investigating potential effect heterogeneity with respect to previous compliance.

All treated taxpayers received the same message which conveyed two main points. First, the treated taxpayers were informed that The Swedish Tax Agency had received CRS-information about their foreign income during 2018. This information was mainly provided for transparency reasons with the goal of removing any potential uncertainty about why a particular person got the message. Although the goal was simply to inform the taxpayers about why the message was sent to them, it cannot be ruled out that receiving the information in practice lead to an increased perceived detection probability, in particular among previously non-compliant individuals. Second, the taxpayers were informed about a recently launched service by The Swedish Tax Agency consisting of an online app which can be used to calculate and subsequently file the proper amount of foreign income in the declaration. The development of the app was motivated by the so called *service paradigm* for tax administration (see for instance Alm 2019). Making the app available is a part of an ongoing effort by The Swedish Tax Agency to provide better service and assist taxpayers in their filing returns and paying taxes. Of course, although the explicit goal was better service, we have no means of knowing how the taxpayers actually perceived being presented with the new app. For instance, the app does not require a login, but taxpayers with limited IT-knowledge or who are critical of surveillance, could potentially have met this new service with some suspicion. It should be noted however that the tax agency did not receive any such signals from the treated taxpayers. There were for instance no calls from worried or upset taxpayers.

Thus, the present study is related to two strands in the tax compliance literature: the role of detection probability and that of improved service to the taxpayers. Improved service could potentially affect compliance via two channels: improved information on how

[4]For the income year of 2019, the final day to file income taxes for private taxpayers was May 4, 2020 and the decision for final tax determination was sent out to the taxpayers on June 12, 2020. The CRS-information about foreign income usually becomes available about three months later, in September/October.

to declare or an improvement in the taxpayers' general attitudes toward the tax agency. There is a large body of empirical research measuring the effect of the probability of detection but we are not aware of any research directly measuring the effect of improved service on compliance. Below, we provide a condensed summary of the available research on the role of detection probability for tax compliance. According to standard neo-classical economics and as first formalized in Alingham and Sandmo (1972), an increased perceived detection probability leads to a higher level of compliance. This theory has found considerable empirical support. Initially, the predictions from the theory were confirmed in a series of laboratory experiments.[5] In addition, starting in the early 1990s, data from the TCMP[6] were used in several observational studies to measure the impact of the audit rate on compliance.[7] In more recent years, a growing number of field experiments have estimated a positive effect of a higher detection probability on tax compliance by manipulating the perceived audit rate in various ways. In their influential study, Kleven et al. (2011) set up a carefully designed field experiment among 40,000 Danish taxpayers, half of whom were randomly assigned to tax audits. The following year, threat-of-audit letters were sent to audited as well as non-audited taxpayers. Kleven et al. find that prior audits and threat-of-audit letters have positive effects on self-reported income, but no effect on third-party reported income. Several other studies in the literature have also documented a positive effect of increased detection risk on tax compliance.[8]

Although the direct cost of the treatment is low, the number of treated was limited to 500 individuals. The reason for this decision was uncertainty on behalf of the team responsible for planning and sending the message about how the taxpayers would react to the message. In particular, there were concerns about potential service desk overload. As the treatment only concerns sending the digital message the design described in detail below is a so called encouragement design. We do not know whether the receiver reads the digital message we can therefore only hope to estimate the intention to treat (ITT) effect. As a consequence of this in combination with the non-intrusive nature of the message, the effect size cannot be expected to be large. For this reason and because of the relative small number of treated, it is useful to consider a more efficient experimental design than a traditional randomized control design (i.e., unconditional randomizing treatment and

[5]In an early experiment, Spicer and Thomas (1982) varied the audit probability and found a positive relationship between the audit rate and compliance. A series of lab experiments have since then replicated these initial results: Becker et al. (1987); Webley (1987); Alm et al. (1992a, 1992b, 1992c, 1995, 1999, 2017); Fortin et al. (2007); Cummings et al. (2009); Kastlunger et al. (2009); Calvet and Alm (2014); Soliman and Cullis (2014); Casagrande et al. (2015).

[6]The TCMP is an abbreviation for the US Internal Revenue Service's Taxpayer Compliance Measurement Program, which involved comprehensive random audits of taxpayers from 1968 to 1988.

[7]Dubin et al. (1990) used variation in the audit rate across US states to measure the relationship between detection risk and compliance, and found a strong positive relationship. Similar results using essentially the same data can be found in Tauchen et al. (1993), Plumley (1996) and Dubin (2007).

[8]See Slemrod et al. (2001), Pomeranz (2015), Shimeles et al. (2017), Bott et al. (2020), Boning et al. (2018), Meiselmann (2018).

controls).

To this end, our chosen experimental design features stratification on two categorical variables (previous compliance and gender) and then rerandomization within each of the four strata. The main idea is to discard allocations of treated and control units in which there is imbalance in observed covariates and then to randomize within the set of allocations where there is balance on these covariates. Although previously mentioned in the literature, the idea of rerandomization was to our knowledge first formalized by Morgan and Rubin (2012). This paper uses the strategy in Morgan and Rubin (2012) who suggested the Mahalanobis distance as the criterion for defining the set of allocations with sufficient balance on the covariates. To our knowledge, ours is the first application using this design in the tax compliance literature. Compared to simple random sampling, there are efficiency gains with both stratification and rerandomization. Due to the uniqueness of the data in this study, we are able to provide numerical evidence on the magnitude of these gains based on real data.

The study closest to ours is Bott et al. (2020) who report the results from a randomized field experiment in Norway among taxpayers deemed by the tax authority to have likely misreported their foreign income. The taxpayers received a letter with information about how to report foreign income that included two types of moral appeal: a fairness argument and a societal benefits argument. In addition, the perceived detection probability was manipulated. Compared to a base letter, including either appeal almost doubled the average foreign income reported compared to a base letter. This effect was similar in size to the one found when letting the taxpayer know that the tax authority had information about income or assets abroad in previous years. Bott et al. (2020) found that the moral appeals worked on the intensive margin, by increasing the amount reported, while the detection probability worked on the extensive margin, by increasing the share of taxpayers who report any foreign income. Only the detection treatment had any effect in the subsequent year.

There are three main differences between the present study and the one by Bott et al. (2020). First, our intervention does not include any moral appeal. Instead, it can be viewed as a combination of unintended increase in detection probability and increased service to the taxpayers. Second, we explore effect differences between taxpayers who were likely compliant and non-compliant, respectively, while Bott et al. (2020) study the effect only among the likely non-compliant. Third, we use a novel experiment design involving stratification and rerandomization. By using non-treated individuals from the same population as was sampled to be treated we are also able to evaluate the efficiency gains in real data. According to our understanding, this is the first study ever attempting to pin down the efficiency gains in this manner. Results from this numerical exercise should be useful for future randomized field experiments, especially, but not exclusively, within the field of tax compliance.

The paper proceeds with a description of the intervention and study sample in the next section. The experimental design is described in section 3 and in the following section 4, we discuss the empirical specification. Section 5 contains the results and in section 6, we provide a simulation exploring the gains from the chosen experimental design. Finally, we summarize our findings in section 7.

2 Intervention and study sample

The intervention consisted of sending the following text to the treated individuals and leaving the controls untreated:

> *Declaring foreign dividends/interest*
>
> Hi!
>
> We get many questions on how to declare foreign income and have therefore developed a new online app in order to make this easier.
>
> The Swedish Tax Agency has obtained information from a foreign tax authority that you have received dividends or interest from abroad during 2018.
>
> If you have received dividends or interest from abroad also during 2019, you can use the online app when you file your taxes. The app will help you with the correct amount to file and how much foreign tax offset you have the right to claim.
>
> You can find the app here:
>
> `https://app.skatteverket.se/klient-sifu-segmentering/`
>
> Sincerely,
>
> The Swedish Tax Agency

This message was sent in digital form to the treated taxpayers via a digital mailbox. In Sweden, about half of the population above 16 years of age have a digital mailbox.[9] This is a free service making it possible to receive mail in digital form from Swedish authorities and some large private firms. With regards to taxes, having a digital box implies that all communication from The Swedish Tax Agency that otherwise would have been sent as paper mail is sent digitally in a secure app. This includes pre-filled tax returns as well as various messages. Taxpayers can file their income tax declaration securely in the app.

The online app mentioned above does not require login and is best described as a calculator. The user fills in the type of foreign income (dividends or interest), amount,

[9]Source: https://svenskarnaochinternet.se/rapporter/svenskarna-och-internet-2019/digitala-samhallstjanster/halften-av-svenskarna-har-en-digital-brevlada/. The information was retrieved on September 30, 2020.

currency, date of receiving the amount, country where the income was received, and if applicable, the amount of foreign tax paid. Upon clicking *Calculate*, the app converts the amount to SEK, calculates the foreign tax offset, and indicates the specific tax declaration boxes where the amounts should be filed.

Data on the outcome variables (to be presented below) comes from the taxpayers' filed income tax returns. The latest day for sending in the income tax declaration for the income year 2019 was May 4, 2020. For various reasons, some taxpayers file their taxes later than the deadline. The most common reason for filing at a later date is that the individual has applied and been approved for respite with income tax returns. As of July 27, 2020, we had access to data on 99.3 percent of the study sample, i.e., the group of individuals who received the message and the control group.

The original population consisted of 10,344 individual taxpayers each of whom had a financial account with accrued foreign dividends summing to over 3,000 SEK (about 292 EUR) during tax year 2018, according to information from the CRS. A small share of the financial accounts had several account holders, which implies that the involved taxpayers' dividends were overestimated in cases where each individual taxpayer's dividends were below 3,000 SEK. We do not have access to the actual numbers due to the tedious CRS-data collection routines at The Swedish Tax Agency, but according to caseworkers, accounts with multiple holders constitute a very small share.

As mentioned previously, for internal Tax Agency reasons, although the direct cost of the treatment is low, the number of treated was limited to 500 individuals. In addition, we expect the effect to be modest as the message is non-intrusive and we do not even know if a receiver reads it. This implies that it is important to have an efficient design, which we describe below.

