Pre-Analysis Plan

The Impact of a Digital Message on Tax Compliance: Evidence from a Swedish Field Experiment Using Model-based Bayesian Inference

Nikolay Angelov* Per Johansson†

May 9, 2022

*The Swedish Tax Agency and Uppsala Center for Fiscal Studies (UCFS), nikolay.angelov@skatteverket.se
†Department of Statistics at Uppsala University, IFAU, and IZA, per.johansson@statistics.uu.se

1 Introduction

The Swedish Tax Agency receives information about parts of Swedish taxpayers' foreign income through OECD's Common Reporting Standard (CRS). The goal of the CRS is to combat tax evasion and it was approved by the OECD council in 2014. Within the CRS, tax authorities obtain information from financial institutions in their own jurisdiction and automatically exchange that information with other jurisdictions on an annual basis.

The purpose of the present study is to estimate the effect of a low-cost intervention (either one of two versions of a digitally disseminated message) on the subsequent tax compliance of Swedish taxpayers who, according to CRS-data, had foreign dividends or interest from any of the three neighbouring countries Norway, Finland, or Denmark, during the previous tax year.

2 The information messages

The intervention consists of randomly sending the two following messages randomly among the treated individuals and leaving the controls untreated:

> **Message A**
>
> *Declaring foreign income*
>
> Hello,
>
> The Swedish Tax Agency has obtained information from a foreign tax authority that you have received dividends or interest from abroad during 2020. If you have received foreign income also during 2021, there is an online app on skatteverket.se which can help you with the correct amount to file and whether you can claim a foreign tax offset.
>
> You can find the app here. (link to https://app.skatteverket.se/klient-sifu-segmentering/)
>
> Sincerely,
>
> The Swedish Tax Agency

Message B

Declaring foreign income

Hello,

The Swedish Tax Agency has obtained information from a foreign tax authority that you have received dividends or interest from abroad during 2020. If you have received foreign income also during 2021, we will receive information about it later this year. Foreign income is therefore not pre-filled in the declaration.

If you have had foreign income also during 2021, you have to declare it yourself.

On skatteverket.se, there is an online app which can help you with the correct amount to file and whether you can claim a foreign tax offset.

You can find the app here. (link to https://app.skatteverket.se/klient-sifu-segmentering/)

Sincerely,

The Swedish Tax Agency

These messages were sent in digital form to the treated taxpayers via a digital mailbox. In Sweden, about 64 percent of individuals aged 15 and above have a digital mailbox.[1] 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 in the messages above does not require login and can be best described as a calculator. The user fills in the type of foreign income (dividends or interest), amount, currency, date of receiving the amount, country, 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.

Our main outcome variable is capital income (*capinc*), where the amount of foreign income should be included. In addition, we plan to estimate the effect on total tax paid (*tax*). An increase in declared foreign income in *capinc* should lead to an increase in total tax paid unless no offsetting adjustments are made in the income tax declaration. Data on

[1] Source: https://www.digg.se/digital-post/offentlig-aktor/statistik-och-prognoser and official population data from Statistics Sweden.

the outcome variables for the tax year 2021 will be collected in August or September 2022. The latest day for sending in the income tax declaration is May 2, meaning that data for many taxpayers will be available after this date. However, in Sweden it is easy for individual taxpayers to apply for a delayed declaration without any penalty, and some taxpayers are late with their declaration for other reasons. Our experience from a similar field study performed two years previously is that declaration data will be missing for a non-negligible share of taxpayers before the summer holidays.[2]

3 Study population and sample

The original population consists of 99,475 individual taxpayers each of whom had foreign dividends and interest from Norway, Finland, or Denmark during tax year 2020, according to information from the CRS. The number of treated was set to 5,000 individuals. The original population was reduced according to the following steps:

1. Taxpayers who had some pre-filled foreign income in the declaration were excluded (70,441 left)
2. Those who had not filed an income tax declaration for 2020 were excluded (69,630 left)
3. Only taxpayers with a digital mailbox as of March 10, 2022 were kept (45,613 left)
4. Remove taxpayers with foreign dividends and interest less than 100 SEK (11,643 left)
5. Trim the sample by removing the top 1% with respect to five variables (11,238 left). The top percentiles for each variable were calculated prior to any trimming. The variables, measured during 2020, were:
   - Foreign dividends and interest (*fdiv*)
   - Earnings including labor income, sick pay, pension, etc. (*earn*)
   - Capital income (*capinc*)
   - Foreign dividends and interest as a share of capital income ($fdivrat = \frac{fdiv}{capinc+1}$)
   - Total tax paid (*tax*)
6. Taxpayers aged below 20 and above 89 were removed (11,095 left)
7. Three taxpayers who had a negative final tax in 2020 were removed (11,092 left)

