

Nudging for Tax Compliance: A Meta-Analysis

Armenak Antinyan^{*†} Zareh Asatryan^{*‡}

First version: November, 2019

This version: July, 2024

Abstract

Governments increasingly use nudges to improve tax collection. We synthesise the growing literature on nudging experiments using meta-analytical methods. We find that, relative to the baseline where about a quarter of taxpayers are compliant, simple reminders increase the probability of compliance by 2.7 percentage points, while tax morale and deterrence nudges increase compliance by additional 1.4 and 3.2 percentage points. Our further results highlight the conditions where nudges are more or less effective. Overall, our findings imply that taxpayers are biased by various informational and behavioural constraints, and that nudges can be of some help in overcoming these frictions.

JEL codes: C93, D91, H26.

Keywords: Tax compliance, Tax evasion, Randomised trials, Nudging, Reminders, Tax morale, Deterrence, Meta-analysis, Publication bias.

^{*}We would like to thank Keith Marzilli Ericson, Annika Havlik, Jost Heckemeyer, Friedrich Heine-mann, Christos Kotsogiannis, Carla Krolage, Tom Lane, Carina Neisser, Justus Nover, Anh Pham, Johannes Rincke, Christian Traxler, as well as Steffen Huck (the editor) and four referees for their valuable comments. We are grateful to Felix Köhler as well as Kerry Neitzel, Agon Topxhiu, David Westerheide and Zeyuan Xiong for excellent research assistance. Data and programs to replicate the results of the paper are available on the journal website.

[†]Thames Water Utilities.

[‡]ZEW Mannheim and CESifo Munich; zareh.asatryan@zew.de; L7 1, Mannheim, Germany, 68161.

1 Introduction

Recent years have seen much excitement around the idea of using “nudges” with the aim of improving individual behaviour. Nudges are interventions that respect freedom of choice and leave economic incentives intact ([Benartzi et al., 2017](#)), and they have been studied in many policy areas such as education ([Dizon-Ross, 2019](#)), healthcare ([Wisdom et al., 2010](#)), environment ([Costa and Kahn, 2013](#)), finance ([Handel, 2013](#)), savings decisions ([Blumenstock et al., 2018](#), [Karlan et al., 2016](#)), and welfare benefits ([Finkelstein and Notowidigdo, 2019](#), [Linos et al., 2022](#)), among others.

In the field of taxation too, nudging has become quite popular in the last decade. This holds both among academics, who strive to understand why people pay taxes, and among policy makers who often claim that the potential payoffs of nudges can be very large in terms of raised revenue. In tax experiments, nudges occasionally take the form of reminders similar to other contexts, and more often they are designed to appeal to either moral motives behind paying taxes or to deterrence reasons behind paying taxes such as threats of audits. In light of the growing number of studies in this field, our paper aims to present a quantitative review of the literature and to provide guidance for further (policy) interventions.

In particular, our meta-analysis attempts to give more systematic answers to questions such as: i) Are nudges effective in curbing tax evasion? ii) If so, by how much on average? iii) Which nudge types work more strongly? iv) Are nudges also effective over a longer time horizon? v) Which groups of taxpayers are more responsive to nudges? vi) Do nudges work only in specific settings (e.g., low-compliance environments) or more generally?

To answer these questions, we collect data on intention to treat (ITT) estimates of nudging interventions on tax compliance from studies that implement randomised control trials (RCTs). Our largest sample consists of up to 71 RCTs, while our baseline sample – which focuses on measurements of extensive margin tax compliance and what we think of main estimates presented by papers – consists of 53 papers and 270 estimates. Our analysis starts with a synthesis of this literature. This appraisal provides a taxonomy of nudging interventions in the field of tax compliance, in particular highlighting the main experimental designs used, the common types of nudges studied, the customary measures of tax compliance used and the important contextual characteristics that define nudging interventions. We then apply meta-analytical techniques to identify the quantitative impact of various types of nudges on tax compliance.

Our main results are threefold. First, our evidence suggests that reminders increase extensive margin compliance, i.e., the share of compliant taxpayers, by 2.7 percentage points compared to a control group of taxpayers not receiving any treatments. Second, we find that non-deterrence nudges, i.e., interventions commonly referring to elements of tax morale, increase extensive margin compliance by another 1.4 percentage points in addition to the reminder effect. Third, we show that deterrence nudges, that is interventions that inform taxpayers about potential audit probabilities and fine rates when caught cheating, increase compliance by an additional 3.2 percentage points on top of reminders.

To put these effects into perspective, we compare them to underlying levels of compliance, that is to the share of compliant taxpayers in the control group that received no communication. In the sample where this information is reported, only about 25% of taxpayers not receiving any nudges are compliant on average, which is low but not surprisingly so, since more than half of estimates in our sample work

with samples of taxpayers which were late in paying their taxes. Our estimates suggest that, compared to this underlying level of compliance, the reminder effect increases the probability of compliance by 10.8% on average, tax morale and other non-deterrence nudges raise compliance by 16.4%, and deterrence nudges are most effective, increasing tax compliance by 23.6%. Thus, in an average experiment, the most comprehensive of nudges, those sending reminders in combination with warning about deterrence, are able to increase the share of compliant taxpayers from 25% in the control group with no communication to about 31%. These headline numbers are based on 44 papers and 218 estimates which is a sub-sample of our baseline sample but restricting it to papers which report the underlying levels of compliance in the control group. The results are generally robust to these and other sample definitions as well as to alternative estimators and measures of tax compliance.

These results are consistent with the idea that taxpayers are biased by various informational and behavioural constraints, and that nudges can help overcome these frictions. Whereas reminders help overcome limited attention, tax morale and deterrence nudges operate by updating taxpayers' beliefs or preferences on the moral and deterrence motives behind paying taxes. As far as the stronger compliance effect of deterrence nudges relative to non-deterrence nudges is concerned, one interpretation is that individual financial motives are more important for compliance decisions than elements of tax morale. However, it is also plausible that nudges implemented by tax authorities are simply more effective at updating perceptions of audit probabilities than perceptions of the various tax morale elements. In terms of the types of tax morale nudges, we consider three main groups – nudges which highlight the importance of paying taxes for the adequate provision of public goods, those about the (positive) behaviour of the majority of taxpayers, and a third group hinting at general appeals of

paying taxes as a moral obligation – and show that neither of them stands out to be as important driver of compliance as deterrence nudges.

Although we are tempted to make comparisons between our results and that of nudges in contexts going beyond tax compliance, such comparisons are not straightforward given the heterogeneity in the behavioural outcomes that nudges are used to target. [Benartzi et al. \(2017\)](#), [Hummel and Maedche \(2019\)](#), [Mertens et al. \(2022\)](#) provide meta-analyses of nudging interventions in various fields, albeit almost always neglecting the tax compliance studies. As suggested by these papers, reminders are one of the more popular nudges in the literature. Deterrence nudges, on the other hand, are used in more special cases such as in law enforcement related contexts. Finally, among the morale nudges, social norm nudges are the ones that are somewhat popular in other fields (for a review of the role of social norms in shaping attitudes and behaviours, see, [Bursztyn and Yang, 2022](#)). However, large heterogeneities in both the treatments and outcomes studied in these papers do not allow for meaningful comparisons of effect magnitudes across the different types of nudges. In addition, a better assessment of whether the effects of nudges that we have identified are small or large requires an understanding of the welfare impacts of nudges as well as an idea of how these effects relate to those of traditional policy tools. Despite the popular belief that the sending of nudges is essentially costless, several papers – such as, [Damgaard and Gravert \(2018\)](#) on reminders for charitable giving, [Huck and Rasul \(2010\)](#) on transaction costs again in the context of giving, [Allcott and Kessler \(2019\)](#) on social norms in energy savings, [Bernheim et al. \(2015\)](#) on default options for savings decisions, [Bhattacharya et al. \(2015\)](#) on commitment devices in health choices, among others – show that nudges may entail significant costs, and suggest that the failure to take these costs into account will overstate the welfare effects of nudges. In addition, [List et al. \(2023\)](#)

studies the welfare effects of policies that combine nudges with more traditional price instruments.

Our additional findings highlight certain design aspects of RCTs that may make nudging more or less effective for tax compliance. We find that nudges are more effective in the very short-run, when targeting late-payers, when communicated through in-person visits, and when implemented in higher income countries. Our final set of results focuses on publication bias. Consistent with [DellaVigna and Linos \(2022\)](#)¹ and [Brodeur et al. \(2020\)](#), we find evidence that the results of this literature, despite being identified through RCTs, are likely to be driven by selection effects both on the basis of statistical significance and also based on the sign of reported treatment effects.

The remainder of the paper is structured as follows. Section 2 presents a taxonomy of tax compliance nudges based on our synthesis of the literature. Section 3 describes the sample of papers and estimates that we collect. Section 4 discusses the meta-analysis framework that we use for estimation. Section 5 presents our main results, and Section 6 discusses our additional results on publication bias. Section 7 concludes with a summary and suggestions for future research.

¹[DellaVigna and Linos \(2022\)](#) compare the impact of nudges in RCTs conducted by nudge units with those found in RCTs conducted and published by academics. The authors find the average impact of nudges to be 1.4 percentage points or 8% in the nudge unit trials, which is one-sixth the magnitude found in academic trials, and explain a large part of this difference to be driven by publication bias.

2 A taxonomy of tax compliance nudges

2.1 Definition of nudges

Unlike standard economic policy interventions, nudging interventions neither prohibit individuals from undertaking a certain action, nor they affect the economic incentives of these individuals (Sunstein, 2014, Thaler and Sunstein, 2008). Thaler and Sunstein (2008) define a nudge as an “aspect of the choice architecture that alters people’s behaviour in a predictable way without forbidding any options or significantly changing their economic incentives”. They continue that for an intervention “to count as a mere nudge, the intervention must be easy and cheap to avoid”. Sunstein (2014) provides a list of popular nudges in various fields. In the context of taxation, experiments are usually conducted in collaboration with the tax authorities of a given jurisdiction. Not only academics, but also dedicated nudge units (e.g., Mind, Behaviour, and Development Unit of the World Bank, the Behavioural Insights Team) conduct such experiments.

In a typical nudging experiment, the agents are randomised into one or several treatment arms, which are exposed to a nudge or nudges of various types, and a control arm in which the agents are either not exposed to any intervention or are exposed to a neutral intervention. The behaviour of agents in the treatment group or groups is then compared with that of agents in the control group some time after the intervention. The exact measurement of taxpayer behaviour, the length of the time horizon over which this behaviour is studied as well as the delivery method of the nudge can vary across experiments. While some experiments may study one type of behaviour (e.g., probability to pay) over one time horizon after sending the nudge using a certain delivery method (e.g., digital letters), others may study multiple types of behaviours (e.g., probability to pay and probability to file) measured over multiple

time horizons (e.g., one month and three months) of nudges delivered through multiple methods (e.g., digital letters and in-person visits), or some combination of these. On the other hand, the taxpayer type (e.g., individual or business), the specific tax (e.g., income tax, property tax or indirect tax) and the country are typically fixed in a given experiment.

2.2 Literature on nudging for tax compliance

The literature on tax compliance is centred around the questions of why taxpayers pay (or do not pay) taxes, and on the effectiveness of enforcement policies in enhancing tax compliance. These questions are of central importance in public economics as the level and nature of tax compliance may have implications for the efficiency and distributive effects of taxes (see, e.g., [Slemrod and Gillitzer, 2014](#)), and what they can say about the level of public good provision. Several excellent qualitative reviews have been written on this extensive literature. Reviews by [Andreoni et al. \(1998\)](#), [Slemrod and Yitzhaki \(2002\)](#) and [Slemrod \(2007\)](#) and, more recently, by [Slemrod \(2019\)](#) and [Alm \(2019\)](#) discuss the literature on the economics of tax compliance. More specifically, [Luttmer and Singhal \(2014\)](#), [Mascagni \(2018\)](#) and [Pomeranz and Vila-Belda \(2019\)](#) review the literatures on the roles of, respectively, tax morale, tax experiments and tax capacity in tax compliance.

We perform a systematic quantitative analysis of the literature on the impact of nudging interventions on tax compliance for the first time.² Unlike the recent qualitative reviews, we not only study the question of whether the nudging interventions

²We are aware of two other meta-studies of tax experiments by [Blackwell \(2007\)](#) and [Alm and Malézieux \(2020\)](#), but both study laboratory experiments while we focus on field work. [Blackwell \(2007\)](#) concludes that increasing the penalty rate, the marginal per capita return to the public good and the probability of audit lead to higher tax compliance, while the tax rate has no significant impact on tax compliance. Focusing on a larger set of papers, [Alm and Malézieux \(2020\)](#) illustrate that audit

are effective or not, but we also provide an estimate of the average effects of nudges on tax compliance. This is important, since, despite the relative similarity of these nudging experiments, summarise their results as having “produced varying results in different contexts”. We also aim to understand which nudge types work best in increasing compliance, and what are the important contextual characteristics that potentially matter for the effectiveness of these nudges.

The three most distinctive features of nudges in the existing literature are their potential to boost tax compliance, first, by referring to deterrence factors such as the threat of audit and fines, second, by using reminders, and third, by appealing to morale elements such as altruism and fairness.³ [Luttmer and Singhal \(2014\)](#) Below we discuss these three groups of nudges one by one.

2.3 Types of nudges

Deterrence nudges: Tax compliance may be driven by a cost-benefit calculation reflecting on the trade-off between higher retained income due to evasion and costs potentially incurred if caught evading. This is the essence of the so-called deterrence approach to tax compliance. The workhorse model dates back to [Allingham and Sandmo \(1972\)](#) and, following the economics of crime literature ([Becker, 1968](#)), articulates that taxpayers are rational utility maximisers who compare the benefits of tax evasion against the costs of detection and punishment when deciding to comply

probability increases tax compliance on the extensive margin, while audit probability and the tax rate influence tax compliance negatively on the intensive margin.

³The nudges we study are arguably the most common types of behavioural interventions in taxation, but governments can nudge in other ways too. For example, policies that publicly recognise the top taxpayers and shame the tax delinquents, as studied by [Slemrod et al. \(2022\)](#) and [Dwenger and Treber \(2018\)](#), or ones that use third-party information reports to pre-fill tax returns, as studied by [Fochmann et al. \(2018\)](#), [Gillitzer and Skov \(2018\)](#), [Kotakorpi and Laamanen \(2016\)](#), might as well be considered as nudges in the broader sense of the word.

with taxes ([Alm, 2012](#)).⁴ This puts forth the fine rate and the audit probability as the two most important policy instruments for enforcing tax compliance ([Alm, 2019](#)). The idea behind deterrence nudges is then to refer to these deterrence factors with the aim of increasing tax compliance, of course without changing the audit probability or the fine rate. A large body of previous evidence, both from the field and the lab, has confirmed that audit and penalty rates do matter for compliance decisions (see, e.g., [Slemrod, 2019](#)). Thus, deterrence nudges can affect compliance by making the audit and penalty rates more salient to taxpayers, or by updating the magnitudes of already salient beliefs.