The original sample of 10,344 taxpayers was reduced according to the following steps:

1. Some of the individuals had not filed a tax declaration for tax year 2018 and were therefore removed from the sample (259 obs removed)

2. Individuals who did not have a digital mailbox on February 27, 2020, were removed (5,730 obs removed)

3. Observations with respect to several pre-experiment variables for the tax year of 2018 were defined as outliers and removed if the value of any single variable was larger than the third quartile plus 1.5 times the interquantile range (IQR) or smaller than the first quartile minus 1.5×IQR. The quartiles for all variables were calculated on the same data, i.e., after the previous step. The following variables were used for outlier detection (with *variable label* and number of removed observations in parentheses):

    - Foreign dividends (*fdiv*, 638)

- Earnings including labor income, sick pay, pension, etc. (*earn*, 355)
- Capital income (*capinc*, 725)
- Foreign dividends as a share of capital income ($fdivrat = \frac{fdiv}{capinc+1}$, 513)
- Total tax paid (*tax*, 439)

The unique number of individuals removed due to outliers amounts to 1,546.

4. Individuals aged below 30 and above 75 were removed (112)

After these steps, there were 2,697 individuals left in the sample.

There is no separate field for declaring foreign dividends. Our main outcome variable is therefore capital income (*capinc*), where according to the tax code, the taxpayer should include the amount of foreign dividends. In addition, we estimate the effect on total tax paid (*tax*). An increase in declared foreign dividends in *capinc* should lead to an increase in total tax paid unless no offsetting adjustments are made in the income tax declaration. Since the treatment is expected to have an effect on *tax* only through changes in foreign income which is a minor part of *tax*, we do not expect a large effect on *tax*. It is nevertheless important to include the final tax as an outcome variable since it is a measure of compliance directly related to tax revenues. Also, although perhaps far fetched, receiving the message could potentially have a positive effect on overall compliance through an unintentional nudge. In other words, merely receiving a message from the Swedish Tax Agency could possibly nudge taxpayers into higher general compliance level which could be manifest through an increase in total tax paid not necessarily stemming from increased declared capital income.

The message was sent to each treated taxpayer's digital box on 25 March, 2020. In addition to measuring the effect of receiving the message on the taxpayers' tax compliance, it is of interest to assess whether receiving the message had an effect on using the online app mentioned in the message. However, since the app does not require the users to log in, this is not feasible.

It is nevertheless informative to see if there was any obvious increase in the number of users on the online app over time. A descriptive plot is shown in Figure 1 below. In addition to the day of the dispatch (March 25) and the final date to send in the income tax declaration (May 4), two additional dates have been highlighted: an online press conference for a broad public arranged by The Swedish Tax Agency on March 13, and a live stream on Facebook on March 17. The online app was promoted on both of these occasions. Clicks on the link to the app from within the digital mailbox should be visible in the dotted line, while copying the link and pasting it in the browser should be visible as a direct visit (the solid line).

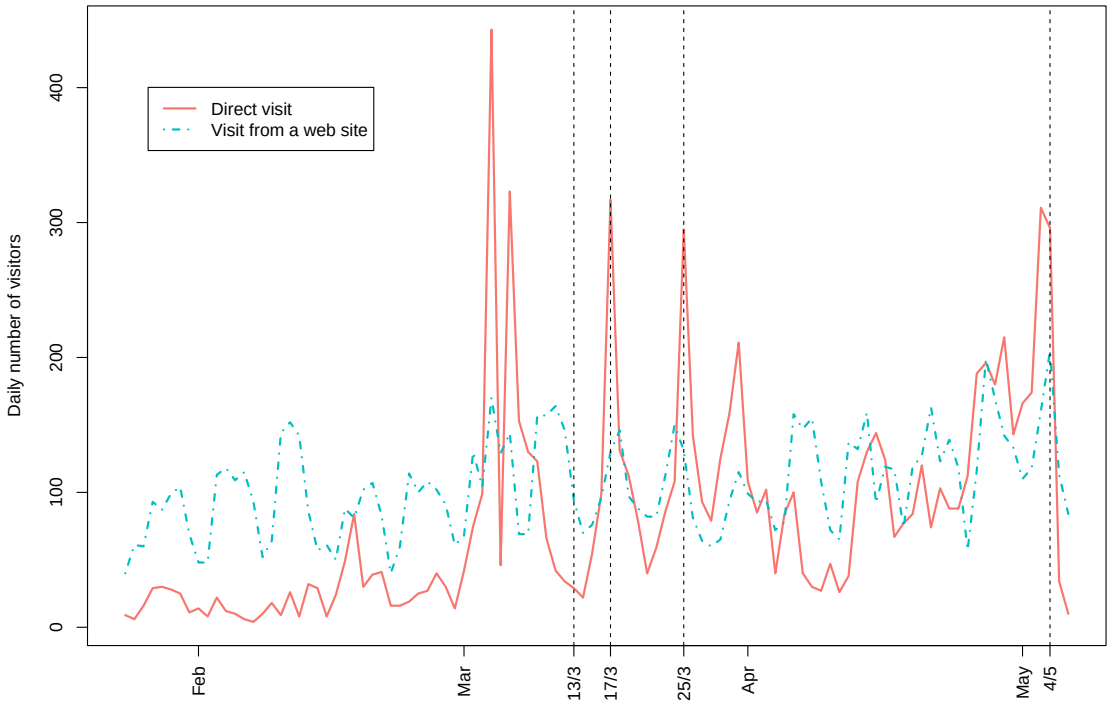

Figure 1: Daily number of users of the online app

On the whole, our take on Figure 1 is that it is hard to see any clear correlation between sending the message and the number of app users. There is certainly an increase of direct visitors during the day of the dispatch, but linking it to the message would be a stretch. Looking at the two preceding dates of promotion, there was an increase on the same day as the Facebook live-stream but no increase (indeed, a fall) in the number of visits during the press conference on March 13. According to the team involved in implementing the online app, the high number of visits in the beginning of March was due to employees at The Swedish Tax Agency fine tuning the app.

3 Experimental design

Balanced designs, that is where the number of treated and controls is equal, are preferable to unbalanced designs in both Fisher (Chung and Romano, 2013) and Neyman-Pearson (Freedman, 2008) inference. For this reason, we decided to equalize the number of treated and controls, meaning that the sampling frame is set to 1,000 individuals.

Even though the estimators from well-conducted experiments are unbiased in expectation, the estimates from any single experiment may still be far from being unbiased (in the colloquial sense) due to an unlucky, albeit random, allocation. For this reason the experimental design is stratified on gender and historical tax compliance; two variables

that we believe can be important determinants for the two outcomes under investigation. A second reason for stratification is that we are interested in examining group differences in behavior as a consequence of the intervention. Hypotheses on the signs of effect differences along previous compliance and gender are discussed in section 5.2.2. As we also have continuous covariates (historical data on the outcomes, earnings, age, etc.) we also would like to balance them within each stratum. To this end, we used the rerandomization strategy suggested by Morgan and Rubin (2012). For details on the procedure, see Appendix A.

We do not have access to a demonstrably reliable measure of previous compliance. As a proxy, we use information about whether the amount of foreign dividends obtained from the CRS-data for the income year of 2018 was less than or equal to total capital income during the same year, i.e., $\mathbf{1}[compliant = 1] \equiv \mathbf{1}[fdiv \leq capinc]$, where $\mathbf{1}[\cdot]$ takes the value one if the expression within brackets is true and zero otherwise. The logic behind this is that for compliance with the tax code, the amount of $fdiv$ should be included along with other capital income sources in the declared *capinc*. Therefore, although $fdiv \leq capinc$ is not necessarily a sign of compliance, $fdiv > capinc$ is a clear measure of non-compliance.

The idea behind rerandomization is the same as with stratification or blocking, that is, to remove from consideration allocations with imbalance in observed covariates between treated and control units and then randomize within the set of allocations with balance on these covariates. Call the set of all allocations within a stratum $\mathcal{A}$ and the set of acceptable allocations $\mathcal{A}_a$. Thus, $Card(\mathcal{A}_a) < Card(\mathcal{A})$, where $Card(\mathcal{A})$ and $Card(\mathcal{A}_a)$ is the cardinality of $\mathcal{A}$ and $\mathcal{A}_a$, respectively. Morgan and Rubin (2012) suggested calculating the Mahalanobis distance between the means of K covariates of the potentially treated and controls and then to accept a specific random allocation only if this measure is less than a, where a is small. The Mahalanobis distance is chi-square distributed with K degrees of freedom, i.e., χ^2_K. The criterion a, determining the set $\mathcal{A}_a$, can, thus, be decided implicitly by accepting a random allocation if the calculated Mahalanobis distance between the two means is smaller than a pre-specified probability derived from the χ^2_K. If we let $\Pr(\chi^2_K \leq a) = p_a$, the specific random allocation is one allocation of the $p_a/100$ % allocations with the smallest difference in means between treated and controls.

As shown in the Appendix, the percent reduction in variance in contrast to complete randomization can be shown to be equal to

$$100 \times R^2(1 - \nu_a), \tag{1}$$

where R^2 is the coefficient of determination of a regression of the outcome on $\mathbf{X}$ and

$$\nu_a = \frac{\Pr(\chi^2_{K+2} \leq a)}{\Pr(\chi^2_K \leq a)}; 0 < \nu_a < 1. \tag{2}$$

ν_a is non-increasing with K and decreasing in p_a. The implication is that the choice of covariates is important. One should not add unnecessary covariates (i.e., covariates that are not partially correlated with the outcome) as this reduces the efficiency of the estimator. Furthermore, a strict criterion for a given K will increase the efficiency in comparison to a less strict criterion.

The following six covariates, measured for the pre-experiment income year 2018, were used to calculate the Mahalanobis distance:

- *age*: the taxpayer's age
- $fdiv$: foreign dividends
- *capinc*: capital income
- tax: total tax paid
- *earn*: earnings including labor income, sick pay, pension, etc.
- $finc \equiv \mathbf{1}[has\ foreign\ income]$: categorical variable based on a check box in the tax declaration which equals one if the box is checked and zero otherwise.