[2]The pre-analysis plan for the previous study is documented in Angelov and Johansson (2020)

Table 1: Group averages by treatment status

Treatment	*woman*	*compliant*	*age*	*fdiv*	*capinc*	*tax*	*earn*	*finc*	*#obs*
0	0.47	0.22	48.67	3.92	8.15	119.53	347.49	0.15	$5,000$
A	0.46	0.22	49.17	3.80	8.31	123.01	346.36	0.15	$2,500$
B	0.47	0.23	48.81	3.87	8.91	120.93	358.65	0.14	$2,500$

Note: The measure units of the variables are as follows: *age* is measured in years, *woman*, $finc$, and *compliant* are valued 0 or 1, *#obs* denotes number of observations, and the rest of the variables are expressed in 1,000s SEK which is approximately equal to 100 EUR.

4 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 10,000 individuals.

Using the 11,092 individuals left in the sample after the procedure described in the previous section, we proceed as follows:

1. Draw a simple random sample of size 10,000 from the 11,092 taxpayers. These 10,000 individuals constitute the sampling frame of the trial.

2. Randomize half of the taxpayers in the sampling frame to treament. 2,500 taxpayers are randomized into treatment A (i.e., receiving message A) and 2,500 are randomized into treatment B.

Table 1 shows group means by treatment status and a number of variables. The variables are pre-treatment (measured in 2020) and most of them were defined in the previous section. One additional variable is a measure of previous compliance ($\mathbf{1}[compliant]$ where $\mathbf{1}[\cdot]$ is the indicator function, or *compliant* for short). We do not have access to a perfect measure of previous compliance. As a proxy, we use information about whether the amount of foreign dividends and interest obtained from the CRS was less than or equal to total capital income, i.e., $\mathbf{1}[compliant] \equiv \mathbf{1}[fdiv \leq capinc]$. 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. $\mathbf{1}[compliant]$, where $\mathbf{1}[\cdot]$ is the indicator function. Below, we list the variables used in Table 1:

- *age*: the taxpayer's age in years
- *woman*: $\mathbf{1}$[taxpayer is a woman]
- *fdiv*: foreign dividends and interest
- *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 income declaration
- *compliant*: $\mathbf{1}[compliant] \equiv \mathbf{1}[fdiv \leq capinc]$, a measure of tax compliance

5 Hypotheses and inference

Our main hypothesis is that being treated (i.e., receiving message A or B) leads to an increase in tax compliance with respect to declared foreign income, measured by an increase in declared *capinc* in the income tax declaration for 2021. Furthermore, we expect the effect to be stronger from the slightly longer and more explicit message B compared to message A. The main effect driver is assumed to be an increase in *perceived detection probability* among taxpayers who *underdeclare their foreign income*. Thus, to the extent that our previous compliance measure is valid, the effect should be larger among taxpayers for whom $\mathbf{1}[compliant] = 0$. Finally, as a large body of literature has documented that women are on average more risk averse than men, we expect a larger effect size among women.

These hypotheses are easily tested under the assumption of random sampling. This could be done by specifying a null of no differences in the average treatment effect in the population. Given a factorial design, this setup would enable us to reject these hypothesis at a pre-specified level of risk (i.e. the significance level). Here, instead we plan to use a model-based Bayesian inference which allows us to calculate the probability that a hypothesis can be rejected under the given model, or to create credible intervals for any given estimand of interest.

We regard the potential outcomes as random variables and specify a super-population model for these potential outcomes that depends on unknown parameters. Since the potential outcomes are considered random variables, any functions of them will also be random variables. This is true for any causal estimand of interest, including for instance the average treatment effect, the median causal effect, quantile effects, etc.

For any individual, let $Y(0)$ be the potential outcome under no message, $Y(1)$ be the potential outcome under message A and $Y(2)$ the potential outcome under message B. Define the causal estimand of interest

$$\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{Y}(2), \mathbf{X}, \mathbf{W}), \tag{1}$$

allowing the estimand to depend on the pre-treatment variables ($\mathbf{X}$), the vector of treatment indicators ($\mathbf{W}$), and the vector of potential outcomes. The potential outcomes, and thus the causal estimands, are well-defined irrespective of the stochastic model for either the treatment assignment or for the potential outcomes.