Consequently, to be considered as a deterrence nudge, the communication between tax administration and taxpayers should contain elements of enforcement. More specifically, the communication should include a threat that highlights the possibility of an audit or the potential penalty if caught evading (or both). A typical example of a deterrence nudge is the following one used by [Castro and Scartascini \(2015\)](#): “Did you know that if you do not pay the CVP on time for a debt of AR\$ 1,000 you will have to disburse AR\$ 268 in arrears at the end of the year and the Municipality can take administrative and legal action?”⁵

Reminder nudges: Tax compliance behaviour may depend on the simple behavioural fallacy of limited attention. Limited attention may lead individuals to forget about the tax payment deadline and simple reminders can help them overcome this issue. [Antinyan et al. \(2021\)](#), [Hernandez et al. \(2017\)](#), [Mascagni et al. \(2017\)](#) discuss the

⁴Recent extensions of this theory include, among others, the possibility that agents are sometimes unable to cheat because of withholding and third-party reporting rules ([Kleven et al., 2011, 2016](#)), or that agents face substantial uncertainties with regards to the (perceived) probabilities of being caught ([Snow and Warren, 2005](#)).

⁵Note that communications including both deterrence and non-deterrence components are classified as a deterrence nudge, given the presence of the threat component.

role of reminders in the tax compliance context. Reminder nudges have also, of course, been studied in other policy areas, such as health ([Altmann and Traxler, 2014](#)), savings ([Calzolari and Nardotto, 2017](#)) and investment ([Karlan et al., 2016](#)) decisions.

The nudges in this sub-category are mainly utilised to “correct” taxpayer non-compliance that stems from limited attention. These nudges use neutral language to remind the taxpayers to comply with taxes, for example: “RRA would like to inform you that your CIT tax return is due by 31st of March 2016. For more information about the filing process and payment methods, contact the call centre (3004) or visit the RRA website (<http://www.rra.gov.rw>)” ([Mascagni et al., 2017](#)).

Tax morale nudges: A number of moral factors such as intrinsic motivation, social norms, altruism, reciprocity, and fairness, among others, may affect the tax compliance decision. “Tax morale” is an umbrella term that encompasses these factors. Individuals may be intrinsically motivated to pay taxes without any enforcement ([Torgler, 2003](#)), and feel shame or guilt in cases of tax evasion ([Coricelli et al., 2010](#), [Dulleck et al., 2016](#)). Individuals may also be guided by concerns of reciprocity and comply with taxes if the state effectively provides public goods or treats the taxpayers fairly ([Kirchler et al., 2008](#)). The tax compliance behaviour of the majority may create a prevailing social norm of compliance and suppress one’s decision to evade taxes. Altruistic concerns, such as improving the welfare of others, may also influence the compliance decision ([Bosco and Mittone, 1997](#)).

We distinguish between the following three types of tax morale related nudges: public goods, social norms and moral appeals. Public good nudges make it clear that the taxes paid by individuals are effectively used to finance public goods and services: “Your tax payment contributes to the funding of publicly financed services in

education, health and other important sectors of society” (Bott et al., 2020). Social norm nudges stress that the majority of individuals in a given country/community are complying with taxes: “Nine out of ten people pay their taxes on time” (Hallsworth et al., 2017).⁶ Moral appeal nudges aim to appeal to morality, fairness, and altruism to influence taxpayer behaviour: “If the taxpayers did not contribute their share, our commune with its 6226 inhabitants would suffer greatly. With your taxes you help keep Trimbach attractive for its inhabitants” (Torgler, 2004).

Other nudges: The sub-category of other nudges includes communications that are relatively rare and are not coherent in the type of content they introduce. Studies introduce distinct types of information content such as sentences on tax-deductible donations (Biddle et al., 2018), instructions on how to file returns (Eerola et al., 2019), various other textual and visual communications (De Neve et al., 2021, Schächtele et al., 2023), among others.

2.4 Tax compliance measures

Extensive margin compliance: Our main dependent variable of interest is the extensive margin of tax compliance. This captures whether a taxpayer is compliant with taxes or not, and is measured with binary outcome variables. More specifically, the studies measure the probability of taxpayers in paying, filing or reporting their taxes,⁷ and they may also distinguish between whether the compliance was full or partial. Following DellaVigna and Linos (2022), we focus on extensive margin of compliance as

⁶As such, our social norm label refers to the descriptive social norm that depicts what most people in a group (or in a society) usually do.

⁷Few studies consider other compliance measures such as whether the taxpayers registered for TV tax, revised the submitted report, or made agreements to pay these taxes.

our main measure for three main reasons. First, this is the most popular measure of compliance used in the literature we study. Second, the binary nature of the outcome variable allows us to measure the impacts of nudges with a common metric, which is the percentage point difference in the outcome relative to the control group. Third, for a meaningful interpretation of the magnitudes of effects, we not only need to measure the effect of the treatment versus the control group, but we also want to have a metric for the level of compliance in the baseline against which these effects can be judged. This metric, labelled as the underlying compliance level, is based on the extensive margin concept and informs us about the share of compliant taxpayers in the control group at the end of the intervention.

Other measures of compliance: Tax compliance can also be measured by focusing on the intensive margin of compliance, which measures the extent or the intensity of compliance. The intensive margin is, however, typically context-dependent and the effect magnitudes are generally not comparable across studies. What is possible instead is to compare the direction and statistical significance of these effects by collecting data on the t-values of treatment effect estimates. The main benefit of this variable is that t-values are available and comparable for all outcomes studied in the literature, including outcome measures at both intensive and extensive margins. We follow other applications of meta-analytical techniques in economics, such as those by [Card et al. \(2010, 2017\)](#), and, in a robustness exercise, study t-values.

2.5 Treatment effect estimates

Experimental designs: The literature uses two main experimental designs to identify the treatment effects of nudges on tax compliance. These design differences have

Table 1: Stylised summary of the framework

(1)	(2)	(3)	(4)	(5)
Treatment nudge	Treatment effect estimates		Post-treatment compliance levels	
	Control group:		In addition to c	In relation to c
	No letter	Reminder	(% of population)	(% change)
Reminder	r	—	$c + r$	$(r)/c$
(Reminder &) Non-deterrence	$r + n$	n	$c + r + n$	$(r + n)/c$
(Reminder &) Deterrence	$r + d$	d	$c + r + d$	$(r + d)/c$

The parameters r, n, d of columns (2) and (3) represent the treatment effect estimates collected from the underlying studies. The c of columns (4) and (5) stands for the underlying compliance level, that is the share of compliant taxpayers in the control group at the end of the intervention, as discussed in Section 2.4.

to do with how the control group is defined. One approach does not treat the taxpayers in the control group in any way, i.e., the taxpayers in this group do not receive any communication from the authority. Another approach always treats the taxpayers in the control group with reminder letters. These two experimental designs are shown in columns (2) and (3) of Table 1, respectively, where we provide a stylised summary of the framework we are in.

Experiments with no communication sent to the control group: The first design typically, but not necessarily, will have a treatment arm that sends reminder letters as shown in column (2) of Table 1. The comparison of this treatment to the control group of taxpayers not receiving any communication will be informative about the reminder effect, r . In these studies, non-deterrence and deterrence nudges are again compared to the control group that did not receive any communication. Therefore, these comparisons lead to treatment effect estimates that also capture the reminder effect, i.e., one treatment effect estimate for reminder and non-deterrence nudges, $r + n$, and another treatment effect estimate for reminder and deterrence nudges, $r + d$.

Experiments with reminders sent to the control group: Studies using the second experimental design always send reminder letters to the control group of taxpayers, as shown in column (3) of Table 1. These studies are not informative about the effects of reminders. Since studies following this design compare non-deterrence and deterrence nudges to the control group that receives reminders, they will be informative about the effects of non-deterrence and deterrence nudges net of the reminder effects, that is n and d , respectively.

Effect magnitudes: To measure the magnitudes of effects, the literature typically reports underlying compliance levels, c , that is the share of compliant taxpayers in the control group at the end of the intervention. We collect this data and calculate the reminder, r , reminder and non-deterrence, $r + n$, and reminder and deterrence, $r + d$, effects compared to the underlying compliance levels. As shown in columns (4) and (5) of Table 1, we calculate this post-treatment compliance levels both in terms of level and relative terms.

2.6 Study characteristics

Basic characteristics: The two basic characteristics that all experiments have are: i) the type of nudges, as defined in Section 2.3, in the most general specification classifying nudges into deterrence or non-deterrence types; and ii) the experimental design, in particular the composition of the control group against which nudges are evaluated as defined in Section 2.5, that is whether the control group received a reminder letter or did not receive any communication.

Additional characteristics: We identify seven additional characteristics as being the defining features of nudging RCTs as follows: iii) late payer sample, i.e., whether the taxpayer is identified as being late in paying her taxes by the official deadline or not; iv) the response horizon of the compliance measure, which we define as a binary variable capturing whether the time interval between the date when the nudge was sent and the date when the outcome variable was measured is shorter or longer than 2 months; v) the year and publication status of the study, i.e., a working paper or a published article; vi) the delivery method used by the tax authority to reach out to the taxpayers, i.e., digital letters, physical letters, or in-person visits; vii) the type of tax being studied, i.e., personal income tax, corporate income tax, property tax, VAT, or other taxes;⁸ viii) the taxpayer type in the sample, i.e., individuals, businesses or a sample mixing both individuals and businesses; and, finally, ix) the income level of the country where the experiment was conducted, i.e., low-, middle- or high-income country.

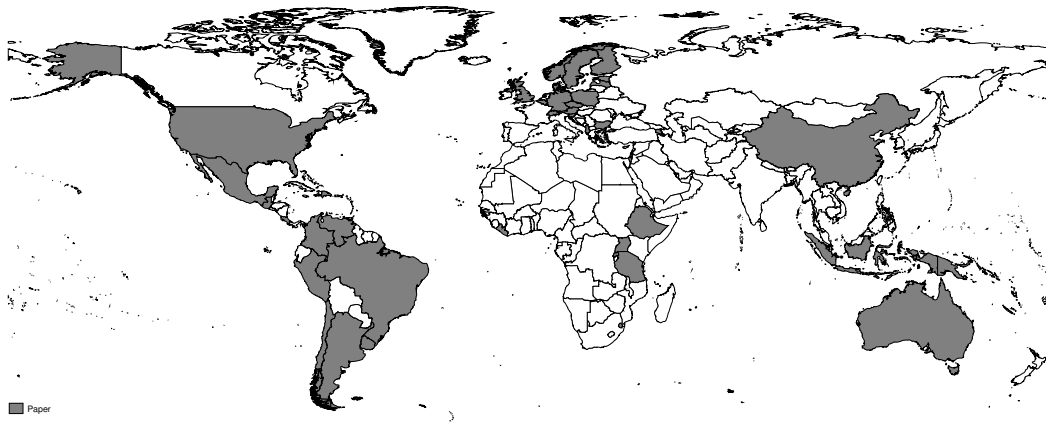
3 Sample of studies and estimates

3.1 Sample of studies

Literature search: We ran a literature search on a rolling basis throughout 2019 to 2023. First we searched for relevant papers using a defined combination of keywords in the main literature databases of the profession. Second, to identify ongoing work, we continued our search in the programs of the main general-interest conferences in economics as well as the main conferences specialising in behavioural or experimental

⁸Other taxes include country-specific taxes or fees, e.g., church tax in Germany, wealth tax in Colombia, TV license fees in Austria, etc.

Figure 1: Country coverage of nudging experiments



Notes: The studies where these experiments come from are listed in Table [A1](#).

economics and public economics.⁹ Third, we carefully looked through the bibliographic information in the papers identified in the last two steps to further refine the study sample. Fourth, we also considered papers sent to us directly by scholars working in the field. In October 2023, we re-visited all of the working papers identified earlier to check their publication status.

Study inclusion criteria: For a paper to be included in our sample, all of the following four criteria need to be fulfilled: i) the study is based on a RCT performed at the level of taxpayers (i.e., individuals or firms rather than, e.g., regions); ii) the trial introduces a nudging intervention which closely follows the definition of [Thaler and Sunstein \(2008\)](#); iii) the dependent variable of interest is the tax compliance behaviour of the taxpayer (either the extensive or the intensive margin or both); and iv) the study

⁹The keywords include: randomised controlled trial, RCT, field experiment, nudging, nudges, behavioural intervention, tax evasion, tax compliance, and tax non-compliance. The literature databases include: Econlit, Google Scholar, and Science Direct. The conferences include ones organised by: AEA, EEA, ESA, SABE, WEAI, NTA, and IIPF.

reports the relevant statistics necessary for our meta-analysis (i.e., the effect size) for at least one treatment effect estimate.

After applying the four filters to the list of papers collected from our extensive search we arrive at an overall sample of 71 studies. These studies are listed in alphabetical order in Table A1. As presented in the map of Figure 1, these experiments were performed in around 40 countries situated mainly in Europe and the Americas, and fewer of them in the developing countries of Africa and Asia.

3.2 Sample of estimates

Treatment effect estimate inclusion criteria: After having defined the sample of studies, we need to decide which treatment effect estimates to collect from these studies. One approach would be to select all estimates that authors report. However, the studies included in our sample differ starkly in the number of estimates, and this approach would run the risk that papers reporting very many estimates, for example from multiple robustness tests, would drive our results. Therefore, we perform our baseline analysis on the “main” estimates reported in studies, while using the full (i.e., main and non-main) sample of treatment effect estimates for robustness checks. In the analysis of publication selection bias of Section 6 we use the latter full sample of treatment effect estimates, since when analysing potential selection effects we are interested in the full sample of estimates that papers report, rather than the main sample of estimates which is chosen by us.