Using the 2,697 individuals left in the sample after the procedure used in the previous section, we proceeded as follows:

1. divide the sample in a compliant (1,759) and a non-compliant group (938) where the number of observations is given within parentheses.
2. draw two simple random samples, each of size 500, from the two groups. These 1,000 individuals constitute the sampling frame of the trial.
3. Create four strata; compliant women, compliant men, non-compliant women, and non-compliant men
4. Within each stratum, randomly select an allocation with a Mahalanobis distance between treated and controls means of the six covariates to be less than 0.17. As $P\left(\chi^2(6) < 0.17\right) = 0.0001$ this means that the specific random allocation is one allocation of the 0.01 % allocations with the smallest differences in means between the treated and the controls.

Table 1 shows group means by stratum after performing the rerandomization as well as the resulting number of observations in each stratum. From this table we can see that we have 167×2 non-compliant men, 83×2 non-compliant women, 200×2 compliant men, and 50×2 compliant women. The mean difference of the covariates between the treated are, as expected, very small within each of the four strata.

Table 1: Group averages after rerandomization

Stratum		*age*		*fdiv*		*capinc*		*tax*		*earn*		*finc*		*#obs*	
$\mathbf{1}[compliant=1]$	$\mathbf{1}[woman=1]$	C	T	C	T	C	T	C	T	C	T	C	T	C	T
0	0	50.93	50.57	7.46	7.39	1.81	1.82	248.59	253.11	615.49	625.63	0.08	0.08	167	167
0	1	51.86	51.58	6.51	6.48	1.58	1.52	207.97	206.92	548.71	544.53	0.07	0.06	83	83
1	0	52.75	52.65	6.35	6.35	11.21	11.23	339.85	341.63	829.46	829.56	0.06	0.07	200	200
1	1	51.44	51.90	7.63	7.76	13.26	13.85	306.35	310.08	769.90	774.46	0.18	0.18	50	50

Note: C denotes controls and T denotes treated individuals. The variables are measured during the pre-intervention income year of 2018 and defined as follows: *age* is the taxpayer's age measured in years, *fdiv* is foreign dividends (1,000s SEK), *capinc* is capital income (1,000s SEK), *tax* is total tax paid (1,000s SEK), *earn* is earnings including labor income, sick pay, pension, etc. (1,000s SEK), and $finc \equiv \mathbf{1}[has\ foreign\ income]$ is a categorical 1/0 variable based on a check box in the tax declaration.

4 Estimation

A drawback with the rerandomization strategy is that the mean difference estimator is no longer asymptotically normally distributed (Li, Ding, and Rubin, 2018). However, Li and Ding (2019) showed that standard asymptotic inference can be conducted using ordinary least squares (OLS) by regressing the outcome on the covariates and the treatment together with the Eicker-Huber-White (EHW) robust standard error estimator (Eicker, 1967; Huber, 1967; White, 1980). To be specific, let $\boldsymbol{x}_i$ be the $K \times 1$ vector of covariates used in the Mahalanobis distance for individual i, W_i be the treatment indicator, and Y_i the outcome. The treatment effect is the estimated coefficient on W_i in the ordinary least squares (OLS) regression of Y_i on W_i, $\boldsymbol{x}_i$ and $W_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$, where $\bar{\boldsymbol{x}}$ is the vector of sample means of the covariates. To construct asymptotically valid confidence intervals one should use the EHW robust standard error estimator. However, the results from Zhang and Johansson (2019) suggest that for sample sizes below 400 the EHW estimator is downward biased. The HC2 covariance matrix estimator is shown to have good small-sample performance. In this estimator the OLS residuals $\widehat{u}_i$ used to estimate the EHW covariance matrix are replaced with $\widehat{\epsilon}_i = \widehat{u}_i/\sqrt{1-h_i}$, where h_i is the ith diagonal element of the projection matrix (for details see MacKinnon, 2013).

A balanced design simplifies the analysis and the tests for effect differences across strata. The reason is that the four effects estimated in the saturated linear regression model are the same as the four between-group differences in mean estimates, which are unbiased (Schultzberg and Johansson, 2019). This holds also in a rerandomization design.

Following our pre-analysis plan (Angelov and Johansson, 2020), the baseline analysis

is conducted by estimating

$$\begin{aligned} Y_i =& \alpha_0 + \tau W_i + \alpha_w \mathbf{1}[woman = 0] + \alpha_c \mathbf{1}[compliant = 1] \\ &+ \alpha_{wc} \mathbf{1}[woman = 0] \times \mathbf{1}[compliant = 1] + \boldsymbol{\beta}_0' \widetilde{\boldsymbol{x}}_i + \boldsymbol{\beta}_1' W_i \widetilde{\boldsymbol{x}}_i + \varepsilon_i, \end{aligned} \tag{3}$$

where $\widetilde{\boldsymbol{x}}_i \equiv \boldsymbol{x}_i - \bar{\boldsymbol{x}}$. Asymptotic inference will be performed using the HC2 covariance matrix.

We have two outcomes (*capinc* and *tax* measured in 1000s SEK) and the test for overall effect in percent for each outcome is $H_0 : \tau = 0$ against the alternative $H_1 : \tau > 0$. We let the overall risk level for judging whether we have an effect or not to be 5% which means that each single test will be conducted at the 2.5% risk level.

5 Results

5.1 Baseline results

Table 2 contains results from estimating equation (3) on the the sample of 998 individuals (out of 1,000) that had filed their taxes by November 26, 2020.[10] The point estimates show an increase of around 5,225 SEK in capital income as a result of the intervention and the corresponding number for final tax paid is a decrease with about 3,856 SEK. None of the point estimates is statistically significant at the 2.5% and the p-values are high (0.23 for *capinc* and 0.76 for *tax*). Thus, we find no evidence of an average effect of the intervention among all treated individuals.

In order to extract as much information as possible from the collected data, we provide an exploratory, hypothesis-generating analysis in the next subsection.

5.2 Exploratory analysis

In this subsection, we explore the data along the four hypotheses on effect heterogeneity described in the pre-analysis plan accompanying this paper. In the pre-analysis plan we stated that we would only test for effect heterogeneity across groups if the parameter estimate $\widehat{\tau}$ from (3) was found to be statistically significant at the 2.5 percent level. The reason for this restriction was to have control over the size of the test and at the same time to have power of detecting the ITT-effect, given that even 'reading the digital message' could have a small effect size.

[10] Out of the 998, 499 were treated and 499 were untreated. In other words, there are no signs of a selection by treatment status into being late or not filing a tax declaration.

Table 2: Baseline specification

	$capinc_{2019}$	tax_{2019}
W (treatment effect)	5.225	−3.856
	(4.335)	(12.760)
woman	−6.108	14.356
	(5.498)	(21.508)
compliant	−9.768	34.853**
	(8.728)	(16.148)
woman × *compliant*	2.537	−15.505
	(6.595)	(31.539)
age	−0.076	−0.896
	(0.110)	(0.627)
$fdiv$	0.011	3.973
	(0.370)	(3.194)
capinc	1.339***	−0.589
	(0.316)	(1.029)
earn	−0.003	0.238***
	(0.006)	(0.091)
tax	0.008	0.511***
	(0.010)	(0.188)
$finc$	8.343	61.637
	(7.687)	(60.878)
$W \times age$	0.405*	0.681
	(0.223)	(1.084)
$W \times fdiv$	1.034	−6.818*
	(1.018)	(3.856)
$W \times capinc$	−0.739	0.608
	(0.619)	(1.296)
$W \times earn$	0.025	0.012
	(0.017)	(0.125)
$W \times tax$	−0.010	0.023
	(0.021)	(0.246)
$W \times finc$	−9.955	−39.750
	(8.927)	(71.631)
Intercept	15.860***	295.468***
	(5.217)	(13.106)
Observations	998	998
R^2	0.031	0.553
Adjusted R^2	0.015	0.546

Note: *p<0.1; **p<0.05; ***p<0.01. Treatment status is denoted by W. Outcomes are valued during the income year of 2019 and are expressed in 1000s SEK which is approximately equal to 97 EUR. The covariates are demeaned and measured during the income year of 2018. The measure units are as follows: *age* is measured in years, $finc$ is valued 0 or 1, and the rest of the covariates are expressed in 1,000s SEK. The results are based on OLS-estimation of equation (3) with HC2 standard errors shown in parentheses.

We have two outcomes and thus two main effects. With an additional four heterogeneous effects on two outcomes, a total of ten tests was of interest. Using the Bonferroni correction, the individual tests would have been at 0.5 % level in order to have an overall risk of 5 %. The implication of this is that properly testing of the hypothesis explored in this subsection would have to be done in future research with new data.

Moreover, an issue that should have been given more attention in the pre-analysis plan

was the choice of model for the heterogeneity analyses. Although it is reasonable to test for the main effect of the intervention using a linear model as was done in the previous section, a log-linear model is arguably more appropriate when discussing effect heterogeneity across strata. Specifically, given the observed pre-intervention mean differences in the outcome variables *capinc* and *tax* (see Table 1), it does not seem meaningful to measure effect differences in levels. Instead, it is reasonable to explore effect heterogeneity with respect to percentage effects, which leads to a log-linear model.

5.2.1 Functional form and outliers

To this end, in the exploratory analysis to follow, we consider two types of log-linear specifications. As there are zeros in the outcome variables (23.1% in *capinc* and 3.4% in *tax*), taking logs is not feasible. One possibility is a log approximation using the inverse hyperbolic sine (IHS) transformation for the outcome variables (see Burbidge et al. 1988). The IHS-transformation of a variable Z is given by $IHS(Z) = \log(Z^2 + (Z^2 + 1)^{1/2})$ and the interpretation of the effect estimate from the OLS-regression is approximately as a percentage effect, i.e., the same as if Z were logged.

Another possibility that we examine is the log-linear model of the expected value, that is

$$\log E[Y|\boldsymbol{g}] = \boldsymbol{\delta}'\boldsymbol{g}, \tag{4}$$

where $\boldsymbol{\delta}$ is a parameter vector estimated by maximum likelihood and $\boldsymbol{g}$ is a vector containing a constant and all covariates including the treatment variable. For estimation, we use Pseudo Maximum Likelihood (PML, see Gourieroux, Monfort and Trognon 1984) using a HC2-robust covariance matrix. The requirement for consistency is that the mean, which in this case is the linear projection, is correctly specified.