The Bayesian approach to model-based inference for causal effects starts with the specification of the super-population model for the potential outcomes, $f(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{Y}(2)|\boldsymbol{X}, \boldsymbol{\theta}, \mathbf{W}) = f(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{Y}(2)|\boldsymbol{X}, \boldsymbol{\theta})$, where the equality sign stems from the randomization. We plan to use simulation-based inference which means that we repeatedly draw values from the posterior distribution and use the postulated model to impute the missing potential outcomes, given the observed data, and use this to create credible intervals for the estimand of interest.

There are four steps involved in the approach. In order to describe these steps, define the potential outcomes in terms of of the missing and observed values:

$$Y_i(0) = \begin{cases} Y_i^{\text{obs}} & \text{if } W_i = 0, \\ Y_i^{\text{mis}} & \text{if } W_i = 1, \\ Y_i^{\text{mis}} & \text{if } W_i = 2, \end{cases} \quad Y_i(1) = \begin{cases} Y_i^{\text{mis}} & \text{if } W_i = 0, \\ Y_i^{\text{obs}} & \text{if } W_i = 1. \\ Y_i^{\text{mis}} & \text{if } W_i = 0. \end{cases}$$

and

$$Y_i(2) = \begin{cases} Y_i^{\text{mis}} & \text{if } W_i = 0, \\ Y_i^{\text{mis}} & \text{if } W_i = 1 \\ Y_i^{\text{obs}} & \text{if } W_i = 2. \end{cases}$$

Here, $W_i = 1$ if taxpayer i is A-treated, $W_i = 2$ if B-treated and $W_i = 0$ if i is untreated.

The first step of the model-based approach involves deriving $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{X}, \mathbf{W}, \theta)$, where $\mathbf{Y}^{\text{obs}}$ and $\mathbf{Y}^{\text{mis}}$ are the vector of observed and missing observations, respectively. The second step involves deriving the posterior distribution for the parameter $\boldsymbol{\theta}$, that is, $f(\boldsymbol{\theta}|\mathbf{Y}^{\text{obs}}, \mathbf{X}, \mathbf{W})$. The third step involves combining the conditional distribution $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{X}, \mathbf{W}, \theta)$ and the posterior distribution $f(\theta|\mathbf{Y}^{\text{obs}}, \mathbf{X}, \mathbf{W})$ to obtain the conditional distribution of the missing data given the observed data. Finally, in the fourth step we use the conditional distribution $f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{X}, \mathbf{W})$ together with the observed outcomes to obtain the conditional distribution of the estimand, $f(\tau|\mathbf{Y}^{\text{obs}}, \mathbf{X}, \mathbf{W})$.

Below, we outline the model that follows a framework from Imbens and Rubin (2015, Ch. 8). We will assume a model allowing for a sizable proportion of zeros in the potential outcomes, to reflect the chosen outcome variables *capinc* and *tax*. Two parts of the conditional distribution are modeled. First, the probability of a positive value for $Y_i(0)$ is

$$\Pr\left(Y_i(0) > 0|\boldsymbol{x}_i, W_i, \boldsymbol{\theta})\right) = \frac{\exp(\boldsymbol{x}_i\gamma_0)}{1+\exp(\boldsymbol{x}_i\gamma_0)}, \tag{2}$$

and similarly for $Y_i(1)$ and $Y_i(2)$. Thus,

$$\Pr\left(Y_i(1) > 0|\boldsymbol{x}_i, W_i, \boldsymbol{\theta}\right) = \frac{\exp(\boldsymbol{x}_i\gamma_1)}{1+\exp(\boldsymbol{x}_i\gamma_1)}$$

and

$$\Pr\left(\mathbf{Y}(2) > 0|\boldsymbol{x}_i, W_i, \boldsymbol{\theta}\right) = \frac{\exp(\boldsymbol{x}_i\gamma_2)}{1+\exp(\boldsymbol{x}_i\gamma_2)}.$$

The vector of pre-treatment covariates $\boldsymbol{x}_i$ consists of the variables used in the description in Table 1 (*woman*; *compliant*; *age*; *fdiv*; *capinc*; *tax*; *earn*; *finc*).

Second, conditional on a positive outcome, the logarithm of the potential outcome is assumed to follow a normal distribution:

$$\ln\left(Y_i(0)\right)|Y_i(0) > 0, \boldsymbol{x}_i, W_i, \boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{x}_i\boldsymbol{\beta}_0, \sigma_0^2\right), \tag{3}$$

$$\ln\left(Y_i(1)\right)|Y_i(1) > 0, \boldsymbol{x}_i, W_i, \boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{x}_i\boldsymbol{\beta}_1, \sigma_1^2\right),$$

and

$$\ln\left(Y_i(2)\right)|Y_i(2) > 0, \boldsymbol{x}_i, W_i, \boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{x}_i\boldsymbol{\beta}_2, \sigma_2^2\right).$$

For the parameters $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_1$, we assume the prior to be independent from each other as well as from the other parameters. The prior distributions will be specified to be normal with zero means and variance equal to 100^2. The prior distributions for σ_0^2 , σ_1^2 and σ_2^2 will be inverse gamma with parameters 1 and 0.01 respectively.