We apply the following seven rules when collecting the estimates from studies and when defining which of these represent the main estimates. First, both in the main and in the full samples, we only consider those estimates that compare the effect of

a nudging intervention to the average compliance in the control group. Thus, we do not consider those estimates that compare the effect of nudging interventions across different treatments. Second, as main estimates, we collect treatment effects utilising the full sample of taxpayers, while those focusing on certain sub-samples, such as for purposes of heterogeneity analysis, are classified among non-main estimates. Third, we consider ITT as the main estimate.¹⁰ Fourth, when studies report results from specifications with (possibly several sets of different) control variables or none at all, we consider as main estimates the latter specifications that do not include any control variables. Fifth, we restrict the main estimates to include only the effects measured in a time horizon of up to around 12 months after the intervention. Sixth, when studies report that their estimates are contaminated by other enforcement activities by tax authorities, we exclude them from both the main and the non-main samples if the time horizon is less than 12 months,¹¹ and include them in the non-main sample when the horizon is 12 months or longer. Seventh, if studies report effects measured over many time horizon after the intervention, we select three of these effects as our main estimate by taking the effects measured at the shortest, the longest and at the middle time horizon.

Dimensions of main estimates within papers: Table [A1](#) reports the number of observations with which each study contributes to the sample. In the extensive margin

¹⁰Studies in our sample either report ITT estimates only, or both ITT and ToT estimates. ITT estimates include every subject that is randomised according to randomised treatment assignment, disregarding non-compliance, protocol deviations, withdrawal, and anything that happens after randomisation ([Gupta, 2011](#)). ToT estimates present the treatment effect on the group of taxpayers who received the communication from the tax administration, using the treatment assignment as an instrumental variable (IV) for the actual treatment ([Mascagni, 2018](#)). The only exception is [Mogollón et al. \(2021\)](#), where the experiment was stopped and failed to treat everyone it intended to. For this specific paper ToT is counted as the main estimate.

¹¹For example, 48-day and 70-day estimates in [Hallsworth et al. \(2017\)](#) are contaminated by external letters sent by tax authorities.

sample, we have 535 estimates from 54 paper of which we consider 270 estimates from 53 papers to be as main estimates. In the t-value sample, which includes both extensive and intensive margin estimates of compliance, we have 928 estimates from 65 paper of which we consider 447 estimates from 62 papers to be as main estimates.

Column (1) of Table [A1](#) suggests that papers in our baseline sample of main estimates measured at extensive margin of compliance can have up to 24 estimates. The range in the number of estimates per paper is determined by the following three dimensions. First, as discussed in Section [2.4](#), studies may use several outcome variables when measuring tax compliance at the extensive margin. In particular, they can report up to four compliance measures – probability to pay, file, report or other as represented in column “compliance measure” of Table [A2](#) – also distinguishing whether the compliance was full or partial as shown in the column “full compliance”. Second, studies will typically have more than one treatment arm in their experimental design. This is primarily driven by the types of nudges, as discussed in Section [2.3](#), which can be up to five in a given study as shown in the column “number of nudge types”. The treatment arm can also vary due to the method of delivering the nudge, which can be up to three – physical letter (L), digital letter (D) or in-person visit (P) – as shown in the column “delivery method”. Third, studies will also tend to report the effects of nudges measured at different time horizons. As explained above, we restrict the baseline sample to not more than three time horizons. These horizons are shown in months for every study in column “time horizons”.

Thus, the largest possible number of estimates per paper is given by interacting all possibilities provided by these three dimensions. In practice, however, the studies will naturally use much fewer and also different combinations of these dimensions. Taking the example of [Chirico et al. \(2019\)](#), the study that contributes with the most number

of observations to our sample, it has 24 main estimates as shown in column (2) of Table A1. These treatment effect estimates are collected from Table 2 of the study, and are available for two outcome variables (full or partial compliance), seven treatment effects that include four nudge types (plus reminder nudges as fifth type), and for effects measured over two time horizons (1 month, 3 months after intervention).

Baseline and other samples of estimates: To conclude, we have 270 treatment effect estimates in our baseline sample coming from 53 studies. These are the compliance effects measured at the extensive margin and represent only the estimates that we consider to be the main ones. Note that this baseline sample only includes the treatment effects of type n , $r + n$, d and $r + d$ of Table 1, and leaves out the effects of reminder nudges, r . Table A2 presents this sample in detail by describing the main characteristics of the papers in this sample paper-by-paper. The summary statistics for this baseline sample are presented in Table 2.

The robustness analysis of Section 5.3 is performed on several further samples. First, for the robustness analysis of the reminder effect we consider again the reminder nudges, r . Two papers only send reminder nudges (Antinyan et al., 2021, Orlett et al., 2017), which leads to the sample increasing to 55 papers. However, since we focus here only on the experiments that do not treat the control group with any communication, this sample consists of 32 studies and 172 estimates. Column (2) of Table A1 shows the studies where these treatment effects r come from. Second, when considering both main and non-main estimates, the sample of extensive margin estimates increases to 535 estimates from 54 papers, as shown in column (3) of Table A1. Third, when considering compliance effects measured by their t-values which contain estimates not only of extensive margin but also of intensive margin compliance, our sample of main

Table 2: Summary statistics

	(1)		(2)	
		Extensive margin	T-values	
Dependent Variable				
Treatment Effect (mean)	270	0.035	447	3.299
Control Group				
Reminder vs No Letter	270	123	447	181
Nudge Type				
Deterrence	270	126	447	212
Public Good	270	39	447	75
Social Norm	270	40	447	49
Moral Appeal	270	16	447	41
Other	270	49	447	70
Late-Payers				
General Sample (vs Late-Payers)	270	115	447	233
Response Horizon				
Short Run	270	138	447	209
Publication Status				
Published	270	174	447	267
Year of Publication (mean)	270	2019	447	2019
Delivery				
Letter	270	175	447	264
Digital	270	81	447	159
In Person	270	14	447	24
Tax Type				
Income Tax	270	116	447	184
Corporate Tax	270	16	447	25
Property Tax	270	76	447	103
VAT	270	11	447	43
Other	270	33	447	69
Multiple	270	18	447	23
Taxpayer Type				
Individual	270	169	447	225
Business	270	65	447	153
Individual and Business	270	36	447	69
Development Level				
Low Income	270	27	447	44
Middle Income	270	108	447	190
High Income	270	135	447	213

Summary statistics show the total number of observations, and the number of observations satisfying the respective criteria. For the dependent variable, its mean values are shown.

estimates increases to 62 papers and 447 estimates. These studies are shown in column (4) of Table A1.¹² Fourth and finally, is our largest sample consisting of both main and non-main estimates of t-values and also including the reminder effects. This sample has 928 estimates coming from 65 papers and is described in column (6) of Table A1. The publication bias analysis of Section 6 is based on this latter largest sample.

4 Empirical design

Baseline specification: We estimate the following equation:

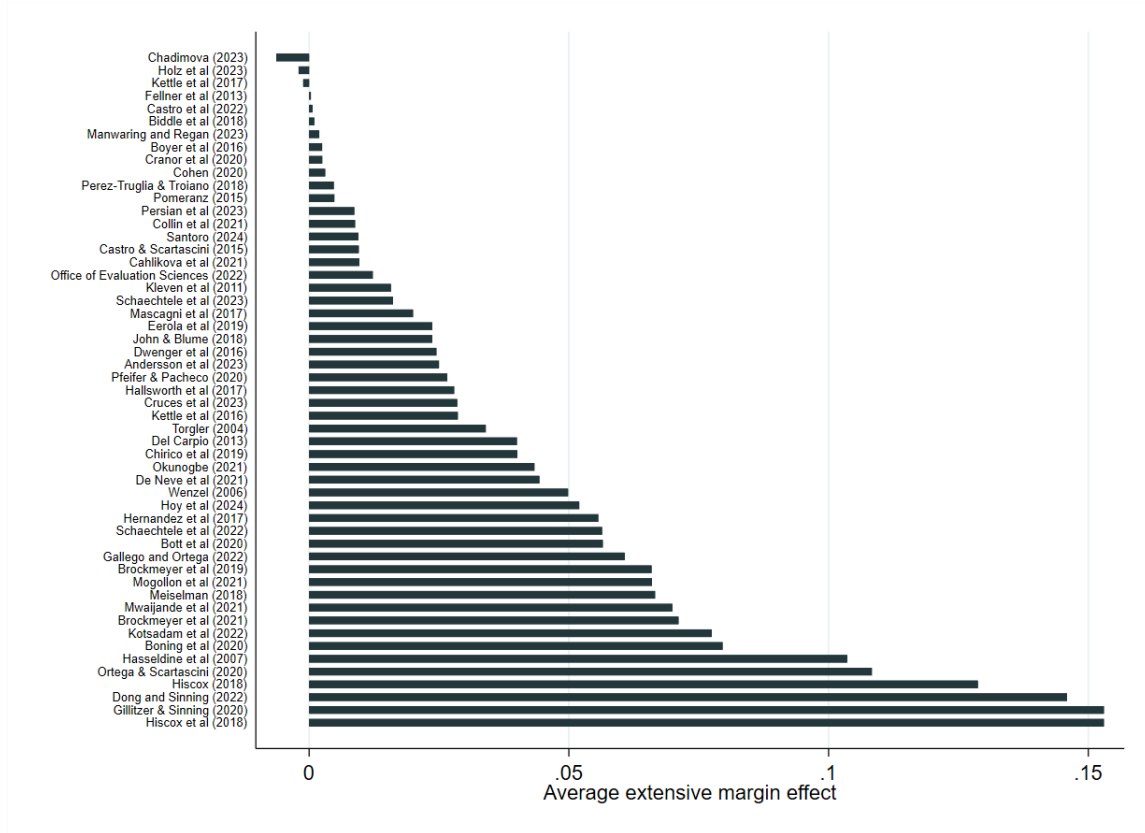
$$ComplianceEffect_{i,p} = \alpha Reminder_p + \beta Deterrence_{i,p} + \gamma NonDeterrence_{i,p} + \epsilon_{i,p} \quad (1)$$

Dependent variable: $ComplianceEffect_{i,p}$ is the i^{th} estimate of extensive margin treatment effect from paper p . The distribution of this variable averaged by study is plotted in Figure 2. In all of our analyses, we winsorise the dependent variable at its 5th and 95th percentiles.¹³ In additional results described in Section 5.3, we run this specification taking the t-values of all treatment effects as an alternative dependent variable.

¹²This t-value sample must be larger than the extensive margin sample since it includes the extensive margin estimates. The only exception is when the standard error for a given extensive or intensive margin coefficient cannot be measured by us. For example, this is the case for all extensive margin estimates coming from Perez-Truglia and Troiano (2018).

¹³This winsorisation strategy is not uncommon in the literature, and is motivated by the possibility of having errors in the data as well as with the aim of reducing the effect of outliers on the average estimates. For example, Card et al. (2017) winsorise the data at the 10th and 90th percentiles. Table 6 provides a robustness exercise by running our baseline specifications on non-winsorised data.

Figure 2: Distribution of average treatment effect estimates across studies



Notes: The figure depicts the average treatment effect estimate on extensive margin compliance among main estimates by paper in ascending order of average effect magnitude.

Independent variables: $Reminder_p$ is a binary variable capturing whether, in a given experiment, the control group received a reminder letter or no communication at all, as defined in Section 2.5. Our sample has about the equal number of these two types of estimates. Consequently, $\hat{\alpha}$ compares the treatment effect estimates across the two experimental designs fixing for the type of the nudge that was sent.¹⁴ Since we do not have studies that utilise both types of experiments, the reminder effect is only identified from across study variation. In Section 5.3, we present an alternative

¹⁴If a non-deterrence nudge was sent $\hat{\alpha} = r + n - n = r$, and if a deterrence nudge was sent $\hat{\alpha} = r + d - d = r$.

estimator for the reminder effect which uses within study variation. However, this is possible to do only using the sample of experiments that have a control group receiving no communication and a send a reminder nudge that is compared to that control group.

$Deterrence_{i,p}$ is again a binary variable equal to one if a nudge is of deterrence type and equal to zero if a nudge is of non-deterrence type. We have about as many deterrence as non-deterrence nudges in the sample. Consequently, $\hat{\beta}$ represents the average treatment effect of a deterrence nudge when the respective control group receives a reminder letter, thus identifying the effect d . Since most studies in our sample send both deterrence and non-deterrence type nudges, we are able to, in Section 5.3, provide an alternative estimator for the deterrence effect identified from within study variation using study fixed effects.

Conversely, $\hat{\gamma}$ captures the average effect of non-deterrence type nudges compared to a control group of taxpayers receiving reminders, that is the effect n . As with the reminder and deterrence effects, we provide further evidence on these non-deterrence nudges in Section 5.3, in particular by dividing them into more detailed types of tax morale nudges.

Error term: $\epsilon_{i,p}$ is the error term. Since the compliance effect estimates may not be independent within studies, we cluster the error term at the level of studies p .

5 Results

5.1 Baseline

Table 3 shows the estimation results of Equation 1. Column (1) runs the regression for the baseline sample of 270 estimates coming from 53 papers, and column (2) repeats

Table 3: Baseline results

	(1) Baseline sample	(2) Reduced baseline (where c is observed)
Reminder $\hat{\alpha}$	0.026*** (0.009)	0.027*** (0.010)
Non-Deterrence $\hat{\gamma}$	0.012** (0.005)	0.014** (0.006)
Deterrence $\hat{\beta}$	0.031*** (0.007)	0.032*** (0.008)
Observations	270	218
Adjusted R^2	0.496	0.507
Papers	53	44
Postestimation p-values:		
Deterrence = Non-Deterrence	0.012	0.026
Deterrence = Reminder	0.702	0.745
Reminder = Non-Deterrence	0.273	0.349

Regressions are estimated according to Equation 1. Column (2) restricts the baseline sample to observations where the underlying compliance level, c , is observed.