In Table 3, we provide estimates of the average percentage treatment effect from the IHS and the log-linear model based on the covariate set from equation (3). The results for *capinc* and *tax* for the two different models are presented in the first four columns of the table. As was the case in Table 2, the point estimates for *tax* are negative and imprecise (-3.4% using IHS and -1.2% using PML). For *capinc*, the point estimates are 13.7% (IHS) and 44.1% (PML) and more precise than was the case in the model estimated in levels.

A drawback with the IHS-transformation is that it is not invariant to the level of the transformed variable. In Table 2, as in the two previous tables, *capinc* and *tax* were expressed in 1000s SEK. Expressing the outcome variables in SEK before transforming them by IHS alters the effect estimates considerably: from 13.7% to 22% for *capinc* and from -3.4% to -14.1% for *tax*.[11] As the choice of units is arbitrary but obviously not innocuous, we consider only the log-linear model in the remainder of the section.

As a final step before turning to the effect heterogeneity analysis, we have examined

[11] The complete results are available upon request from the authors.

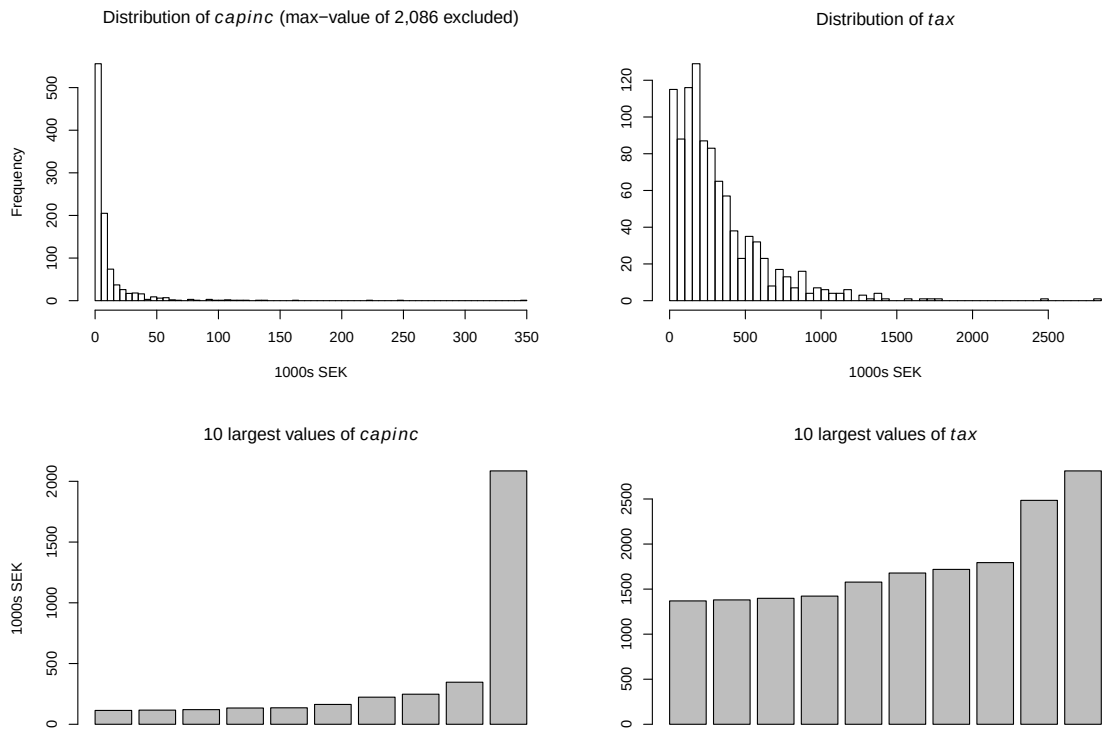

Figure 2: The distributions of the outcome variables

the outcome variables with respect to extreme values. For both *capinc* and *tax*, the distributions are skewed with extreme values on the right. This is expected given the nature of the variables. However, we wish to make sure that a single observation does not alter the results significantly. Figure 2 shows the distributions of *capinc* and *tax*. The largest value of *capinc* is slightly above two million SEK and it is not meaningful to include it in the histogram for *capinc* (see the top-left panel of the plot). This observation is for a previously non-compliant man who was treated in the intervention. The second-largest value for *capinc* is considerably lower (about 350 thousand SEK). To illustrate how extreme the maximum value is, the bottom-left panel of Figure 2 depicts the ten largest values for *capinc*. It is clear from the bar plot that the maximum value stands out considerably from the rest of the observations. As seen in the top-right and bottom-right panels of Figure 2, the maximum value of *tax* is in line with the immediately preceding values.

To see how sensitive the results are to the inclusion of the largest value, in the last two columns of Table 3, we present effect estimates using the log-linear model where we have removed a single observation pertaining to the maximum value of *capinc* and *tax*, respectively. For *tax*, the effect estimate changes from -1.2% to -5.1% and is still imprecisely estimated. For capinc, the point estimate drops from 44.1% to 14.4% and is no longer statistically significant at the 5%- or even 10%-level. Since removing the

Table 3: Alternative specifications for the estimation of the main effect

	All observations				Maximum value removed	
	OLS with IHS		PML		PML	
	$capinc_{2019}$	tax_{2019}	$capinc_{2019}$	tax_{2019}	$capinc_{2019}$	tax_{2019}
W (treatment effect)	0.137*	−0.034	0.441**	−0.012	0.144	−0.051
	(0.072)	(0.073)	(0.181)	(0.056)	(0.149)	(0.043)
woman	−0.280**	−0.063	−0.797	−0.017	−0.236	0.022
	(0.119)	(0.136)	(0.535)	(0.088)	(0.336)	(0.080)
compliant	0.884***	0.337***	−0.483	0.212***	0.284	0.244***
	(0.110)	(0.091)	(0.528)	(0.059)	(0.214)	(0.052)
woman × *compliant*	0.188	−0.002	0.654	0.003	−0.150	−0.039
	(0.161)	(0.168)	(0.658)	(0.114)	(0.364)	(0.107)
age	0.008	0.004	−0.003	−0.003	−0.006	−0.003
	(0.005)	(0.005)	(0.013)	(0.003)	(0.014)	(0.003)
$fdiv$	0.047***	0.011	0.019	0.007	0.029*	0.008
	(0.016)	(0.012)	(0.018)	(0.010)	(0.017)	(0.010)
capinc	0.067***	−0.002	0.061***	−0.003	0.049***	−0.004
	(0.012)	(0.006)	(0.012)	(0.004)	(0.008)	(0.004)
earn	0.0004	0.003***	−0.0003	0.001**	−0.0005	0.001**
	(0.0003)	(0.001)	(0.001)	(0.0004)	(0.0005)	(0.0004)
tax	−0.001	−0.002*	0.001	0.0005	0.001	0.0005
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$finc$	0.132	−0.375	0.569	0.170	0.564	0.172
	(0.240)	(0.281)	(0.434)	(0.239)	(0.420)	(0.238)
$W \times age$	0.010	0.003	0.026	0.003	0.028*	0.001
	(0.008)	(0.008)	(0.016)	(0.005)	(0.017)	(0.004)
$W \times fdiv$	0.010	−0.041**	0.016	−0.018	0.007	−0.017
	(0.022)	(0.019)	(0.025)	(0.013)	(0.021)	(0.013)
$W \times capinc$	−0.013	0.008	−0.035	0.003	−0.010	0.005
	(0.012)	(0.009)	(0.025)	(0.005)	(0.009)	(0.005)
$W \times earn$	−0.0001	−0.001	0.001*	−0.0003	0.001	−0.0001
	(0.0004)	(0.001)	(0.001)	(0.0005)	(0.001)	(0.0005)
$W \times tax$	−0.00005	0.002	−0.001	0.001	−0.002**	0.0003
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
$W \times finc$	−0.083	0.342	−0.679	−0.142	−0.505	−0.143
	(0.299)	(0.363)	(0.626)	(0.277)	(0.457)	(0.279)
Intercept	1.470***	5.675***	2.373***	5.453***	1.924***	5.430***
	(0.088)	(0.078)	(0.294)	(0.051)	(0.159)	(0.047)
Observations	998	998	998	998	997	997

Note: *p<0.1; **p<0.05; ***p<0.01. The effect estimates can be approximately interpreted as percentage effects. Outcomes are valued during the income year of 2019. The covariates are demeaned and valued pre-intervention during the previous income year. The measure units are as follows: *age* is measured in years, fin is valued 0 or 1, and the rest of the variables are expressed in 1,000s SEK which is approximately equal to 97 EUR. HC2 standard errors are shown in parentheses.

maximum value considerably affects the results, we think it is unjustified to keep this single observation in the estimation sample when exploring effect heterogeneity in the next subsection. For full transparency, we will however include the results for the complete sample in an appendix.

5.2.2 An exploratory analysis of effect heterogeneity

To fix ideas on effect heterogeneity with respect to previous compliance, let $\widetilde{Y}_{it}$ denote the latent, true, value of an outcome (e.g., *capinc*) for individual i at time t, and let $\lambda_{it} \in [0, 1]$ denote the level of compliance. When $\lambda_{it} = 0$, the taxpayer is fully non-compliant and consequently does not report any income, and on the other extreme, $\lambda_{it} = 1$ means that all income is correctly filed. For each individual, we observe $Y_{it} = \lambda_{it}\widetilde{Y}_{it}$. Assume a latent income $\widetilde{Y}_{it} = \widetilde{Y}_i^*$ and therefore $Y_{it} = \lambda_{it}\widetilde{Y}_i^*$. Fixing the level of the latent variable over time means that there is no relationship between the level of compliance and the revealed capital income. The assumption is not strong for individuals with stable income streams but, of course, less realistic for individuals with large fluctuations in revealed capital incomes. Suppose we have access to a measure of past compliance, $\hat{\lambda}_{it-1}$. In the following, we consider two broad, simplified mechanisms leading to a certain level of compliance. At the one extreme, the taxpayer decides on their compliance at some level λ_i^* (e.g., starting when entering the labor or capital market) and keeps this level throughout unless an external shock occurs (i.e., an audit or other intervention from The Swedish Tax Agency). In absence of receiving the digital message this means that knowing $\hat{\lambda}_{it-1}$ implies knowing λ_{it}. Under this scenario we expect a larger positive effect (i.e., compliance to the treatment) of previously non-compliant taxpayers than for the compliant taxpayers. The reason is simply that if one is a previous compliant, there is no reason to respond upon reading the digital message. At the other extreme, each year the taxpayer makes completely independent decisions about the compliance level which means that we would not observe any effect heterogeneity with respect to previous compliance.