The Bayesian method relies on model assumptions. Depending on the characteristics of the data, more complicated Bayesian models can be specified, e.g. a nonparametric Bayesian regression like the Bayesian Additive Regression Tree (Chipman et al., 2010) . However, as the data stems from a randomized experiment designed to balance the means of the covariates between the treated and control groups, we expect that the Bayesian method is in general fairly robust to model misspecification. The sensitivity to the model assumptions on covariates and parameter priors will nevertheless be checked.

To be more explicit on the simulation approach, let N be the number of observations (10,000 in our case) and focus on the average treatment effect of treatment A against no message as the estimand. The average treatment effect for treatment A for this population of size N can be written as

$$\tau_{\text{fs}} = \tau(\mathbf{Y}(0), \mathbf{Y}(1)) = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(1) - Y_i(0)\Big). \tag{4}$$

When the potential outcomes are linearly related to the covariates, one example of the model for $f(\mathbf{Y}(1), \mathbf{Y}(0)|\boldsymbol{X}, \boldsymbol{\theta})$ is

$$\begin{aligned} Y_i(0) &= \alpha_0 + \boldsymbol{\beta}_0'\boldsymbol{x}_i + \varepsilon_{i0}, \\ Y_i(1) &= \alpha_0 + \gamma + \boldsymbol{\beta}_0'\boldsymbol{x}_i + \varepsilon_{i1}, \end{aligned} \tag{5}$$

where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2), j = 0, 1$, are independent across units.

Given a prior for $\boldsymbol{\theta} = (\alpha_0, \gamma, \boldsymbol{\beta}_0, \sigma_0^2, \sigma_1^2)'$ we obtain posterior samples of the parameters. For each posterior sample, we impute the missing potential outcomes under the unassigned treatment arms. Because there is no information in the data about the correlation between $Y_i(0)$ and $Y_i(1)$ for the same unit we take a conservative approach by assuming that the two potential outcomes are perfectly correlated (Imbens and Rubin, 2015). The posterior samples of τ are then calculated as follows.

Suppose that the hth posterior sample of the parameters is

$$\left(\alpha_0^{(h)}, \gamma^{(h)}, \boldsymbol{\beta}_0^{(h)}, \sigma_0^{2(h)}, \sigma_1^{2(h)}\right).$$

For units with $W_i = 1$, we let $[Y_i(1)]^{(h)} = Y_i^{\text{obs}}$, and impute $Y_i(0)$ as

$$[Y_i(0)]^{(h)} = \alpha_0^{(h)} + [\boldsymbol{\beta}_0^{(h)}]'\boldsymbol{x}_i + \frac{\sigma_0^{(h)}}{\sigma_1^{(h)}}\left(Y_i^{\text{obs}} - \alpha_0^{(h)} - \gamma^{(h)} - [\boldsymbol{\beta}_0^{(h)}]'\boldsymbol{x}_i\right).$$

For units with $W_i = 0$, we let $[Y_i(0)]^{(h)} = Y_i^{\text{obs}}$, and impute $Y_i(1)$ as

$$[Y_i(1)]^{(h)} = \alpha_0^{(h)} + \gamma^{(h)} + [\boldsymbol{\beta}_0^{(h)}]'\boldsymbol{x}_i + \frac{\sigma_1^{(h)}}{\sigma_0^{(h)}}\left(Y_i^{\text{obs}} - \alpha_0^{(h)} - [\boldsymbol{\beta}_0^{(h)}]'\boldsymbol{x}_i\right).$$

The hth posterior sample of τ is

$$\tau^{(h)} = \frac{1}{N} \sum_{i=1}^{N} \left\{ [Y_i(1)]^{(h)} - [Y_i(0)]^{(h)} \right\}.$$

With a total number of H posterior samples, the posterior mean $\bar{\tau} = \sum_{h=1}^{H} \tau^{(h)}/H$ is used to estimate τ. A 95% credible interval for τ is constructed using the .025 and .975 quantiles of the posterior samples of τ.

References

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.

Chipman, H. A., George, E. I., and McCulloch, R.E. (2010) BART: Bayesian additive regression trees. The Annals of Applied Statistics, 4(1): 266-298.

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

Imbens G, Rubin D. (2015) Causal Inference for Statistics, Social, and Biomedical Sciences: an Introduction. Cambridge University Press.

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

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