Standard errors in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

the analysis for the sub-sample of this where the underlying compliance level, c , is available leading to a reduced sample of 218 estimates coming from 44 papers. First, regarding the reminder effect, $\hat{\alpha}$ suggests a 2.6 percentage point difference between the treatment effects of the experiments sending no communication to the control group and those sending reminder letters to the control group. Second, $\hat{\beta}$ suggests that deterrence nudges increase the probability of compliance by 3.1 percentage points compared to the control group receiving reminder letters. And third, $\hat{\gamma}$ suggests that non-deterrence nudges increase the probability of compliance by 1.2 percentage points compared to the control group receiving reminder letters. All of these effects are statistically distinguishable from zero at least at the 5% level. Post-estimation tests across the three parameters suggest that neither deterrence nor non-deterrence effects

Table 4: Effect magnitudes

(1)	(2)	(3)	(4)	(5)
Treatment effect	Estimator	Estimated effects Table 3 col. 2 (percentage points)	Post-treatment compliance levels In addition to c (% of population)	In relation to c (% change)
Reminder	$r = \hat{\alpha}$	=2.7p.p	27.7%	10.8%
Rem. & Non-det.	$r + n = \hat{\alpha} + \hat{\gamma}$	=4.1p.p	29.1%	16.4%
Rem. & Deterrence	$r + d = \hat{\alpha} + \hat{\beta}$	=5.9p.p	30.9%	23.6%

Parameters r, n, d of column (2) represent the reminder, non-deterrence and deterrence effects, as discussed in Table 1. The effects, $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ are estimated according to Equation 1 in Table 3 (column 2). Columns (4) and (5) show the post-treatment compliance levels in addition and in relation to the underlying compliance level, c .

are statistically distinguishable from the reminder effect, but the deterrence effect is distinguishable from the non-deterrence effect. The results from the reduced sample where the underlying compliance levels are observed are plotted in column (2). These are very similar to those from the baseline sample of column (1). We present our headline calculations on the post-treatment compliance levels using the estimates from this latter sample.

5.2 Effect magnitudes

To understand the magnitudes of the effects identified so far, we compare them to underlying levels of compliance in our sample, c . The average share of compliant taxpayers in the sample which have not received any communication is 25%. As discussed earlier this number is low and also, with a standard deviation of 21.8, masks quite some heterogeneity. However, this is not surprising, since more than half of RCTs in our sample work with samples of late-payers where compliance is close to zero. Another reason is that about one-third of RCTs in our sample were conducted in low- and middle-income countries, where baseline compliance rates can be quite low.

Table 4 summarises the discussion on the magnitudes of our effects, building on the structure of Table 1. Column (2) shows how the estimates we identify map into the treatment effect estimates of the underlying studies. Column (3) shows the three treatment effects using the parameters estimated in Table 3: Reminder effect, r , is identified by $\hat{\alpha}$, the effect of reminder and non-deterrence nudges, $r + n$, is identified by $\hat{\alpha} + \hat{\gamma}$, and the effect of reminder and deterrence nudges, $r + d$, is identified by $\hat{\alpha} + \hat{\beta}$. Columns (4) and (5) then show the post-treatment compliance levels in addition and in relation to the underlying compliance level. Compared to the share of compliant taxpayers of 25%, reminders shift the compliance level to 27.7% of the population or increase compliance by 10.8% relative to the underlying average, while reminders combined with non-deterrence and deterrence nudges increase compliance to, respectively, the levels of 29.1% and 30.9% of the population or by 16.4% and 23.6% relative to the baseline.

5.3 Robustness tests

Reminder effect: In our baseline approach we identify the reminder effect using variation across papers. An alternative strategy is to focus on the sub-sample of experiments that do not treat the control group with any communication, include the estimates on the effects of reminder nudges r in the data and compare their effects to the control group of untreated taxpayers. The specification is as follows:

$$ComplianceEffect_{i,p} = \alpha_1 ReminderNudge_{i,p} + \beta_1 Deterrence_{i,p} + \gamma_1 NonDeterrence_{i,p} + \epsilon_{i,p}^1 \quad (2)$$

where $\hat{\alpha}_1$, $\hat{\gamma}_1$ and $\hat{\beta}_1$ respectively identify the effects of reminder r , non-deterrence $r + n$ and deterrence $r + d$ nudges compared to the control group that receives no communication. Column (1) of Table 5 runs Equation 2 on this sub-sample of experiments. The finding of a 2.2 percentage point effect of reminders is similar to the baseline results on the effects of reminders presented in Table 3.¹⁵

Deterrence effect: Almost all of the studies in our sample implement several nudge interventions. Therefore, we can exploit the substantial within-study variation in the data and, as a robustness exercise, study the effects of deterrence nudges compared to that of non-deterrence nudges on tax compliance within studies. The specification is as follows:

$$ComplianceEffect_{i,p} = \beta_2 Deterrence - vs - NonDeterrence_{i,p} + \lambda_p + \epsilon_{i,p}^2 \quad (3)$$

where λ_p is a new term capturing the study fixed effects. $\hat{\beta}_2$ is the main coefficient of interest which identifies the effect of deterrence nudges on tax compliance compared to that of non-deterrence nudges. Column (2) of Table 5 suggests that deterrence nudges increase the probability of compliance by 2.4 percentage points compared to the effects of sending non-deterrence nudges controlling for study fixed effects. This estimate is consistent with the baseline results of Table 3 where the deterrence effect compared to the non-deterrence, i.e. $\hat{\beta} - \hat{\gamma}$, was 1.9 percentage points and statistically different from zero.

¹⁵The deterrence and non-deterrence effects are likewise similar to the baseline results. Note that $\hat{\beta}_1$ should be compared to $\hat{\alpha} + \hat{\beta}$ and $\hat{\gamma}_1$ to $\hat{\alpha} + \hat{\gamma}$.

Table 5: Robustness tests for reminder, deterrence and tax morale nudges

	(1) Reminder	(2) Deterrence	(3) Non-Deterrence
Reminder $\hat{\alpha}_1$	0.022*** (0.004)		
Non-Deterrence $\hat{\gamma}_1$	0.033*** (0.008)		
Deterrence $\hat{\beta}_1$	0.061*** (0.009)		
Det-vs-NonDet $\hat{\beta}_2$		0.024*** (0.006)	
Non-Deterrence Effect $\hat{\gamma}^\kappa$ of type:			
Public Good			-0.029*** (0.007)
Social Norm			-0.023*** (0.006)
Moral Appeal			-0.028*** (0.007)
Other			-0.019*** (0.006)
Paper FE	No	Yes	Yes
Observations	172	270	270
Adjusted R^2	0.534	0.647	0.645
Papers	32	53	53
Postestimation p-values:			
Public Good = Social Norm			0.175
Public Good = Moral Appeal			0.922
Social Norm = Moral Appeal			0.294

Regressions of columns (1), (2) and (3) are estimated according to Equations 2, 3 and 4.

Standard errors in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Tax morale effects: Instead of grouping nudges into the general deterrence and non-deterrence categories, we unpack non-deterrence nudges according to the detailed elements of tax morale that they target. Following our definitions of Section 2.3, we divide non-deterrence nudges into public good, social norm, moral appeal and other nudges. As shown in Table 2 public good, social norm and moral appeal nudges are

about as numerous as deterrence nudges. Additionally, a tenth of the nudges are grouped into the category of other nudges.¹⁶ Since the exact definitions of these detailed types of nudges can vary substantially across studies, we run this specification using study fixed effects as introduced in Equation 3. The variable *NonDeterrence – vs – Deterrence* now equals zero for deterrence nudges and serves as the omitted category, and it has separate categories for the four types of detailed non-deterrence nudges identified by γ_2^κ .

$$ComplianceEffect_{i,p} = \gamma_2^\kappa \sum_{\kappa=1}^4 NonDeterrence-vs-Deterrence_{i,p}^\kappa + \lambda_p + \epsilon_{i,p}^2 \quad (4)$$

Column (3) of Table 5 presents the estimation results. They suggest that the individual effects of public good, moral appeal and social norm nudges are smaller than those of deterrence nudges. The magnitudes of these three effects are tested for equality in the bottom of the table. These post-estimation tests do not find robust differences across the three types of non-deterrence nudges. Overall, this evidence suggests that not only the average tax morale nudge but also its three sub-categories – public good, social norm and moral appeal nudges – do not stand out as being as important drivers of tax compliance as deterrence nudges.

Alternative estimators: Our baseline specification is estimated using an ordinary least squares (OLS) estimator. In columns (1) and (2) of Table 6, we follow a number of recent applications of meta-analytical techniques in economics (see, e.g., Card et al., 2010, 2017, Heinemann et al., 2018, Lichter et al., 2015, Neisser, 2021) and the

¹⁶Several interventions mix different types of non-deterrence nudges. Our analysis classifies such instances in the following hierarchical order: deterrence, moral appeal, social norm, and public good. That is, if, for example, a message contains both a moral appeal and a public good nudge, we classify it as a moral appeal nudge.

Table 6: Robustness to alternative estimators, samples and compliance measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Estimator		Sample definition			Compliance measure		Main &	
	WLS (s.e.)	Ran- dom effects	Non- winsor- ised	Non- nega- tive	<90th %tile	T- value	Pr. to pay	Pr. to file	non-main esti- mates
Reminder $\hat{\alpha}$	0.017* (0.010)	0.023* (0.013)	0.029*** (0.010)	0.020** (0.009)	0.030** (0.012)	1.582 (0.998)	0.029** (0.012)	0.016 (0.026)	0.031*** (0.011)
Non-Det $\hat{\gamma}$	0.009* (0.006)	0.013 (0.011)	0.012** (0.006)	0.021*** (0.005)	0.016** (0.007)	1.489** (0.719)	0.015** (0.007)	0.013 (0.021)	0.010 (0.006)
Det $\hat{\beta}$	0.023*** (0.007)	0.035*** (0.013)	0.030*** (0.008)	0.040*** (0.007)	0.029*** (0.009)	3.320*** (0.758)	0.035*** (0.010)	0.036 (0.025)	0.029*** (0.009)
Obs.	238	238	270	228	185	447	157	58	535
R^2	0.373		0.455	0.565	0.505	0.457	0.579	0.428	0.500
Papers	46	46	53	53	47	62	31	12	54

Regressions are estimated according to Equation 1. Column (1) estimates a weighted least squares model, where the inverse of squared standard errors are the weights. Column (2) estimates a Random Effects model. Column (3) uses non-winsorised data as the outcome variable. Column (4) excludes negative values from the outcome variable. Column (5) excludes estimates from studies that present 12 (90th percentile) or more estimates. Column (6) uses t-values instead of extensive margin treatment effect estimates as the outcome variable. Column (7) and (8) focus on the sub-sample where the outcome variable is either the probability to pay or the probability to file. Column (9) includes non-main estimates in addition to main ones. Standard errors in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

literature reviewing these methods (see, e.g., [Stanley, 2001](#), [Stanley and Doucouliagos, 2012](#)), and show the robustness of the OLS results to those using weighted least squares (WLS) and random effects estimators. We use a WLS estimator since meta-analytical regressions are known to be heteroscedastic.¹⁷ As analytical weights we take the inverse of the squared standard error of the parameter estimates, which, unlike the OLS approach, yields to precision-weighted estimates. We also adopt a random effects model. This estimator assumes the existence of a distribution of true effects for distinct studies and populations. Thus, we relax the assumption that for each nudge type there exists a single “true” effect which is common to all studies under consideration. In general, the results are similar to the baseline findings.¹⁸

¹⁷One form of heteroskedasticity arises because the variance in the individual estimates is negatively related to the size of the underlying sample and this correlation is likely to be different between the primary studies.

¹⁸Columns (1) and (2) of Table 6 have fewer observations than the baseline sample of 270 observations because for some estimates in our data the standard errors are missing.

Sample definitions: We present four additional tests for the sensitivity of our results to the choice of the sample. First, in column (3) of Table 6, we replicate the baseline results using non-winsorised data. Second, in column (4) Table 6, we test whether the effects we have identified are driven by the few negative treatment effects in our sample. Third, in column (5) of Table 6, we drop papers which present extraordinarily many treatment effects estimates to test whether a few papers that have unusually many estimates drive our results. To that end, six papers are excluded that have more than 12 estimates which is the 90th percentile in the distribution of the number of estimates. Fourth, in Figure B1 we implement jackknife-type robustness tests by excluding papers one-by-one from the sample. Overall, these tests suggest that for all of our three effects of interest, the confidence intervals estimated in the robustness exercises always include the average estimate of the baseline approach.

Tax compliance measures: We perform two exercises. First, we study a larger sample of estimates that measure compliance both at the extensive and intensive margins, and, second, we narrow down the measurement of extensive margin compliance to two more specific definitions. As discussed earlier, the magnitudes of the estimates that measure compliance at the intensive margin cannot be compared to each other due to the heterogeneity in measurements. However, we can compare their direction and statistical significance by using the t-values of treatment effect estimates as our dependent variable of interest. The main benefit of this exercise is that t-values are available for both types of compliance measures increasing our sample from 270 to 447 estimates. This approach is not uncommon in the field of meta-analysis, and is used by many applications in economics (Card et al., 2010, 2017), sometimes even as their primary outcome variable of interest (Baskaran et al., 2016, Heinemann et al., 2018).

The results presented in column (6) of Table 6 suggest that the direction of all three main effects, that is of reminder, deterrence and non-deterrence nudges, are robust in this larger sample. Columns (7) and (8) of Table 6 then narrow down the measurement of extensive margin compliance to two more specific definitions of compliance: probability to pay, which make about 60% of our data, and probability to file, which makes another about 20% of the data. Although we lose statistical power in the latter and reduced sample, the estimates are not contradictory to the baseline findings.

Main and non-main estimates: Our baseline results use the sample of main treatment effect estimates, and here we extend the analysis to using both main and non-main treatment effect estimates. The sample increases from 270 to 535 treatment effect estimates. The results from running the baseline regressions on this sample are presented in column (9) of Table 6. Overall, these results are not different from the baseline findings of Table 3.

5.4 Study characteristics

We are interested in the role of a number of important study characteristics in explaining the heterogeneity in the estimates of the compliance effect $ComplianceEffect_{i,p}$. To do so, we introduce vector $\mathbf{X}_{i,p}$ of these study characteristics to Equation 1 on top of the baseline regressors.

$$ComplianceEffect_{i,p} = \alpha Reminder_p + \beta Deterrence_{i,p} + \gamma NonDeterrence_{i,p} + \delta \mathbf{X}_{i,p} + \epsilon_{i,p} \quad (5)$$

The study characteristics enter the regression first one-by-one and then jointly.¹⁹ These seven characteristics are defined in Section 2.6. Table 2 shows that in our sample more than half of estimates deal with samples of taxpayers who were late in paying their taxes; that about half of the responses are measured in the short time horizon of less than two months after treatment; that more than half of the estimates come from papers that have been published already; that the mean study year is 2019; that most of the communication takes place through either physical or digital letters as opposed to through few in person visits; that estimates are most likely to come from experiments undertaken with personal income taxpayers but substantial part also coming from property taxpayers, and fewer corporate income and VAT taxpayers; that about two-third of our sample interventions target individuals; and that half of estimates come from experiments conducted in high-income countries and fewer come from middle-income and especially low-income countries.²⁰

The regressions are presented in Table 7. The basic characteristics are always included. Reassuringly, we find that the point estimates on these characteristics, that is $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$, remain robust to the inclusion of the additional study characteristics as control variables. Taken together, these study characteristics can explain an additional 14 percentage points of the heterogeneity in the treatment effect estimates of nudges.²¹

Our findings are as follows. First, nudges are more effective when addressing sub-samples of taxpayers who missed their deadline for paying taxes. The magnitude of the effect is 2.7 percentage points. Second, the treatment effects are stronger in the

¹⁹This strategy follows DellaVigna and Linos (2022) for example, and is motivated by the fact that the inclusion of all characteristics is too demanding of a specification given our data such that, due to multi-collinearity of regressors, we are likely to be left with little variation to exploit.