A large experimental literature has shown systematic differences in risk preferences between men and women. The results are broadly consistent with women being more risk averse than men (see Croson and Gneezy, 2009 and Eckel and Grossman, 2008). Most of the experimental work consists of comparing how men and women (mostly college students) value risky gambles or choose between gambles with often small stakes. In a study combining field and laboratory evidence, Antonovics et al. (2009) illustrate the importance of understanding the gender differences for larger stakes. Finally, there is clear empirical evidence that women on average have a lower level of financial literacy than men in Sweden (Almenberg and Dreber, 2015) and in many other countries (Hasler and Lusardi, 2017).

A potential positive effect (i.e. increased compliance) of receiving the digital message could be due to an increased knowledge about how to properly file foreign income stemming from information provided in the online app or due to an increase in the perceived risk of non-compliance detection. Both of these effect mechanisms should imply that the effect from reading the digital message among women is higher than the effect

among men. First, if women have on average a lower level of financial literacy, information about how to file foreign income could affect women more than men. Second, the empirically well established gender gap in risk aversion should imply that women, being more risk averse, are more likely to react to reading the message than men. In particular, the second sentence in the message could affect risk-averse individuals more, since it is unclear exactly how much and what type of information The Swedish Tax Agency has ("The Swedish Tax Agency has obtained information from a foreign tax authority that you have received dividends or interest from abroad during 2018."). Of those two potential sources of effect heterogeneity, we believe that the risk aversion mechanism would be more significant than financial literacy for a potential gender difference in effect as declaring foreign dividends is relatively uncomplicated. Of course, a potential effect could also be due to some mechanism other than the two above mentioned, for instance an unintentional moral nudge. However, also in this case there is reason to believe that women should react more to the intervention. For instance, Fumagalli et al. (2010) document a gender difference in moral judgments, whereby females moral reasoning appears to be directed toward placing higher value on social relationships and on fulfilling other individuals' expectations.

To sum up, we see two potential sources of effect heterogeneity with respect to gender and they both point toward a greater positive effect from receiving the digital message on compliance among women. Of course, should we find any effect heterogeneity, we cannot know which of those sources would dominate given the data in this study. It should also be noted that the study closest to ours, Bott et al. (2020), contains a heterogeneity analysis with respect to, among other groups, gender. The general conclusion drawn in Bott et al. is that the main findings in the paper are robust across subgroups and they do not find any effect differences between women and men. Furthermore, with respect to previous compliance, if any, the heterogeneity would mean a larger positive effect of receiving the digital message for previously non-compliant taxpayers. To explore whether there is any support for this in data we estimate a version of the model described in the pre-analysis plan (Angelov and Johansson 2020). The previously specified linear model is displayed below:

$$
\begin{aligned}
Y_i &= \alpha_0 + \tau W_i + \alpha_w \mathbf{1}[woman = 1] + \alpha_c \mathbf{1}[compliant = 1] \\
&\quad + \alpha_{wc} \mathbf{1}[woman = 1] \times \mathbf{1}[compliant = 1] + \tau_w \mathbf{1}[woman = 1] W_i + \tau_c \mathbf{1}[compliant = 1] W_i \\
&\quad + \tau_{wc} \mathbf{1}[woman = 1] \times \mathbf{1}[compliant = 1] W_i + \boldsymbol{\beta}_0' \widetilde{\boldsymbol{x}}_i + \boldsymbol{\beta}_1' W_i \widetilde{\boldsymbol{x}}_i + \varepsilon_i \\
&\equiv \boldsymbol{\beta}' \boldsymbol{g}_i + \varepsilon_i, \qquad (5)
\end{aligned}
$$

where $\boldsymbol{\beta}$ is a parameter vector and $\boldsymbol{g}_i$ is a vector consisting of all covariates including a constant and the treatment variable. As explained previously, in this subsection we use the log-linear model $\log E[Y|\boldsymbol{g}] = \boldsymbol{\theta}'\boldsymbol{g}$, where $\boldsymbol{\theta}$ is a parameter vector estimated with

PML. Inference is performed using a HC2-robust covariance matrix.

The results by stratum are displayed in Table 4. As in the last two columns of Table 3, we have excluded the observations for the maximum value of *capinc* and *tax*, respectively. For ease of exposition, we only present the effect estimates along with 95% confidence intervals.[12] With respect to previous compliance, the point estimates for our main outcome variable *capinc* show a 141% and 50.4% increase in compliance for previously non-compliant women and men, respectively. For previously compliant women and men, the point estimates suggest a negative effect on compliance (-38.2% and -24.2% for men and women, respectively). As seen from the table, the 95% confidence intervals are large and do not cover zero except for non-compliant women. The results for *tax* show negative effect estimates ranging between -14.8% (for non-compliant women) and 5.8% (for compliant men). Each of the 95% confidence intervals for *tax* covers zero. Even for non-compliant taxpayers, we note that the estimates on *tax* are negative (-14.8% and -13.5% for women and men, respectively) while they were positive for *capinc*. If we take these point estimates at face value, they seem so suggest that a sort of offsetting behavior on behalf of the taxpayers. For instance, it might bet that treated taxpayers compensate for their increased capital income following the intervention by reporting less of other income sources such that the tax paid is at least not increased. Of course, this is purely speculative given the low precision in our results, but it would perhaps be interesting to study in the future.

As noted above, the results in Table 4 are from estimation samples in which one observation (for the maximum value of *capinc* and *tax*, respectively) has been removed, since the value for *capinc* in particular is very large in comparison to the rest of the distribution and affects the results considerably. For transparency, we provide the results using the complete estimation sample in Appendix B (see Table 6). The numbers change quite a bit when the whole sample is used. The only qualitatively important change is for non-compliant men and *capinc* where the point estimate increases from 50.4% to 137.5% and the 95% confidence interval does not cover zero in Table 6. This is expected due to the large outlier in *capinc* (see Figure 2) which was for a non-compliant man. The main result, that is, a positive effect estimate with a confidence interval not covering zero for non-compliant women, holds also for the whole estimation sample and the parameter estimate for this group has hardly changed (149.1% compared to 141%). Finally, Table 6 Appendix B presents results from OLS-regression using equation 5. The effect sizes in different strata are not easily comparable in this case, but the main result holds qualitatively: the point estimate for non-compliant women is positive for *capinc* (10,400 SEK) and the corresponding 95% confidence interval does not cover zero. Note that this point estimate is large in relation to the mean capital income for non-treated non-compliant women which is 2,157 SEK, but the corresponding estimate from the

[12]The complete set of results is available from the authors upon request.

Table 4: Treatment effect estimates from the specification with interaction effects based on a log-linear model

	Women		*Men*	
	Compliant	*Non − compliant*	*Compliant*	*Non − compliant*
	capinc			
Treatment effect	-0.382	1.41	-0.242	0.504
CI	[-0.877, 0.112]	[0.511, 2.308]	[-0.648, 0.164]	[-0.132, 1.141]
	tax			
Treatment effect	-0.142	-0.148	0.058	-0.135
CI	[-0.381, 0.096]	[-0.409, 0.114]	[-0.053, 0.17]	[-0.291, 0.022]

Note: *p<0.1; **p<0.05; ***p<0.01. The effects were estimated using transformed PML-estimates from a log-linear model based on the covariate set from equation (5). For instance, the estimate for non-compliant men is given by $\widehat{\tau}$, the estimate for non-compliant women is given by $\widehat{\tau} + \widehat{\tau}_w$, etc. The effect estimates can be approximately interpreted as percentage effects. Observations for the largest value in *capinc* and *tax*, respectively, have been removed from the estimation sample (997 out of 998 observations left). 95% confidence intervals formed using the HC2 covariance matrix are shown in square brackets.

log-linear model also points to an economically significant effect (141%) . All other confidence intervals cover zero.

In summary, the point estimates for *capinc* are in line with our previous discussion on the effect heterogeneity with respect to previous compliance and gender. Our expectations were of no positive effect of the intervention for previously compliant taxpayers which is what the table suggests (negative point estimates for both compliant women and men). We also conjectured that the effect for women would be larger in magnitude than for men, and this is also supported by the point estimates for non-compliant taxpayers. The only case of a confidence interval that does not cover zero is for non-compliant women, and this finding is robust to functional form as it is supported both in the log-linear PML-model and in linear regression based on OLS. Given that this is the stratum for which we expected the largest effect on compliance from receiving the digital message, this result is suggestive of an effect. In light of our expectations about the effect differences, the signs and magnitudes of the estimates for *tax* do not make much sense. In this case, all confidence intervals cover zero, which is in line with the negative and imprecise estimate for *tax* in table 2. Thus, we find suggestive evidence of an effect on *capinc*, but not on *tax*.

6 Efficiency gains from the statistical design

Stratification (see e.g. Imbens and Rubin 2015) and rerandomization (Morgan and Rubin 2012) can be shown to be in general more efficient than standard randomized experiments (e.g. tossing a coin in Bernoulli trials or complete randomization with a fixed number of

treated and controls). Schultzberg and Johansson (2019) show that the combination of stratification and rerandomization in general is more efficient than only stratification.

As the efficiency gains depend on both the correlation of the covariates with the outcome and the degree of effect heterogeneity, the relative efficiency gains will depend on the context of the application. Since the experimental design is somewhat complex it is also of interest to compare its efficiency against post-stratification, that is, the use of regression as an estimator in a standard randomized experiment. The present data provides an excellent opportunity of studying these issues in the context of experiments analyzing tax compliance behavior. The reason for this is the possibility of conducting a simulation involving a hypothetical experiment on the same population as was randomly sampled to be part of the experiment.

Thus, we perform the simulation using the subset of non-treated individuals from the original experiment. To recap, the experiment was based on 2,697 individuals (see section 3), from which a 1,000 individuals were sampled and 500 were treated. In the simulation, we exclude the 500 treated and remove those few who had not filed their declaration as of July 27, 2020, resulting in a sample of 2,190 taxpayers which we denote the *simulation sample*. As in the original experiment, we divided the sample in a compliant and non-compliant group. The first group which we denote (a) consists in 1,509 individuals and the second, (b) of 681 individuals.