²⁰According to the definition and data of World Bank (2021).

²¹This refers to the difference in adjusted R^2 in regressions without and with the additional study characteristics. Adjusted R^2 in Table 3, column (1) is 49.6% and in Table 7, column (8) it is 63.2%.

Table 7: Role of study characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Reminder $\hat{\alpha}$	0.028*** (0.008)	0.029*** (0.009)	0.025*** (0.009)	0.027*** (0.009)	0.025*** (0.008)	0.023*** (0.009)	0.030*** (0.009)	0.027*** (0.008)
Non-Deterrence $\hat{\gamma}$	0.024*** (0.007)	0.016** (0.006)	0.014** (0.006)	0.012* (0.006)	0.045 (0.029)	0.012** (0.005)	0.019*** (0.006)	0.034 (0.028)
Deterrence $\hat{\beta}$	0.040*** (0.008)	0.038*** (0.008)	0.033*** (0.008)	0.029*** (0.008)	0.066** (0.029)	0.031*** (0.008)	0.041*** (0.009)	0.052* (0.028)
Late-Payers (omitted: Late-Payer Sample)								
General Sample	-0.027*** (0.008)							-0.016** (0.008)
Response Horizon (omitted: Short Run)								
		-0.014* (0.009)						-0.012* (0.007)
Publication Status (omitted: Published)								
Unpublished			-0.003 (0.009)					0.002 (0.009)
Year			-0.002 (0.001)					-0.001 (0.001)
Delivery (omitted: Physical Letter)								
Digital Letter				-0.004 (0.009)				0.017** (0.008)
In Person				0.026* (0.015)				0.058*** (0.012)
Tax Type (omitted: VAT)								
Income Tax					-0.032 (0.031)			-0.008 (0.027)
Corporate Tax					-0.064** (0.030)			-0.035 (0.025)
Property Tax					-0.034 (0.030)			-0.006 (0.029)
Other					-0.035 (0.031)			-0.018 (0.028)
Multiple					-0.017 (0.041)			-0.009 (0.029)
Taxpayer Type (omitted: Individual)								
Business						0.012 (0.012)		0.024** (0.011)
Individual and Business						-0.008 (0.011)		-0.004 (0.010)
Development Level (omitted: High Income)								
Low Income							-0.023* (0.013)	-0.043*** (0.015)
Middle Income							-0.019* (0.010)	-0.021*** (0.008)
Observations	270	270	270	270	270	270	270	270
Adjusted R^2	0.553	0.510	0.505	0.505	0.527	0.505	0.524	0.631
Papers	53	53	53	53	53	53	53	53

Regressions are estimated according to Equation 5.

Standard errors in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

short-run, that is within a horizon of two months after the intervention, compared to effects measured two months after the intervention. Third, we do not find evidence for significant differences between working papers and published papers, and nor by the year of the study. Fourth, nudges communicated by in person visits to taxpayers relative to nudges delivered through physical letters have 2.6 percentage point larger effects on tax compliance. Fifth, we study whether the effects differ across tax types and, relatedly, across individuals or businesses. We do not find robust evidence for such differences. Sixth, and finally, when comparing experimental results across countries where the RCTs were conducted, we find that experiments seem to be less effective in low- and middle-income countries compared to high-income countries. We note that these findings remain suggestive since our inference is correlational and is often based on small samples. Future work may try to use experiments to study the role of these characteristics more directly. This can be done by designing interventions that go beyond short horizons, tax bases and country borders, among other dimensions.

6 Publication selection bias

We study whether treatment effects reported by the studies in our sample are systematically selected towards having the “right” sign and towards being more statistically significant. The two underlying hypotheses are that researchers tend to present results that show: i) positive effects because it is generally expected, either according to theory or due to conventional beliefs, that nudges should only have positive effects (sign bias), and ii) statistically significant effects because of a predisposition to treat significant results more favourably, for example, due to the belief that non-significant effects are harder to publish (p-hacking). As discussed earlier, in this section we focus

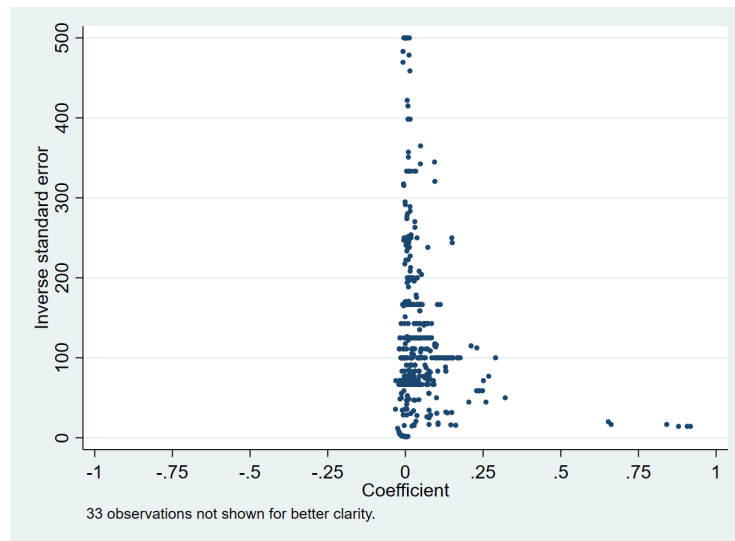
on our largest sample consisting of 928 estimates coming from 65 papers, which is presented in column (6) of Table [A1](#).

6.1 Sign bias

We use a funnel plot to provide visual checks for asymmetries in the relationship between treatment effect directions and magnitudes on the one hand, and measures of their precision on the other hand. The idea is that, absent publication bias, very imprecise estimates should be randomly distributed around zero rather than being skewed to one direction, resembling an inverted funnel. In our particular case, the hypothesis is that the sign bias will lead to imprecise estimates being skewed to the right, that is towards positive treatment effect estimates. We present a funnel plot in Figure [3](#) where the x-axis plots the size of the treatment effect at the extensive margin and the y-axis plots the inverse standard error of the treatment effect as a measure of precision. We observe that the imprecisely estimated treatment effects, i.e., those at the bottom of the funnel plot, tend to be skewed towards positive values. This visual evidence provides suggestive evidence for the presence of sign bias in our sample.

More formally, we use the method of [Egger et al. \(1997\)](#) to test for funnel-plot asymmetry. Table [B1](#) regresses the normalised coefficient, i.e., the point estimate divided by the standard error, on a measure of its precision, i.e., the inverse of the standard errors, and an intercept. The coefficient of interest is the intercept, which provides a measure of asymmetry. If there is asymmetry, with smaller studies showing effects that differ systematically from larger studies, the intercept will be different from zero. Our results reject the null hypothesis that there is no publication bias of this type. This evidence suggests that the published estimates on the treatment effects of nudges

Figure 3: Funnel plot



Notes: The figure plots a correlation between the size of the treatment effect and the inverse of its standard error. For visual clarity, outliers larger than 500 on the y-axis are dropped, arriving at a sample of 490 estimates.

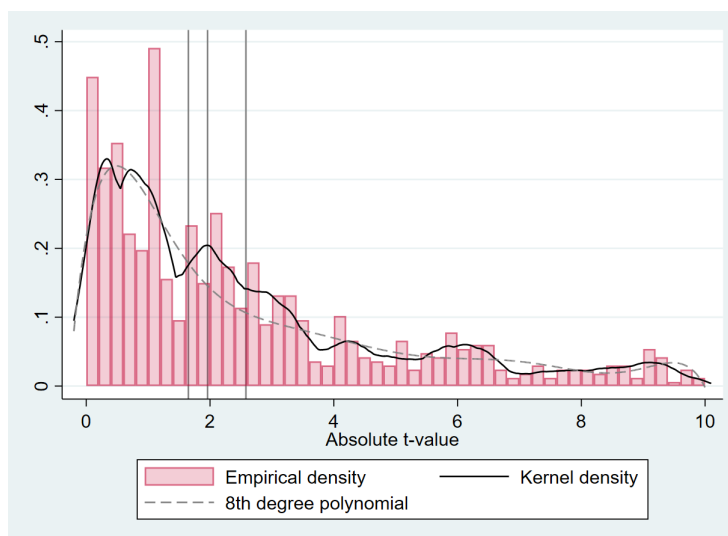
are systematically selected towards showing the “right” sign, that is towards showing effects that increase tax compliance and ignoring those that decrease tax compliance.

6.2 P-hacking bias

We now turn to the study of p-hacking, and check for unusual patterns in the distribution of t-values around their critical values. [Brodeur et al. \(2016\)](#) use a large dataset comprising of tests published in top economics journals, and show a disproportionately large share of tests that narrowly reject the null hypothesis. We follow this approach and plot the distribution of absolute t-values in Figure 4. This first visual evidence suggests some bunching in the number of observations of t-values situated just to the right of the three critical values (which are denoted by vertical lines).

More formally, we follow [Brodeur et al. \(2020\)](#) and implement three exercises. First, in Figure 4 we estimate a counterfactual distribution by fitting a polynomial

Figure 4: Distribution of absolute t-values



Notes: This figure plots a histogram of absolute t-values for treatment effect estimates in 0.2 wide bins. The kernel density line is estimated according to an Epanechnikov function with bandwidth 0.2. The 8th degree polynomial is estimated for the counterfactual case by dropping absolute t-values in the [1.5, 3] range. For visual clarity, the figure drops outliers outside the (-10, 10) range and includes 835 observations. Vertical lines denote critical values for two-sided significance tests at t-values of 1.645, 1.96 and 2.58.

function to data that drops t-values in the region around the three critical values, and contrasting this to the kernel density estimated on the whole distribution. This exercise shows visually the presence of excess mass in the density around the area of the critical values. Second, to statistically confirm the visually observable discontinuities, we use the randomisation test, and check for discontinuities in the probability of a t-value appearing above or below the critical values. The idea is that, absent p-hacking, the probability of being just above versus just below any threshold should be equal. Panel A of Table B2 performs this test using the data of t-values centred around the three significance thresholds and several local bandwidths going from 0.075 to 0.1575 (the largest value before we cross the next closest critical value). The test shows that there are discontinuities in the distribution of t-values, with over 60% of observations

in these bandwidths being skewed towards showing statistically significant effects. In columns (4) to (6) of Table B2, we implement a version of this test that studies several bandwidths around the 5% significance threshold rather than centring the data around the three thresholds. We find similar results. This evidence suggests that the studies in our sample choose to report results that are statistically significant at conventional levels, and ignore reporting treatment effect estimates that narrowly fail to reject the null hypothesis. Third and finally, we use the caliper tests again following Brodeur et al. (2020), to study whether p-hacking is more likely to take place in certain sub-samples. We consider the type of the nudge and the experimental design, and also compare published papers versus working papers, new versus old papers (split at the median study year in our sample) and main estimates versus other estimates. The coefficients shown in Panel B of Table B2 represent increases in the probability of finding a statistically significant effect relative to the baseline category. We do not find robust evidence for large differences in p-hacking patterns across any of these dimensions.

Overall, our evidence suggests that our sample is biased due to both sign as well as p-hacking type biases. We find it unlikely that the existing biases can explain the difference in the effects of reminder and deterrence nudges that we document. This analysis suggests that empirical studies implementing RCTs, which are otherwise believed to have relatively sound methodologies, may not be immune to biased reporting of results.

7 Conclusions

Policy interventions that nudge taxpayers with the aim of increasing compliance have become a popular tool among many governments owing to their ease of implementation and low monetary costs. This easy adoption of the policy is demonstrated, for example, by [Hjort et al. \(2021\)](#), who inform randomly selected Brazilian mayors about research on the positive tax compliance effects of reminder letters and find that the treated municipalities are more likely to implement nudging interventions. However, little is known about the effectiveness of nudges beyond the evidence presented in individual experiments carried out in different contexts.

In this paper we quantitatively summarise the knowledge accumulated from tax nudging interventions in a systematic way. We estimate the average effects of reminder, tax morale, and deterrence nudges on tax compliance. We find positive effects of all these nudges and provide precise numbers on their magnitudes. These estimates may help policymakers form more realistic expectations about the impact of nudges rather than having to rely on the outcomes of individual studies. Our evidence on the particular design features of interventions that make them more or less effective can provide further guidance for potentially more effective policy interventions in the future.

This review highlights a number of opportunities for researchers by directing attention towards gaps in the literature where the evidence has been weak so far. Few papers test whether nudges work in the longer run, and when implemented repeatedly. Although it is plausible that nudges shift decisions from the future to the present, we are not aware of studies that identify such potential inter-temporal (crowding) effects of nudges. Having better measurements about taxpayers' priors, perhaps by borrowing

techniques from the literature on survey experiments (for a review, see, [Fuster and Zafar, 2022](#)), would help understand the mechanism through which nudges operate more exactly. We are also not aware of studies that try to measure and then take into account the costs of nudging in the tax compliance context. Although the direct or implementation related costs are probably negligible, the indirect costs of nudges – such as the annoyance costs of reminders, the psychological costs of tax morale nudges, or the potentially risk-aversion inducing effects of “intimidating” deterrence nudges – can be substantial, thereby degrading the positive effects of nudges from a welfare perspective. Finally, we do not have much knowledge of how interventions interact with the context they operate in. This is not surprising given that randomised control trials tend to narrowly focus on local environments where the context is fixed. Cross-study comparisons such as the one adopted in this paper, on the other hand, are limited owing to methodological concerns when comparing different experiments. Future interventions, possibly ones that span borders or institutional environments, could study whether, for example, tax morale nudges work more effectively in contexts with higher levels of trust, and whether deterrence nudges work better in uncorrupted environments where audits can be enforced more credibly than in institutionally less mature environments.

References

- Allcott, H. and J. B. Kessler (2019). The welfare effects of nudges: A case study of energy use social comparisons. *American Economic Journal: Applied Economics* 11(1), 236–276.
- Allingham, M. and A. Sandmo (1972). Income tax evasion: A theoretical analysis. *Journal of Public Economics* 1(3-4), 323–338.
- Alm, J. (2012). Measuring, explaining, and controlling tax evasion: Lessons from theory, experiments, and field studies. *International Tax and Public Finance* 19(1), 54–77.
- Alm, J. (2019). What motivates tax compliance? *Journal of Economic Surveys* 33(2), 353–388.
- Alm, J. and A. Malézieux (2020). 40 years of tax evasion games: A meta-analysis. *Experimental Economics*, 1–52.
- Altmann, S. and C. Traxler (2014). Nudges at the dentist. *European Economic Review* 72, 19–38.
- Andersson, H., P. Engström, K. Nordblom, and S. Wanander (2023). Nudges and threats: Soft vs hard incentives for tax compliance. *Economic Policy*, eiad017.
- Andreoni, J., B. Erard, and J. Feinstein (1998). Tax compliance. *Journal of Economic Literature* 2(36), 818–860.
- Antinyan, A. and Z. Asatryan (2020). Tax compliance nudges in Armenia. *Mimeo*.

- Antinyan, A., Z. Asatryan, Z. Dai, and K. Wang (2021). Does the frequency of reminders matter for their effectiveness? A randomized controlled trial. *Journal of Economic Behavior & Organization* 191, 752–764.
- Ariel, B. (2012). Deterrence and moral persuasion effects on corporate tax compliance: Findings from a randomized controlled trial. *Criminology* 50(1), 27–69.
- Baskaran, T., L. P. Feld, and J. Schnellenbach (2016). Fiscal federalism, decentralization, and economic growth: A meta-analysis. *Economic Inquiry* 54(3), 1445–1463.
- Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy* 2(76), 169–217.
- Benartzi, S., J. Beshears, K. L. Milkman, C. R. Sunstein, R. H. Thaler, M. Shankar, W. Tucker-Ray, W. J. Congdon, and S. Galing (2017). Should governments invest more in nudging? *Psychological science* 28(8), 1041–1055.
- Bergolo, M., R. Ceni, G. Cruces, M. Giacobasso, and R. Perez-Truglia (2023). Tax audits as scarecrows: Evidence from a large-scale field experiment. *American Economic Journal: Economic Policy* 15(1), 110–153.
- Bernheim, B. D., A. Fradkin, and I. Popov (2015). The welfare economics of default options in 401(k) plans. *American Economic Review* 105(9), 2798–2837.
- Bhattacharya, J., A. M. Garber, and J. D. Goldhaber-Fiebert (2015). Nudges in exercise commitment contracts: A randomized trial. *National Bureau of Economic Research, Working Paper 21406*.

- Biddle, N., K. M. Fels, and M. Sinning (2018). Behavioral insights on business taxation: Evidence from two natural field experiments. *Journal of Behavioral and Experimental Finance* 18, 30–49.
- Blackwell, C. (2007). A meta-analysis of tax compliance experiments. In Martinez-Vazquez and J. Alm (Eds.), *Tax Compliance and Evasion*.
- Blumenstock, J., M. Callen, and T. Ghani (2018). Why do defaults affect behavior? Experimental evidence from Afghanistan. *American Economic Review* 108(10), 2868–2901.
- Blumenthal, M., C. Christian, J. Slemrod, and M. G. Smith (2001). Do normative appeals affect tax compliance? Evidence from a controlled experiment in Minnesota. *National Tax Journal* 54(1), 125–138.
- Boning, W. C., J. Guyton, R. Hodge, and J. Slemrod (2020). Heard it through the grapevine: The direct and network effects of a tax enforcement field experiment on firms. *Journal of Public Economics* 190, 104261.
- Bosco, L. and L. Mittone (1997). Tax evasion and moral constraints: Some experimental evidence. *Kyklos* 50(3), 297–324.
- Bott, K. M., A. W. Cappelen, E. Sorensen, and B. Tungodden (2020). You’ve got mail: A randomized field experiment on tax evasion. *Management Science* 66(7), 2801–2819.
- Boyer, P. C., N. Dwenger, and J. Rincke (2016). Do norms on contribution behavior affect intrinsic motivation? Field-experimental evidence from Germany. *Journal of Public Economics* 144, 140 – 153.

- Brockmeyer, A., A. Estefan, K. R. Arras, and J. C. S. Serrato (2021). Taxing property in developing countries: Theory and evidence from Mexico. *National Bureau of Economic Research Working Paper 28637*.
- Brockmeyer, A., M. Hernandez, S. Kettle, and S. Smith (2019). Casting a wider tax net: Experimental evidence from Costa Rica. *American Economic Journal: Economic Policy* 11(3), 55–87.
- Brodeur, A., N. Cook, and A. Heyes (2020). Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review* 110(11), 3634–2660.
- Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics* 8(1), 1–32.
- Bursztyn, L. and D. Y. Yang (2022). Misperceptions about others. *Annual Review of Economics* 14, 425–452.
- Cahlikova, J., L. Cingl, K. Chadimova, and M. Zajicek (2021). Carrots, sticks, or simplicity? Field evidence on what makes people pay TV fees. *Max Planck Institute for Tax Law and Public Finance Working Paper 2021-12*.
- Calzolari, G. and M. Nardotto (2017). Effective reminders. *Management Science* 63(9), 2915–2932.
- Card, D., J. Kluve, and A. Weber (2010). Active labour market policy evaluations: A meta-analysis. *The Economic Journal* 120(548), F452–F477.

- Card, D., J. Kluve, and A. Weber (2017). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association* 16(3), 894–931.
- Castro, J. F., D. Velásquez, A. Beltrán, and G. Yamada (2022). The direct and indirect effects of messages on tax compliance: Experimental evidence from Peru. *Journal of Economic Behavior & Organization* 203, 483–518.
- Castro, L. and C. Scartascini (2015). Tax compliance and enforcement in the Pampas: Evidence from a field experiment. *Journal of Economic Behavior & Organization* 116, 65 – 82.
- Chadimova, K. (2023). Timing, deterrence & simplicity in repetitive nudges. *Available at SSRN 4455664*.
- Chirico, M., R. Inman, C. Loeffler, J. MacDonald, H. Sieg, J. A. Mortenson, H. R. Schramm, A. Whitten, D. Shoag, C. Tuttle, et al. (2019). Deterring property tax delinquency in philadelphia: An experimental evaluation of nudge strategies. *National Tax Journal* 72(3), 479–506.
- Cohen, I. (2020). Low-cost tax capacity: A randomized evaluation on tax compliance with the Uganda Revenue Authority. *NTA Annual Conference, Mimeo*.
- Coleman, S. (1996). The Minnesota income tax compliance experiment: State tax results. *MPRA Paper No. 4827*.
- Collin, M., V. Di Maro, D. K. Evans, and F. Manang (2021). Property tax compliance in Tanzania. *Global Economy and Development at Brookings Working Paper No 161*.

- Coricelli, G., M. Joffily, C. Montmarquette, and M. C. Villeval (2010). Cheating, emotions, and rationality: An experiment on tax evasion. *Experimental Economics* 13(2), 226–247.
- Costa, D. L. and M. E. Kahn (2013). Energy conservation “nudges” and environmentalist ideology: Evidence from a randomized residential electricity field experiment. *Journal of the European Economic Association* 11(3), 680–702.
- Cranor, T., J. Goldin, T. Homonoff, and L. Moore (2020). Communicating tax penalties to delinquent taxpayers: Evidence from a field experiment. *National Tax Journal* 73(2), 331–360.
- Cruces, G., D. Tortarolo, and G. Vazquez-Bare (2023). Design of partial population experiments with an application to spillovers in tax compliance. *Institute for Fiscal Studies*.
- Damgaard, M. T. and C. Gravert (2018). The hidden costs of nudging: Experimental evidence from reminders in fundraising. *Journal of Public Economics* 157, 15–26.
- De Neve, J.-E., C. Imbert, J. Spinnewijn, T. Tsankova, and M. Luts (2021). How to improve tax compliance? Evidence from population-wide experiments in Belgium. *Journal of Political Economy* 129(5), 1425–1463.
- Del Carpio, L. (2013). Are the neighbors cheating? evidence from a social norm experiment on property taxes in Peru. Working paper, Princeton University.
- DellaVigna, S. and E. Linos (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica* 90(1), 81–116.

- Dizon-Ross, R. (2019). Parents' beliefs about their children's academic ability: Implications for educational investments. *American Economic Review* 109(8), 2728–65.
- Doerrenberg, P., A. Pfang, and J. Schmitz (2023). How to improve small firms' payroll tax compliance? Evidence from a randomized field experiment. *Saïd Business School Working Paper 2023-04*.
- Doerrenberg, P. and J. Schmitz (2017). Tax compliance and information provision: A field experiment with small firms. *Journal of Behavioral Economics for Policy* 1(1), 47–54.
- Dong, S. X. and M. Sinning (2022). Trying to make a good first impression: A natural field experiment to engage new entrants to the tax system. *Journal of behavioral and experimental economics* 100, 101900.
- Dulleck, U., J. Fooker, C. Newton, A. Ristl, M. Schaffner, and B. Torgler (2016). Tax compliance and psychic costs: Behavioral experimental evidence using a physiological marker. *Journal of Public Economics* 134, 9–18.
- Dwenger, N., H. Kleven, I. Rasul, and J. Rincke (2016). Extrinsic and intrinsic motivations for tax compliance: Evidence from a field experiment in Germany. *American Economic Journal: Economic Policy* 8(3), 203–32.
- Dwenger, N. and L. Treber (2018). Shaming for tax enforcement: Evidence from a new policy. *CEPR Discussion Papers No. 13194*.
- Eerola, E., T. Kosonen, T. Lyytikäinen, and J. Tuimala (2019). Tax compliance in the rental housing market: Evidence from a field experiment. *VATT Institute for Economic Research Working Papers No 122*.

- Egger, M., G. D. Smith, M. Schneider, and C. Minder (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 315(7109), 629–634.
- Fellner, G., R. Sausgruber, and C. Traxler (2013). Testing enforcement strategies in the field: Threat, moral appeal and social information. *Journal of the European Economic Association* 11(3), 634–660.
- Finkelstein, A. and M. J. Notowidigdo (2019). Take-up and targeting: Experimental evidence from SNAP. *The Quarterly Journal of Economics* 134(3), 1505–1556.
- Fochmann, M., N. Müller, and M. Overesch (2018). Less cheating? The effects of prefilled forms on compliance behavior. *ARQUS Discussion Paper No. 227*.
- Fuster, A. and B. Zafar (2022). Survey experiments on economic expectations. *National Bureau of Economic Research, Working Paper 29750*.
- Gallego, J. and F. Ortega (2022). Can facebook ads and email messages increase fiscal capacity? Experimental evidence from Venezuela. *Economic Development and Cultural Change* 70(4), 1531–1563.
- Gillitzer, C. and M. Sinning (2020). Nudging businesses to pay their taxes: Does timing matter? *Journal of Economic Behavior & Organization* 169, 284–300.
- Gillitzer, C. and P. E. Skov (2018). The use of third-party information reporting for tax deductions: Evidence and implications from charitable deductions in Denmark. *Oxford Economic Papers* 170(3), 892–916.
- Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research* 2(3), 109.

- Hallsworth, M., J. A. List, R. D. Metcalfe, and I. Vlaev (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics* 148, 14 – 31.
- Handel, B. R. (2013). Adverse selection and inertia in health insurance markets: When nudging hurts. *American Economic Review* 103(7), 2643–82.
- Harju, J., T. Kosonen, and O. Ropponen (2018). Do honest hairdressers get a haircut? *mimeo*.
- Hasseldine, J., P. Hite, S. James, and M. Toumi (2007). Persuasive communications: Tax compliance enforcement strategies for sole proprietors. *Contemporary Accounting Research* 24(1), 171–194.
- Heinemann, F., M.-D. Moessinger, and M. Yeter. (2018). Do fiscal rules constrain fiscal policy? A meta-regression-analysis. *European Journal of Political Economy* 51, 69–92.
- Hernandez, M., J. Jamison, E. Korczyk, N. Mazar, and R. Sormani (2017). Applying behavioral insights to improve tax collection: Experimental evidence from Poland. Working paper, The World Bank.
- Hiscox, M. (2018). Improved compliance with the deferred GST scheme. *Behavioural Economics Team of the Australian Government, Working Paper*.
- Hiscox, M., J. Bialecki, H. Cotching, M. Daffey, H. Greenwell, S. M. Kyaw-Myint, K. Rajah, R. Slonim, and B. Weeks (2018). Improving tax compliance: Deductions for work-related expenses. *Behavioural Economics Team of the Australian Government, Working Paper*.

- Hjort, J., D. Moreira, G. Rao, and J. F. Santini (2021). How research affects policy: Experimental evidence from 2,150 Brazilian municipalities. *American Economic Review* 111(5), 1442–1480.
- Holz, J. E., J. A. List, A. Zentner, M. Cardoza, and J. E. Zentner (2023). The \$100 million nudge: Increasing tax compliance of firms using a natural field experiment. *Journal of Public Economics* 218, 104779.
- Hoy, C., L. McKenzie, and M. Sinning (2024). Improving tax compliance without increasing revenue: Evidence from population-wide randomized controlled trials in Papua New Guinea. *Economic Development and Cultural Change* 72(2), 691–723.
- Huck, S. and I. Rasul (2010). Transactions costs in charitable giving: Evidence from two field experiments. *The BE Journal of Economic Analysis & Policy* 10(1).
- Hummel, D. and A. Maedche (2019). How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics* 80, 47–58.
- John, P. and T. Blume (2018). How best to nudge taxpayers? The impact of message simplification and descriptive social norms on payment rates in a central London local authority. *Journal of Behavioral Public Administration* 1(1), 1–11.
- Karlan, D., M. McConnell, S. Mullainathan, and J. Zinman (2016). Getting to the top of mind: How reminders increase saving. *Management Science* 62(12), 3393–3411.
- Karver, J. G., H. Shijaku, and C. T. Ungerer (2022). Nudging in the time of the coronavirus. *World Bank Policy Research Working Paper* 9961.