In each replication of the simulation ($r = 1, 2, \ldots, 1,000$), the following steps were performed:

1. Draw two simple random samples, each of size 500, from (a) and (b) respectively, resulting in 1,000 individuals.

2. Draw 500 individuals assumed to be treated according to three different experimental designs:

   A. *Rerandomization within stratum*: Perform steps 3 and 4 described in Section 3.

   B. *Complete randomization within stratum*: Create the four strata and randomly allocate 50% to be 'treated' within each stratum.

   C. *Complete randomization*: Randomly allocate 500 individuals to be 'treated'.

3. Using the simulated data according to each design, estimate equation (3) with and without covariates and store the corresponding estimate for τ, $\widehat{\tau}_r$.

The simulation thus results in 12 simulated distributions of $\widehat{\tau}$, each of size 1,000 (three different designs, two outcome variables, and two versions with respect to inclusion of covariates).

Table 5: Simulation: standard deviations of point estimates under the null of no effect

	capinc	*tax*
	I. No covariates	
A: rerandomization within stratum	2.8927	12.1143
B: complete randomization within stratum	2.9861	17.8315
C: complete randomization, no stratification	3.0741	18.8310
$\%\Delta:\ 100 \times (B - A)/A$	3.2310	47.1940
$\%\Delta:\ 100 \times (C - A)/A$	6.2700	55.4452
$\%\Delta:\ 100 \times (C - B)/B$	2.9439	5.6056
	II. Covariates included	
D: rerandomization within stratum	2.5993	12.8921
E: complete randomization within stratum	2.6041	12.7851
F: complete randomization, no stratification	2.7297	13.5777
$\%\Delta:\ 100 \times (E - D)/D$	0.1846	-0.8295
$\%\Delta:\ 100 \times (F - D)/D$	5.0186	5.3184
$\%\Delta:\ 100 \times (F - E)/E$	4.8251	6.1993

Note: The table shows standard deviations along with percentage differences between standard deviations for point estimates of τ from equation (3) with and without covariates, respectively. The simulation is performed on non-treated individuals, i.e., under the null hypothesis of no effect.

The results from the simulation are presented in Table 5. For each of the 12 combinations of type of design, outcome variable, and presence or absence of covariates, we present the standard deviation of $\widehat{\tau}$. In addition, the pairwise difference in percent between different cells is presented as a measure of relative efficiency. Our interest lies in the relative efficiency of the most efficient design according to theory (rerandomization within stratum) compared to each of the two remaining designs.

The total gain from stratification and rerandomization is displayed in rows 5 and 11 in Table 5. Without the regression adjustment, the total efficiency gains for capital income and income tax are 6.27 and 55.44 percent, respectively. The corresponding efficiency gains when the analysis is based on regression adjustment are only 5.02 and 5.32 percent, respectively. Furthermore, the table displays efficiency gains from (i) rerandomization in contrast to the stratified design for the specification but only in the case without regression adjustment and (ii) stratification in comparison with complete randomization. (i) is seen by comparing rerandomization within stratum (A) with complete randomization within stratum (B). The efficiency gain is expressed in percent in rows 4 and 10. Without the regression adjustment (row 4), the efficiency gains for capital income and tax are 3.23 and 47.19 percent, respectively. The corresponding efficiency gain when the analysis is based on regression adjustment (row 10) is only 0.18 percent for *capinc*, and there is an efficiency loss for *tax* (-0.83 percent). (ii) can be seen by relating the standard

deviations of $\widehat{\tau}$ in B to those in C (row 6), and in F to those in E (row 12). The results reveal that stratification provides efficiency gains over complete randomization both in the specification without covariates (a gain of 2.94% for *capinc* and 5.61% for *tax*) and with covariates (a gain of 4.83% for *capinc* and 6.20% for *tax*).

It might be interesting to compare the efficiency gains displayed in Table 5 with the theoretical efficiency gains. The percentage reduction in variance in a rerandomization design is $100 \times R^2(1-\nu_a)$ (cf. equation (1)). As $\nu_a = Pr(\chi^2_8 \leq 0.17)/\Pr(\chi^2_6 \leq 0.17)) = 0.12$, the efficiency gains in our design should be approximately $100 \times R^2(1-0.12)$ within each stratum. From Table 2, we have that the overall R^2 for capital income and income tax are $R^2 = 0.031$ and $R^2 = 0.553$, respectively. This amounts to an expected gain of 2.73% and 48.66%, respectively. The observed efficiency gains displayed in panel I of Table 5 against stratification were 3.23% and 47.19%, which is very close to the theoretical numbers.

To sum up, we find efficiency gains from both stratification and rerandomization relative to complete randomization. The gains are far greater in a scenario where no covariates are included in the regressions, compared to including the same covariates as were used in the rerandomization routine. Interestingly, our results suggest that using the same covariates in the estimation as would have been used in the rerandomization design is essentially as efficient as estimation based on a rerandomization design without covariates (compare design F in panel II with design A in panel I). Thus, it can be argued that the choice of relevant covariates (including stratification variables) is paramount, irrespective of whether one chooses an experiment design involving stratification and rerandomization or uses the covariates for post-stratification. Finally, the more relevant the covariates are for the outcome, as is the case for *tax* with an $R^2 = 0.553$ compared to $R^2 = 0.031$ for *capinc* (see Table 2), the larger the efficiency gain is from rerandomization and stratification. However, this only holds in the case without regression adjustment: there is essentially no difference in efficiency gain between *capinc* and *tax* in panel II, while the difference is substantial in panel I.

7 Summary

The purpose of this study has been to measure the effect of a digitally disseminated message on the subsequent compliance of taxpayers who had foreign dividends during the income year of 2018. Data on foreign dividends was obtained through the automatic international exchange of information (CRS). We performed a field experiment consisting of a message sent to a group of taxpayers in March 2020, about six weeks before the final day to file income taxes for the income year of 2019. The message contained two chief points: that The Swedish Tax Agency had received CRS-information about the specific taxpayer's foreign income during the income year of 2018, and that the tax agency had

developed an online app which facilitates filing in the correct amount of foreign income to be filed for 2019.

We have not found any clear indication that receiving the digital message leads to an increase in the usage of the online app, although it should be noted that this is solely based on a descriptive analysis as we do not have access to visitor data on the individual level. If this conjecture is correct, any positive effect of the intervention on tax compliance would imply that the taxpayers are in fact well-informed and that a potential change in behavior is not due to the extra service provided by the app but due to an increased perceived detection risk. In particular, the following sentence indicates that the Tax Agency has some information on the individual taxpayer: *"The Swedish Tax Agency has obtained information from a foreign tax authority that you have received dividends or interest from abroad during 2018."* Of course, compliance behavior could be potentially affected via a mechanism other than the two mentioned above. For instance, merely receiving a message could function as an unintentional nudge toward better compliance. As the message does not include any explicit moral or societal appeal it is however difficult to argue that there is any nudge content in the message, besides it being a message sent by the Swedish Tax Agency.

Turning to the effect of the intervention on compliance, the point estimates for capital income and final tax are 5,225 SEK (p-value = 0.23) and -3,856 SEK (p-value = 0.76), respectively. These imprecise estimates imply that we find no evidence of an average effect of the intervention among all treated individuals. In an exploratory analysis, we examine the data from the experiment further along two hypotheses. First, departing mainly from the research literature documenting a gender gap in risk aversion, we hypothesize that women should react more than men to the intervention as they are on average more risk averse. Second, since compliant taxpayers cannot further increase their compliance, we expect a larger effect magnitude among taxpayers who were deemed to be non-compliant during the previous income tax year. When using capital income as outcome variable, we find support for both of these conjectures in the data.

When we explore the effect heterogeneity by gender and previous compliance, we measure percentage effects for ease of group comparisons. We find that the point estimates for capital income are higher in magnitude among the previously non-compliant taxpayers and this is true for both women and men. Furthermore, among the previously non-compliant, the effect estimate on capital income among women is larger in magnitude than among men: 141 percent (95% CI: [51.1%, 230.8%]) compared to 50.4 percent (95% CI: [-13.2%, 141.1%]). The only case of a confidence interval that does not cover zero is for non-compliant women and this result also holds in a sensitivity analysis where we use OLS to estimate a linear model instead of PML on the log-linear model. Given that this is the group for which we expected the largest effect on compliance from receiving the digital message, this result is suggestive of an effect. Taken at face value, the estimated

effect for non-compliant women 141 percent is large, especially taking into consideration the non-intrusive nature of the intervention. The estimated effect using a linear model appears even larger when it is related to the mean capital income among non-treated: the effect on *capinc* for non-compliant women is 10,400 SEK and the mean capital income for non-treated non-compliant women is 2,157 SEK. It is hard to draw any conclusions from the exploratory analysis when we use final tax as the outcome variable, which is in line with the previously mentioned imprecise average effect estimate for tax.

The point estimates for non-compliant women and men discussed above are large in economic terms. Thus, the relatively mild low-cost intervention seems promising policy-wise. For future research, we therefore think that it is important to collect a new and larger experimental data set and see whether the results hold. As expected, our data seems to confirm the expectation of a lack of effect among previously compliant taxpayers. Indirectly, this provides evidence that our chosen measure of previous compliance contains important information about current compliance. This has practical consequences: a tax authority could use a measure similar to ours in order to target information toward the group where it is most effective, that is, taxpayers who are estimated to be non-compliant. Similarly, to increase the statistical precision in future field experiments along the lines suggested above, it is reasonable to use the previous compliance measure in order to focus solely on non-compliant taxpayers. Finally, we believe that the apparent relevance of our chosen previous compliance measure suggests that it could be used also in risk-based tax audits, as there appears to be quite strong time correlation in individual compliance.