- Kettle, S., M. Hernandez, S. Ruda, and M. Sanders (2016). Behavioral interventions in tax compliance: Evidence from Guatemala. World bank policy research working papers 7690.
- Kettle, S., M. Hernandez, M. Sanders, O. Hauser, and S. Ruda (2017). Failure to captcha attention: Null results from an honesty priming experiment in Guatemala. *Behavioral Sciences* 7(2), 1–21.
- Kirchler, E., E. Hoelzl, and I. Wahl (2008). Enforced versus voluntary tax compliance: The “slippery slope” framework. *Journal of Economic psychology* 29(2), 210–225.
- Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica* 79(3), 651–692.
- Kleven, H. J., C. T. Kreiner, and E. Saez (2016). Why can modern governments tax so much? An agency model of firms as fiscal intermediaries. *Economica* 330(83), 219–246.
- Kotakorpi, K. and J.-P. Laamanen (2016). Prefilled income tax returns and tax compliance: Evidence from a natural experiment. *University of Tampere Working Paper No. 104*.
- Kotsadam, A., K. Løyland, O. Rauum, G. Torsvik, and A. Øvrum (2022). Does perceived risk of future audits explain the behavioral effects of tax enforcement? *12th Norwegian German Seminar on Public Economics*.
- Lichter, A., A. Peichl, and S. Siegloch (2015). The own-wage elasticity of labor demand: A meta-regression analysis. *European Economic Review* 80, 94–119.

- Linos, E., A. Prohofsky, A. Ramesh, J. Rothstein, and M. Unrath (2022). Can nudges increase take-up of the EITC: Evidence from multiple field experiments. *American Economic Journal: Economic Policy* 14(4), 432–452.
- List, J. A., M. Rodemeier, S. Roy, and G. K. Sun (2023). Judging nudging: Understanding the welfare effects of nudges versus taxes. *National Bureau of Economic Research, Working Paper No. 31152*.
- Luttmer, E. F. and M. Singhal (2014). Tax morale. *Journal of Economic Perspectives* 28(4), 149–68.
- Manwaring, P. and T. Regan (2023). Public disclosure and tax compliance: Evidence from Uganda. *Centre for Economic Performance Discussion paper No 1937*.
- Mascagni, G. (2018). From the lab to the field: A review of tax experiments. *Journal of Economic Surveys* 32(2), 273–301.
- Mascagni, G., C. Nell, and N. Monkam (2017). One size does not fit all: A field experiment on the drivers of tax compliance and delivery methods in Rwanda.
- Meiselman, B. (2018). Ghostbusting in Detroit: Evidence on nonfilers from a controlled field experiment. *Journal of Public Economics* 158, 180–193.
- Mertens, S., M. Herberz, U. J. Hahnel, and T. Brosch (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences* 119(1).
- Mogollón, M., D. Ortega, and C. Scartascini (2021). Who's calling? The effect of phone calls and personal interaction on tax compliance. *International Tax and Public Finance*, 1–27.

- Mwaijande, F., M. Kachwamba, J. Mwakalikamo, and D. Shirima (2021). Local authorities and tax collection: Experimental evidence from Tanzania.
- Neisser, C. (2021). The elasticity of taxable income: A meta-regression analysis. *The Economic Journal* 131(640), 3365–3391.
- Office of Evaluation Sciences (2022). Increasing voluntary tax compliance for return preparers. Available at <https://oes.gsa.gov/assets/abstracts/2007-abstract.pdf>.
- Okunogbe, O. (2021). Becoming legible to the state: The role of detection and enforcement capacity in tax compliance.
- Orlett, S., R. Javaid, V. Koranda, M. Muzikir, and A. Turk (2017). Impact of filing reminder outreach on voluntary filing compliance for taxpayers with a prior filing delinquency.
- Ortega, D. and P. Sanguinetti (2013). Deterrence and reciprocity effects on tax compliance: Experimental evidence from Venezuela. CAF Working Papers 08/2013, Development Bank Of Latin America.
- Ortega, D. and C. Scartascini (2020). Don't blame the messenger. The delivery method of a message matters. *Journal of Economic Behavior & Organization* 170, 286–300.
- Perez-Truglia, R. and U. Troiano (2018). Shaming tax delinquents. *Journal of Public Economics* 167, 120–137.
- Persian, R., G. Prastuti, D. Bogiatzis-Gibbons, M. H. Kurniawan, G. Subroto, M. Mustakim, L. Scheunemann, K. Gandy, A. Sutherland, et al. (2023). Behavioural prompts to increase early filing of tax returns: A population-level randomised controlled trial of 11.2 million taxpayers in Indonesia. *Behavioural Public Policy* 7(3), 701–720.

- Pfeifer, F. F. and T. S. Pacheco (2020). Increasing tax compliance with behavioral insights: Evidence from São Paulo. *Sao Paulo School of Business Administration, Mimeo*.
- Pomeranz, D. (2015). No taxation without information: Deterrence and self-enforcement in the value added tax. *American Economic Review* 105(8), 2539–2569.
- Pomeranz, D. and J. Vila-Belda (2019). Taking state-capacity research to the field: Insights from collaborations with tax authorities. *Annual Review of Economics* 11, 755–781.
- Santoro, F. (2024). Income tax payers are not all the same: A behavioral letter experiment in Eswatini. *Economic Development and Cultural Change* 72(2), 000–000.
- Saulitis, A. and P. Chapkovski (2023). Investigating tax compliance with mixed-methods approach: The effect of normative appeals among the firms in Latvia. *Available at SSRN 4373889*.
- Scartascini, C. and E. Castro (2019). Imperfect attention in public policy: A field experiment during a tax amnesty in Argentina. *IDB Discussion Paper No 665*.
- Schächtele, S., H. Eguino, and S. Roman (2022). Improving taxpayer registration through nudging? Field experimental evidence from Brazil. *World Development* 154, 105887.

- Schächtele, S., H. Eguino, and S. Roman (2023). Fiscal exchange and tax compliance: Evidence from a field experiment. *Journal of Policy Analysis and Management* 42(3), 796–814.
- Shimeles, A., D. Z. Gurara, and F. Woldeyes (2017). Taxman’s dilemma: Coercion or persuasion? Evidence from a randomized field experiment in Ethiopia. *American Economic Review: Papers and Proceedings* 107(5), 420—424.
- Slemrod, J. (2007). Cheating ourselves: The economics of tax evasion. *Journal of Economic Perspectives* 21(1), 25–48.
- Slemrod, J. (2019). Tax compliance and enforcement. *Journal of Economic Literature* 57(4), 904–954.
- Slemrod, J., M. Blumenthal, and C. Christian (2001). Taxpayer response to an increased probability of audit: Evidence from a controlled experiment in Minnesota. *Journal of Public Economics* 79(3), 455 – 483.
- Slemrod, J. and C. Gillitzer (2014). *Tax systems*. MIT Press Cambridge, MA.
- Slemrod, J., O. U. Rehman, and M. Waseem (2022). How do taxpayers respond to public disclosure and social recognition programs? Evidence from Pakistan. *The Review of Economics and Statistics* 104(1), 116–132.
- Slemrod, J. and S. Yitzhaki (2002). Tax avoidance, evasion, and administration. In *Handbook of public economics, Vol. 3*, Volume 2, pp. 1423–1470. Elsevier.
- Snow, A. and R. S. Warren (2005). Tax evasion under random audits with uncertain detection. *Economics Letters* 88(1), 97–100.

- Stanley, T. D. (2001). Wheat from chaff: Meta-analysis as quantitative literature review. *Journal of Economic Perspectives* 15(3), 131–150.
- Stanley, T. D. and H. Doucouliagos (2012). *Meta-regression analysis in economics and business*. New York & London: Routledge.
- Sunstein, C. R. (2014). Nudging: a very short guide. *Journal of Consumer Policy* 37(4), 583–588.
- Thaler, R. H. and C. R. Sunstein (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven & London: Yale University Press.
- Torgler, B. (2003). To evade taxes or not to evade: That is the question. *The Journal of Socio-Economics* 32(3), 283–302.
- Torgler, B. (2004). Moral suasion: An alternative tax policy strategy? Evidence from a controlled field experiment in Switzerland. *Economics of Governance* 5(3), 235–253.
- Vainre, M., L. Aaben, A. Paulus, H. Koppel, H. Tammsaar, K. Telve, K. Koppel, K. Beilmann, A. Uusberg, et al. (2020). Nudging towards tax compliance: A fieldwork-informed randomised controlled trial. *Journal of Behavioral Public Administration* 3(1).
- Wenzel, M. (2006). A letter from the tax office: Compliance effects of informational and interpersonal justice. *Social Justice Research* 19(3), 345–364.
- Wenzel, M. and N. Taylor (2004). An experimental evaluation of tax-reporting schedules: A case of evidence-based tax administration. *Journal of Public Economics* 88(12), 2785–2799.

Wisdom, J., J. S. Downs, and G. Loewenstein (2010). Promoting healthy choices: Information versus convenience. *American Economic Journal: Applied Economics* 2(2), 164–78.

World Bank (2021). World bank income classifications. Retrieved August 18, 2021 from <https://datacatalogfiles.worldbank.org/ddhpublished/0037712/DR0090754/OGHIST.xlsx>.

Appendix

Appendix A: Sample of studies

Table A1: Samples of studies

No.	Paper	Country	(1)	(2)	(3)	(4)	(5)	(6)
			Main	Extensive margin Main & Reminders	Main & Non-Main	Main	T-value Main & Non-Main	Main & Non-Main & Reminders
1	Andersson et al. (2023)	Sweden	5	5	10	5	10	10
2	Antinyan and Asatryan (2020)	Armenia	0	0	3	0	9	9
3	Antinyan et al. (2021)	China	0	1	0	0	0	2
4	Ariel (2012)	Israel	0	0	0	8	8	8
5	Bergolo et al. (2023)	Uruguay	0	0	0	6	9	9
6	Biddle et al. (2018)	Australia	3	3	3	6	6	6
7	Blumenthal et al. (2001)	USA	0	0	0	4	4	4
8	Boning et al. (2020)	USA	2	2	8	6	24	24
9	Bott et al. (2020)	Norway	3	3	12	6	24	24
10	Boyer et al. (2016)	Germany	2	2	2	4	4	4
11	Brockmeyer et al. (2019)	Costa Rica	12	12	24	20	32	32
12	Brockmeyer et al. (2021)	Mexico	2	2	4	4	8	8
13	Cahlíkova et al. (2021)	Czech Republic	12	12	12	9	9	9
14	Castro and Scartascini (2015)	Argentina	9	9	18	9	18	18
15	Scartascini and Castro (2019)	Argentina	0	0	0	3	9	9
16	Castro et al. (2022)	Peru	12	12	48	12	48	48
17	Chadimova (2023)	Czech Republic	6	6	18	6	18	18
18	Chirico et al. (2019)	USA	24	28	36	36	54	63
19	Cohen (2020)	Uganda	2	3	4	4	8	12
20	Coleman (1996)	USA	0	0	0	8	8	8
21	Collin et al. (2021)	Tanzania	4	6	8	8	16	24
22	Cranor et al. (2020)	USA	6	6	15	6	15	15
23	Cruces et al. (2023)	Argentina	6	6	6	6	6	6
24	De Neve et al. (2021)	Belgium	3	3	5	3	5	5
25	Del Carpio (2013)	Peru	3	4	6	3	6	8
26	Doerrenberg and Schmitz (2017)	Slovenia	0	0	0	2	4	4
27	Doerrenberg et al. (2023)	Bulgaria	0	0	0	18	18	18
28	Dong and Sinning (2022)	Australia	2	2	4	2	4	4
29	Dwenger et al. (2016)	Germany	1	1	1	2	2	2
30	Eerola et al. (2019)	Finland	3	4	6	9	18	24
31	Fellner et al. (2013)	Austria	3	3	3	3	3	3
32	Gallego and Ortega (2022)	Venezuela	3	3	6	6	12	12
33	Gillitzer and Sinning (2020)	Australia	3	3	6	6	12	12
34	Hallsworth et al. (2017)	UK	13	14	31	13	31	33
35	Harju et al. (2018)	Finland	0	0	0	2	2	2
36	Hasseldine et al. (2007)	UK	5	5	5	0	0	0
37	Hernandez et al. (2017)	Poland	9	9	18	27	54	54
38	Hiscox (2018)	Australia	4	4	4	6	6	6
39	Hiscox et al. (2018)	Australia	1	1	1	0	0	0
40	Holz et al. (2023)	Dominican Republic	5	5	5	10	10	10
41	Hoy et al. (2024)	Papua New Guinea	2	2	2	4	4	4
42	John and Blume (2018)	UK	3	3	3	0	0	0
43	Karver et al. (2022)	Albania	0	0	0	4	8	16
44	Kettle et al. (2016)	Guatemala	8	10	32	16	64	80
45	Kettle et al. (2017)	Guatemala	6	6	6	30	30	30

46	Kleven et al. (2011)	Denmark	2	2	2	4	4	4
47	Kotsadam et al. (2022)	Norway	2	2	2	3	3	3
48	Manwaring and Regan (2023)	Uganda	5	5	10	10	20	20
49	Mascagni et al. (2017)	Rwanda	8	12	16	8	17	26
50	Meiselman (2018)	USA	5	6	10	5	10	12
51	Mogollón et al. (2021)	Colombia	2	2	8	3	12	12
52	Mwaijande et al. (2021)	Tanzania	4	4	4	4	4	4
53	Office of Evaluation Sciences (2022)	USA	4	4	4	4	4	4
54	Okunogbe (2021)	Liberia	8	8	12	12	20	20
55	Orlett et al. (2017)	USA	0	8	0	0	0	0
56	Ortega and Sanguinetti (2013)	Venezuela	0	0	0	0	8	8
57	Ortega and Scartascini (2020)	Colombia	9	9	36	11	45	45
58	Perez-Truglia and Troiano (2018)	USA	4	4	4	0	0	0
59	Persian et al. (2023)	Indonesia	5	5	5	5	5	5
60	Pfeifer and Pacheco (2020)	Brazil	5	5	5	5	5	5
61	Pomeranz (2015)	Chile	2	2	2	4	4	4
62	Santoro (2024)	Eswatini	12	12	24	12	24	24
63	Saulitis and Chapkovski (2023)	Latvia	0	0	0	1	5	8
64	Schächtele et al. (2022)	Brazil	1	1	2	1	2	2
65	Schächtele et al. (2023)	Argentina	2	2	10	2	10	10
66	Shimeles et al. (2017)	Ethiopia	0	0	0	2	4	4
67	Slemrod et al. (2001)	USA	0	0	0	3	3	3
68	Torgler (2004)	Switzerland	1	1	2	1	2	2
69	Vainre et al. (2020)	Estonia	0	0	0	2	2	2
70	Wenzel and Taylor (2004)	Australia	0	0	0	3	3	4
71	Wenzel (2006)	Australia	2	2	2	0	0	0
Estimates			270	296	535	447	856	928
Papers			53	55	54	62	64	65

The table presents the list of studies used in this paper along with the country where the RCT comes from and the number of estimates coming from each study. Column (1) – main estimates on the extensive margin – represents our baseline sample. Columns (2) and (3) additionally include reminder estimates and non-main estimates for this extensive margin. The sample of t-values – based on both extensive and intensive margin estimates – is presented in columns (4) to (6), first restricted to main estimates, then to main and non-main estimates, and finally in column (6) additionally including reminders which makes the sample used for the publication bias analysis.

Table A2: Study-level characteristics of the baseline sample

Paper	No. of estimates	Average estimate	Full compliance	Compliance measure	Control group <i>ReminderP</i>	No. of nudge types	Country	Published	Avg. compliance, c.	Time horizons in months	Share late payer	Delivery methods	Tax type	Taxpayer type
Andersson et al. (2023)	5	.025		PTP	2	2	Sweden	1	.66	1; 3	1	D	Income tax	Ind.
Biddle et al. (2018)	3	.081	Partial	Other	2	3	Australia	1	3	3	0	L	VAT	Bus.
Boning et al. (2020)	2	.01	Partial	PTP	1	1	USA	1	.58	3	0	L; P	Other	Bus.
Bott et al. (2020)	3	.057	Partial	PTP	2	3	Norway	1	.2	.75	0	L	Income tax	Ind.
Boyer et al. (2016)	2	.002	Partial	PTP	2	1	Germany	1	.02	4.6	0	L	Other	Ind.
Brockmeyer et al. (2019)	12	.066	Partial	PTP; PTF; PTP	1	2	Costa Rica	1	.03	3.75	0	D	Income tax	Bus.
Brockmeyer et al. (2021)	2	.071	Partial	PTP	1	2	Mexico	0	.06	1.3	1	L	Property tax	Ind.
Chalikova et al. (2021)	12	.01	Full	PTP	1	4	Czech Rep.	0	.18	1.3; 3; 3.8	1	L	Other	Ind.
Castro and Scartascini (2015)	9	.01	Full	PTP	2	3	Argentina	1	.33; 1; 2	.33; 1; 2	0	L	Property tax	Ind.
Scartascini and Castro (2019)	12	.001	Partial	PTP	2	3	Peru	1	.04	.5; 5.5; 12.5	0	D	Income tax	Ind.
Chadimova (2023)	6	-.006	Full	PTP	2	1	Czech Rep.	0	.18	1; 2	1	L	Other	Ind.
Chirco et al. (2019)	24	.04	Full; Partial	PTP	1	4	USA	1	.37	1; 3	1	L	Property tax	Ind.
Cohen (2020)	2	.003	Partial	PTP	1	2	Uganda	0	.05	3.1	0	D	Income tax	Bus.
Collin et al. (2021)	4	.009	Partial	PTP	1	1	Tanzania	0	.09	.33; 1.33	0	D	Property tax	Ind.
Granor et al. (2020)	6	.003	Full; Partial	PTP	2	2	USA	1	.19	1	1	L	Income tax	Ind.
Cruces et al. (2023)	6	.029	PTP	PTP	1	1	Argentina	0	.2	.15; 1	.5	L	Property tax	Bus.
De Neve et al. (2021)	3	.044	Partial	PTP; PTF; PTP	2	1	Belgium	1	.5	.46; .7; 2	.67	L	Income tax	Ind.
Del Carpio (2013)	3	.04	Partial	PTP	1	2	Peru	0	.29	1	1	L	Property tax	Ind.
Dong and Sinning (2022)	2	.146	Partial	PTP	1	2	Australia	1	.04	1.5	1	L	Income tax	Ind.
Dwenger et al. (2016)	1	.024	Partial	PTP	2	1	Germany	1	.21	5	0	L	Other	Ind.
Erola et al. (2019)	3	.024	Partial	PTP	1	2	Finland	0	.75	1	0	L	Income tax	Ind.
Felner et al. (2013)	3	0	Partial	Other	2	3	Austria	1	.07	1.6	0	L	Other	Ind.
Gallego and Ortega (2022)	3	.061	Full	PTP	1	1	Venezuela	1	.02	1.5	1	D	Other	Ind.; Bus.
Gillitzer and Sinning (2020)	3	.153	Partial	PTP	2	1	Australia	1	.79; 1.05; 1.28	.75	1	L	VAT	Bus.
Hallsworth et al. (2017)	13	.028	Partial	PTP	1	4	UK	1	.34	.4	1	L	Income tax	Ind.
Hasseldine et al. (2007)	5	.104	Partial	PTP	1	3	UK	1	.4	1	0	L	Income tax	Ind.
Hernandez et al. (2017)	9	.056	Partial	PTP	2	5	Poland	0	.4	1	1	L	Income tax	Ind.
Hiscox (2018)	4	.129	Partial	PTP; PTF; PTP	1	2	Australia	0	.29	.5; .7	1	D	Multiple	Bus.
Hiscox et al. (2018)	1	.153	Partial	Other	1	1	Australia	0	.01	2	0	L	Income tax	Ind.
Holz et al. (2023)	5	-.002	Partial	PTF	2	2	Dominican Rep.	1	.51	4	0	D	Corporate tax	Bus.
Hoy et al. (2024)	2	.052	Partial	PTF	2	1	Papua New Guinea	1	.21	.467	1	D	VAT	Bus.
John and Blume (2018)	3	.024	Full	PTP	2	2	UK	1	.43	1	0	L	Property tax	Ind.
Kettle et al. (2016)	8	.029	Partial	PTF; PTF	1	1	Guatemala	0	0	2.5	1	L	Income tax	Ind.; Bus.
Kettle et al. (2017)	6	-.001	Partial	PTF	1	3	Guatemala	1	0	0	0	D	Multiple	Ind.; Bus.
Kleven et al. (2011)	2	.016	Partial	Other	1	1	Denmark	1	.14	1	0	L	Income tax	Ind.
Kotsadam et al. (2022)	2	.078	Partial	Other	1	1	Norway	0	0	4; 10	0	D	Income tax	Ind.
Manwaring and Regan (2023)	5	.002	Partial	PTP	2	4	Uganda	0	.04	1.4	0	D	Property tax	Ind.; Bus.
Mascagni et al. (2017)	8	.02	Partial	PTF	1	2	Rwanda	0	2	2	0	D; P	Income tax	Bus.
Meiselman (2018)	5	.067	Partial	PTP	1	3	USA	1	0	2.4	1	L	Income tax	Ind.
Mogollón et al. (2021)	2	.066	Full; Partial	PTP	1	1	Colombia	1	.05	2	1	L	Multiple	Ind.; Bus.
Mwajande et al. (2021)	4	.07	Partial	PTP	1	1	Tanzania	0	.04	.47	0	D	Property tax	Ind.; Bus.
Office of Evaluation Sciences (2022)	8	.043	Partial	Other	1	1	USA	0	.29	5	0	L	Income tax	Ind.
Okunogbe (2021)	9	.108	Full; Partial	PTP; PTF; Other	2	2	Liberia	0	.08	6	1	P	Property tax	Ind.; Bus.
Ortega and Scartascini (2020)	4	.005	Partial	Other	1	2	Colombia	1	3; 4	3; 4	1	L; D; P	Other	Bus.
Perez-Truglia and Troiano (2018)	5	.009	Partial	PTF	2	3	USA	1	.66	1.15; 2.5	1	L	Income tax	Ind.
Persian et al. (2023)	5	.027	Partial	PTP	1	4	Indonesia	1	.49	9	0	D	Income tax	Ind.
Pfeifer and Pacheco (2020)	2	.005	Partial	PTP	1	2	Brazil	0	0	1	1	L	Property tax	Ind.
Pomeranz (2015)	12	.009	Partial	PTF	1	4	Chile	1	.16	4	0	L	VAT	Bus.
Santorio (2024)	1	.056	Full	PTP	2	1	Eswatini	1	.01	3	0	L	Income tax	Ind.; Bus.
Schächtele et al. (2022)	2	.016	Full	PTP	2	2	Brazil	1	.77	2	0	L	Property tax	Ind.
Schächtele et al. (2023)	1	.034	Partial	PTP	1	1	Argentina	1	.89	2	0	L	Property tax	Ind.
Torgler (2004)	2	.05	Partial	PTF	2	2	Switzerland	1	.46	1.15	1	L	Income tax	Ind.
Wenzel (2006)	2	.05	Partial	PTF	2	2	Australia	1	.46	1.15	1	L	VAT	Bus.

The table presents summary statistics for the main characteristics at level of individual studies. These characteristics are defined in detail in Section 2. The sample includes the 53 papers in our baseline sample (i.e. column (1) of Table A1).

Appendix B: Additional tables

Table B1: Test of asymmetry

	(1)
Inverse of standard error	0.006 (0.006)
Constant	2.909*** (1.074)
Observations	490
Adjusted R^2	0.015

Normalised coefficients are regressed on inverse standard errors and an intercept following Egger et al. (1997).

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

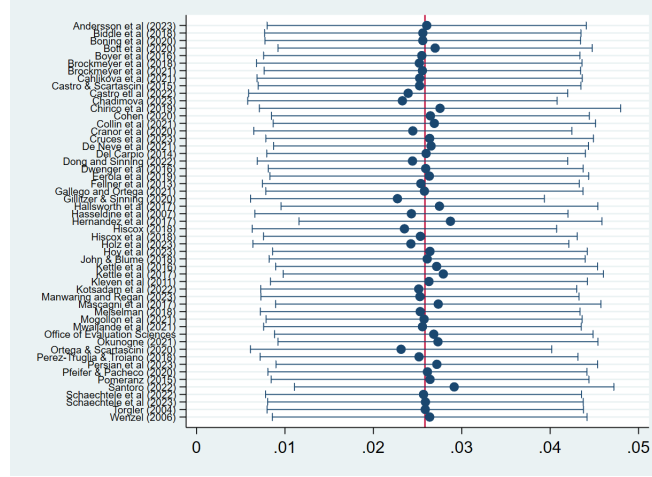
Table B2: P-hacking in t-values around significance threshold

	Normalised t-values			5% significance threshold		
	(1)	(2)	(3)	(4)	(5)	(6)
Bandwidth	0.1575	0.1	0.075	0.25	0.2	0.1
Number of tests in bandwidth	141	103	82	82	67	40
Panel A: Randomisation test of discontinuities in t-values						
Share significant	0.652	0.641	0.634	0.610	0.642	0.675
p-value (one-sided)	0.000	0.003	0.010	0.030	0.014	0.019
Panel B: Caliper tests						
Deterrence vs non-deterrence	0.039	-0.031	-0.040	0.057	-0.104	-0.193
p-value	0.589	0.715	0.710	0.617	0.299	0.080
Reminder vs no letter	-0.051	-0.106	-0.100	-0.041	-0.074	-0.006
p-value	0.501	0.380	0.516	0.715	0.582	0.969
Published vs working paper	0.064	0.172	0.195	-0.098	0.003	0.143
p-value	0.455	0.107	0.102	0.306	0.973	0.337
Main vs other estimate	-0.064	-0.157	-0.126	0.029	0.089	-0.184
p-value	0.277	0.066	0.249	0.760	0.373	0.110
New vs old paper	-0.106	-0.083	-0.066	0.041	0.064	0.191
p-value	0.137	0.402	0.620	0.736	0.648	0.199

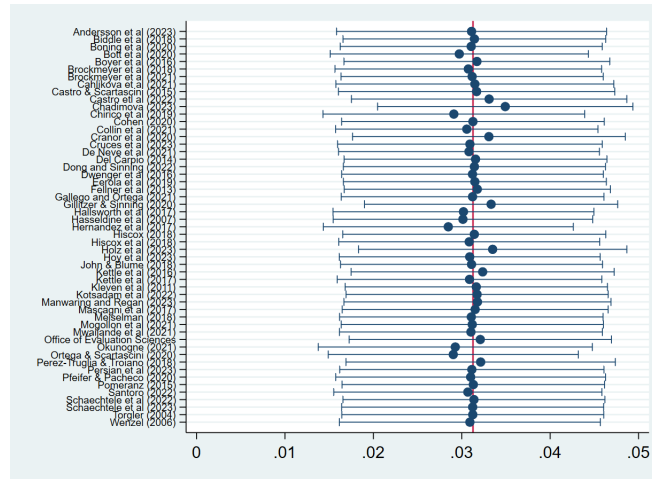
The table shows two exercises in Panels A and B on detecting p-hacking for three different bandwidths around normalised significance thresholds in columns (1) to (3) and around the 5% significance threshold in columns (4) to (6). Panel A tests whether the empirical distribution corresponds to a binomial distribution with equal probability below and above the significance threshold. Panel B presents marginal effects from a probit regression of a significance dummy on a set of control variables. Standard errors are clustered at the level of papers.

Figure B1: Jackknife exercise

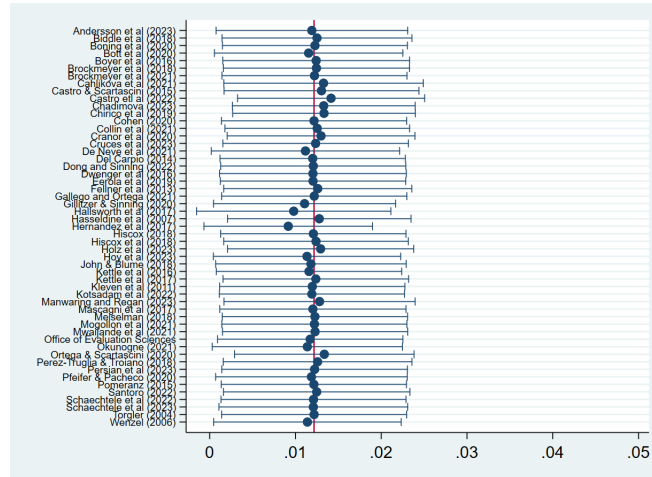
(a) Reminder, $\hat{\alpha}$



(b) Deterrence, $\hat{\beta}$



(c) NonDeterrence, $\hat{\gamma}$



Notes: The figures presents jackknife-type robustness tests by excluding papers one-by-one from the sample. The specification follows Equation 1. The exercise is performed for the three coefficients – $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ – separately in the three sub-figures. Horizontal lines represent 95% confidence intervals.