As mentioned in the introduction, the number of treated taxpayers had to be restricted to 500 individuals. As our ambition was to study how the effect varies among different groups of taxpayers, choosing an efficient experimental design was imperative. To this end, we used stratification and rerandomization within each stratum. The efficiency gains from this design compared to complete randomization depend on both the correlation of the covariates with the outcome and the degree of effect heterogeneity. Thus, the relative efficiency gain will depend on the specific application. As the design is somewhat complex to implement in practice, it is of interest to compare its efficiency relative to post-stratification, that is, the use of regression as an estimator in a standard randomized experiment. We were able to study these issues because of the opportunities provided by the unique data used in the present study. In short, we performed a simulation involving hypothetical experiments on the same population as was randomly sampled to be part of the original experiment. The results show that the total efficiency gains from our chosen design compared to complete randomization are substantive for final tax (55.4 percent) but less so for capital income (6.3 percent) if no covariates are included in the regression. The difference in efficiency gain between the two outcomes is in line with the observation that the covariates are much more relevant for the tax outcome ($R^2 = 0.553$) compared to capital income ($R^2 = 0.031$). There are gains also when we use the same covariates in the

estimation as were used in the rerandomization design, but the numbers are more modest and range between two and four percent. In this case, the above mentioned difference in efficiency gain between the two outcomes is not present. Finally, we find that there are gains both from stratification and rerandomization.

How is a practitioner to use these results in an actual application? Making the assumption that there is no extra cost for implementing a rerandomization design relative to a standard randomized experiment, the results suggest that a rerandomization design should be chosen as it is more efficient. However, assuming no extra cost for rerandomization might be unwarranted. In our experience, in a large organization, there might be administrative costs beyond the relatively small cost of writing the source code for the experiment, meaning that a standard randomized experiment would be a more practical option. For this and similar applications, our conclusion is that regression-based post-stratification along with including relevant covariates might be a reasonably efficient alternative in the case where proper stratification and rerandomization is costly.

It should be stressed that if one chooses the covariates in the regression models after the experiment is conducted, the resulting inference may be flawed as the researchers are prone to searching for statistically significant results (see Mutz, Pemantle and Pham (2019) for a recent discussion of p-hacking in randomized experiments). To avoid difficult post-experiment decisions and for transparency, we recommend a careful choice process documented in a pre-analysis plan.

References

Alm, J. (2019). What Motivates Tax Compliance?, Journal of Economic Surveys 33(2), pp. 353–388.

Alm, J., McClelland, G.H. and Schulze,W.D. (1992a) "Why do people pay taxes?", *Journal of Public Economics* 48(1): pp. 21–38.

Alm, J., Cronshaw, M. B. and McKee, M. (1992b) "Tax compliance with endogenous audit selection rules", *KYKLOS*, 1, 27–45.

Alm, J., Jackson, B., and McKee, M. (1992c) "Estimating the determinants of taxpayer compliance with experimental data", *National Tax Journal*, Vol. 45(1), pp. 107–114.

Alm, J., Sanchez, I. and de Juan, A. (1995) "Economic and Noneconomic Factors in Tax Compliance", *KYKLOS*, vol. 48(1), pp. 3-18.

Alm, J., McClelland, G.H. and Schulze, W.D. (1999) "Changing the Social Norm of Tax Compliance by Voting", *KYKLOS*, Vol. 52(2), pp. 141-171.

Alm, J., Bloomquist, K.M., and McKee, M. (2017) "When You Know Your Neighbour Pays Taxes: Information, Peer Effects and Tax Compliance", *Fiscal Studies*, vol. 38(4), pp. 587–613.

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

Almenberg, Johan and Anna Dreber. "Gender, Financial Literacy and Stock Market Participation." Economics Letters, 137: 140-142.

Alstadsæter, A., Johannesen, N., and Zucman, G. (2019), Tax Evasion and Inequality, American Economic Review, 109(6): 2073–2103 https://doi.org/10.1257/aer.20172043

Angelov, N. and Johansson, P. (2020) Using Intelligence from International Tax Cooperation to Improve Voluntary Tax Compliance: Evidence from a Swedish Field Study. AEA RCT Registry. May 04. https://doi.org/10.1257/rct.5800-1.0.

Becker, W., Büchner, H.-J. and Sleeking, PP. (1987), "The impact of public transfer expenditures on tax evasion: an experimental approach", *Journal of Public Economics*, vol 34, pp. 243—252.

Boning, W.C., Guyton, J, Hodge, R.H., Slemrod, J. and Troiano, U. (2018), "Heard it through the grapevine: direct and network effects of a tax enforcement field experiment", NBER WP 24305.

Bott, K.M., Cappelen, A.W., Sørensen, E.Ø., and Tungodden, B. (2020) You've Got Mail: A Randomized Field Experiment on Tax Evasion. Management Science 66(7):2801-2819. https://doi.org/10.1287/mnsc.2019.3390

Burbidge, J.B., Magee, L., and Robb, A.L. (1988) Alternative Transformations to Handle Extreme Values of the Dependent Variable, Journal of the American Statistical Association 83(401), pp. 123–127.

Calvet Christian, R. and Alm, J. (2014) "Empathy, sympathy, and tax compliance", *Journal of Economic Psychology*, vol. 40, pp. 62-82

Casagrande, A., Di Cagno, D., Pandamiglio, A. and Spallone, M. (2015) "The effect of competition on tax compliance: the role of audit rules and shame", *Journal of Behavioral and Experimental Economics* vol. 59 pp. 6—110

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. Annals of Statistics, 41(2):484–507.

Cummings, R.G., Martinez-Vazquez, J., McKee, M. and Torgler, B. (2009) "Tax morale affects tax compliance: Evidence from surveys and an artefactual field experiment", *Journal of Economic Behavior & Organization*, vol. 70 447–457

Dubin, J. (2007) "Criminal Investigation Enforcement Activities and Taxpayer Noncompliance.", *Public Finance Review*, vol. 35 (4), 500–529.

Dubin, J.A., Graetz, M.J. and Wilde, L.L. (1990) "The effect of audit rates on the federal individual income tax, 1977-1986", *National Tax Journal*, vol. 43(4), pp. 395–409.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume I, pages 59–82. University California Press, Berkeley, CA.

Fortin, B., Lacroix, G. and Villeval, M.-C. (2007) "Tax evasion and social interactions", *Journal of Public Economics*, vol. 91, pp. 2089—2112.

Freedman, D. A. (2008a). "On regression adjustments to experimental data."Adv. in Appl. Math. 40 180–193.

Gourieroux, C., Monfort, A., and Trognon, A. (1984) "Pseudo Maximum Likelihood Methods: Applications to Poisson Models", *Econometrica*, vol. 52(3), pp. 701–720.

Fumagalli, M., Ferrucci, R., Mameli, F., Marceglia, S., Mrakic-Sposta, S., Zago, S., Lucchiari, C., Consonni, D., Nordio, F., Pravettoni, G., Cappa, S., and Priori, A. (2010) "Gender-related differences in moral judgments." *Cognitive Processing* vol. 11, pp. 219–226. https://doi.org/10.1007/s10339-009-0335-2

Hasler, A. and Lusardi, A. (2017) The Gender Gap in Financial Literacy: A Global Perspective. Global Financial Literacy Excellence Center, The George Washington University School of Business

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume I, pages 221–233. University California Press, Berkeley, CA.

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

Kleven, H.J., Knudsen, M.B., Kreiner, C.T., Pedersen, S., Saez, E., (2011) "Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark", *Econometrica*, vol. 79 (3), 651–692.

Li, X. and Ding, P. (2019). Rerandomization and regression adjustment. To appear in Journal of the Royal Statistical Society, Series B.

Li, X., Ding, P., and Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment Control experiments. Proceedings of the National Academy of Sciences of the United States of America, 115(37):9157–9162.

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In Chen, X. and Swanson, N. R., editors, Recent Advances and Future Directions in Causality, Prediction, and Specication Analysis: Essays in Honor of Halbert L. White Jr., pages 437–461. Springer, New York.

Meiselmann, B. (2018), "Ghostbusting in Detroit: Evidence on nonfilers from a controlled field experiment", *Journal of Public Economics*, vol. 158, pp. 180–193.

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. Annals of Statistics, 40(2):1263–1282.

Mutz, D. C., Pemantle, R., and Pham, P. (2019). The Perils of Balance Testing in Experimental Design: Messy Analyses of Clean Data. The American Statistician, 73(1):32-42.

OECD and Council of Europe (2011), The Multilateral Convention on Mutual Administrative Assistance in Tax Matters: Amended by the 2010 Protocol, OECD Publishing. http://dx.doi.org/10.1787/9789264115606-en

Ortega, D. and Sanguinetti, P. (2013) "Deterrence and reciprocity effects on tax compliance: experimental evidence from Venezuela", CAF WP 2013/08.

Rubin, D. B. (1980). "Discussion of Basu's paper. "Journal of the American Statistical Association, vol. 75, pp. 591–593.

Plumley, A.H. (1996) "The Determinants of Individual Income Tax Compliance."Internal Revenue Service Publication 1916 (Rev. 11-96) Internal Revenue Service, U.S. Department of the Treasury, Washington, DC.

Pomeranz, D. (2015) "No Taxation without Information: Deterrence and Self-Enforcement in the Value Added Tax", *American Economic Review*, vol. 105(8), pp. 2539–2569.

Schultzberg, M. and Johansson, P. (2019) "Rerandomization: a complement or substitute for stratification in randomized experiments?", Working Paper 2019:4, Department of Statistics, Uppsala University.

Shimeles, A., Zerfu Gurara, D, Woldeyes, F. (2017) "Taxman's Dilemma: Coercion or Persuasion? Evidence from a Randomized Field Experiment in Ethiopia", *American Economic Review: Papers & Proceedings*, 107(5), pp. 420–424

Slemrod, J., Blumenthal, M., Christian, C. (2001) "Taxpayer response to an increased probability of audit: evidence from a controlled experiment in Minnesota", *Journal of Public Economics* 79, pp. 455–483

Soliman, A., Jones, P. and Cullis, J. (2014) "Learning in experiments: Dynamic interaction of policy variables designed to deter tax evasion", *Journal of Economic Psychology*, vol. 40, pp. 175–186.

Spicer, M.W. and Thomas, J.E. (1982) Audit probabilities and the tax evasion decision: an experimental approach, Journal of Economic Psychology vol. 2 p. 241-245.

Tauchen, H.V., Witte, A.D. and Beron, K.J. (1993) "Tax Compliance: An Investigation Using Individual TCMP Data.", *Journal of Quantitative Criminology* 9 (2), 177–202.

Webley, P. (1987) "Audit probabilities and tax evasion in a business simulation", *Economics Letters*, vol. 25, pp. 267-270.

White, H. (1980). Using least squares to approximate unknown regression functions. International Economic Review, 21(1):149–170.

Zhang, J. L.; Johansson, P (2019). A Comparison of Methods of Inference in Randomized Experiments from a Restricted Set of Allocations Working Paper 2019:5 Department of Statistics

Zucman, G. (2013). The Missing Wealth of Nations: Are Europe and the U.S. net Debtors or net Creditors?, The Quarterly Journal of Economics, Oxford University Press, vol. 128(3), pp 1321–1364.

Appendix A: Experimental design, rerandomization and inference

Consider a Randomized Control Trial (RCT) with n units in the sample, indexed by i, with n_1 assigned to treatment and n_0 assigned to control. Let $W_i = 1$ or $W_i = 0$ if unit i is assigned treatment or control, respectively, and define $\mathbf{W} = (W_1, ..., W_n)'$. Furthermore, let $\mathbf{X}$ be the $n \times K$ matrix of fixed covariates in the sample $(\mathbf{x}_i, i = 1, ..., n)$, with sample covariance $cov(\mathbf{X})$.

There are $\binom{n}{n_1} = A$ possible treatment allocation (assignment) vectors labeled $\mathbf{W}^j = (W_1^j, ..., W_n^j)'$, $j = 1, ..., A$, where $A = \text{card}(\mathcal{A})$, i.e., the cardinality of the set $\mathcal{A}$. The Mahalanobis distance for allocation j is

$$M(\mathbf{W}^j, \mathbf{X}) = \frac{n}{4} \widehat{\tau}_X^{j\prime} cov(\mathbf{X})^{-1} \widehat{\tau}_X^j, \ j = 1, ..., A,$$

where

$$\widehat{\tau}_X^j = \frac{1}{n_1} \sum_{i=1}^{n_1} W_i^j \mathbf{x}_i' - \frac{1}{n_0} \sum_{i=1}^{n_0} (1 - W_i^j) \mathbf{x}_i' = \overline{\mathbf{X}}_T^j - \overline{\mathbf{X}}_C^j.$$

Morgan and Rubin (2012) proposed accepting the jth allocation when its treatment assignment vector $\mathbf{W}^j$ satisfies

$$M(\mathbf{W}^j, \mathbf{X}) \leq a,$$

where a is a positive constant.

By the central limit theorem, the sample means of the covariates will be normally distributed across random samples, so that $M(\mathbf{W}^j, \mathbf{X}) \sim \chi_K^2$. Letting

$$p_a = \Pr(\chi_K^2 \leq a) \simeq \Pr(M(\mathbf{W}^j, \mathbf{X}) \leq a), \tag{6}$$

we see that a is determined from the choice of p_a. Because the number of rerandomizations is geometrically distributed, the expected number of randomizations needed to obtain an acceptable allocation is $1/p_a$. This means for instance that for $p_a = 0.001$, the expected number of randomizations before drawing an allocation that fulfills the criterion is $1,000$.

Morgan and Rubin (2012) show that since $M(\mathbf{W}^j, \mathbf{X}) \sim \chi_K^2$,

$$Cov(\overline{\mathbf{X}}_T^j - \overline{\mathbf{X}}_C^j | \mathbf{X}, M(\mathbf{W}^j, \mathbf{X}) < a) = \nu_a Cov(\overline{\mathbf{X}}_T - \overline{\mathbf{X}}_C | \mathbf{X}), \tag{7}$$

with

$$\nu_a = \frac{\Pr(\chi_{(K+2)}^2 \leq a)}{\Pr(\chi_K^2 \leq a)}; \ 0 < \nu_a < 1. \tag{8}$$

This result implies that the variance in the covariate mean differences across allocations in $\mathcal{A}_a$ is reduced relative to its variance across the allocations in $\mathcal{A}$ by the factor ν_a, and

the percent reduction in variance of each of the covariates in $\mathbf{X}$ (or any linear combination of them) is equal to $100(1-\nu_a)$.

Let $Y_i(w)$ be the potential outcome under treatment w for individual i. Under the Stable Unit Treatment Value Assumption (SUTVA, Rubin 1980), the observed outcome when i is assigned W_i is equal to $Y_i = Y_i(W_i)$ The mean difference estimator is defined as

$$\widehat{\tau} = \overline{Y}_1 - \overline{Y}_0 \tag{9}$$

where $\overline{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n} W_i Y_i(1)$ and $\overline{Y}_0 = \frac{1}{n_0}\sum_{i=1}^{n}(1-W_i)Y_i(0)$.

Let $\widehat{\tau}^{CR}$ and $\widehat{\tau}^{RR}$ be the estimators defined in (9) under complete randomization and Mahalanobis-based rerandomization, respectively. These estimators are unbiased for the estimation of the sample average treatment effect (SATE) and also of the population average treatment effect (PATE) under random sampling of the n units from the population.

The variance of $\widehat{\tau}^{CR}$ is given by

$$V(\widehat{\tau}^{CR}) = \frac{S^2_{Y(1)}}{n_1} + \frac{S^2_{Y(0)}}{n_0} - \frac{S_{Y(1)Y(0)}}{n} \tag{10}$$

where

$$S^2_{Y(w)} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(w) - \overline{Y}(w))^2, \quad \overline{Y}(w) = \frac{1}{n}\sum_{i=1}^{n}(Y_i(w)$$

and

$$\begin{aligned} S_{Y(1)Y(0)} &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(1) - Y_i(0) - (\overline{Y}(1) - \overline{Y}(0))^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(\tau_i - \tau^s)^2 = S^2_\tau \end{aligned}$$

that is, the sample variance of the unit-level treatment effects. Note that with homogeneous treatment effects, i.e. $\tau_i = \tau$, $S_{Y(1)Y(0)} = 0$ and $S^2_{Y(1)} = S^2_{Y(0)} = S^2_Y$. This means that

$$V(\widehat{\tau}^{CR}) = \frac{n_0 n_1}{n} S^2_Y$$

Thus, all else equal, the variance of the estimator will be larger with heterogeneous effects than with homogeneous effects.

The asymptotic distribution is given by

$$\sqrt{n}(\widehat{\tau}^{CR} - \tau) \xrightarrow{d} N(0, V_{\tau\tau})$$

where

$$V_{\tau\tau} = \frac{S^2_{Y(1)}}{n_1} + \frac{S^2_{Y(0)}}{n_0} - \frac{S^2_\tau}{n}.$$

Under the superpopulation assumption and inference to the PATE, the third term vanishes since treated and controls are sampled independently.

Li et al (2018) derive the asymptotic results for Mahalanobis-based rerandomization. It is shown that the asymptotic distribution of the SATE and PATE (under random sampling) estimators after rerandomization is generally non-normal. Instead, the asymptotic distribution is a linear combination of a normal distributed variable and a truncated normal variable.

Let $\mathbf{Y}(w) = (Y_1(w), Y_2(w), ..., Y_n(w))', w = 0, 1$, and let R^2 be the squared multiple correlation of $\mathbf{Y}(0)$ on $\mathbf{X}$. Under the assumptions that (i) the residual in the linear projection of $\mathbf{Y}(0)$ on $\mathbf{X}$ is normally distributed and that (ii) treatment effects are additive (so that R^2 is also the squared multiple correlation of $\mathbf{Y}(1)$ on $\mathbf{X}$)), it holds that the percentage reduction in variance (PRIV) of $\widehat{\tau}^{RR}$ against the corresponding estimators under complete randomization is

$$PRIV = \frac{V(\widehat{\tau}^{CR}) - V(\widehat{\tau}^{RR})}{V(\widehat{\tau}^{CR})} = 100 \times R^2(1 - \nu_a), \tag{11}$$

where $V(.)$ denotes the variance of the estimators. From this expression together with Equations 6 and 8, it becomes clear that the variance reduction from Mahalanobis-based rerandomization relative to complete randomization is decreasing in p_a, the strictness of the rerandomization criterion, and non-increasing in K, the dimension of $\mathbf{X}$.

Appendix B: Supplementary results

Table 6: Treatment effect estimates from the specification with interaction effects based on a log-linear model: full estimation sample

	Women		*Men*	
	Compliant	*Non − compliant*	*Compliant*	*Non − compliant*
	capinc			
Treatment effect	-0.148	1.491	-0.48	1.375
CI	[-0.652, 0.356]	[0.514, 2.468]	[-1.12, 0.16]	[0.414, 2.336]
	tax			
Treatment effect	-0.115	-0.142	0.079	-0.047
CI	[-0.361, 0.131]	[-0.405, 0.12]	[-0.038, 0.197]	[-0.269, 0.175]

Note: *p<0.1; **p<0.05; ***p<0.01. The effects were estimated using transformed PML-estimates from a log-linear model based on the covariate set from equation (5). For instance, the estimate for non-compliant men is given by $\widehat{\tau}$, the estimate for non-compliant women is given by $\widehat{\tau} + \widehat{\tau}_w$, etc. The effect estimates can be approximately interpreted as percentage effects. The full sample has been used in the estimation (998 observations). 95% confidence intervals formed using the HC2 covariance matrix are shown in square brackets.

Table 7: Treatment effect estimates from the specification with interaction effects based on OLS-regression

	Women		*Men*	
	Compliant	*Non − compliant*	*Compliant*	*Non − compliant*
	capinc			
Treatment effect	-2.414	10.415	-5.639	18
CI	[-8.683, 3.855]	[0.133, 20.697]	[-12.979, 1.701]	[-9.689, 45.69]
	tax			
Treatment effect	-19.432	-42.774	25.546	-15.216
CI	[-97.518, 58.653]	[-113.347, 27.798]	[-9.833, 60.925]	[-65.338, 34.905]

Note: *p<0.1; **p<0.05; ***p<0.01. The effects were estimated using transformed OLS-estimates from the specification in equation (5). For instance, the estimate for non-compliant men is given by $\widehat{\tau}$, the estimate for non-compliant women is given by $\widehat{\tau} + \widehat{\tau}_w$, etc. Outcomes are expressed in 1000s SEK and are valued during the income year of 2019. Observations for the largest value in *capinc* and *tax*, respectively, have been removed from the estimation sample (997 out of 998 observations left). 95% confidence intervals formed using the HC2 covariance matrix are shown in square brackets.