// NO.19-055 | 02/2024

DISCUSSION PAPER

// ARMENAK ANTINYAN AND ZAREH ASATRYAN

Nudging for Tax Compliance: A Meta-Analysis

Leibniz Association

ZEW

Nudging for Tax Compliance: A Meta-Analysis

Armenak Antinyan*† Zareh Asatryan*‡

First version: November, 2019

This version: February 2024

Abstract

Governments increasingly use nudges to improve tax collection. We synthesize the growing literature that evaluates nudging experiments using meta-analytical methods. We find that simple reminders increase the probability of compliance by 2.7 percentage points relative to the baseline where about a quarter of taxpayers are compliant. Nudges that commonly refer to elements of tax morale increase compliance by another 1.4 percentage points. Deterrence nudges, which inform taxpayers about enforcement parameters, increase compliance the most, amounting to an additional 3.2 percentage points increase on top of reminders. Our additional findings highlight some of the conditions where nudges are more effective, such as their potential when targeting sub-population of late-payers, and also suggest that even this sample of randomized trials may be susceptible to selective reporting of results. Overall, our findings imply that taxpayers are biased by various informational and behavioral constraints, and that nudges can be of some help in overcoming these frictions.

JEL codes: C93, D91, H26.

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

*We would like to thank Keith Marzilli Ericson, Annika Havlik, Jost Heckemeyer, Friedrich Heinemann, 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.

†Cardiff University and Zhongnan University of Economics and Law, antinyan.armenak@gmail.com.

‡ZEW Mannheim and CESifo Munich, zareh.asatryan@zew.de.

1 Introduction

Recent years have seen much excitement around the idea of using "nudges" with the aim of improving individual behavior. 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 effectively? iv) Are nudges also effective over a longer time horizon? v) Which groups of taxpayers are more responsive to nudges? vi) Do nudges work in specific settings (e.g., low-compliance environments) or more generally?

To answer these questions we collect data from up to 71 randomized control trials (RCTs). Our analysis starts with a synthesis of the 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 important contextual characteristics that define nudging interventions, and the customary measures of tax compliance used. 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. These results are robust to various alternative estimators and sample definitions.

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 our sample only about 25% of taxpayers not receiving any nudges are compliant on average, which is low but not surprisingly so, since about two-thirds of RCTs in our sample work with samples of taxpayers which were late in paying their taxes. Our estimates suggest that, compared to this baseline 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 baseline to about 31%.

These results are consistent with the idea that taxpayers are biased by various informational and behavioral 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) behavior 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 behavioral 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 behaviors, 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 relates to those of traditional policy tools. Despite the popular belief that the sending 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, while targeting late-payers, when communicated through in-person visits, and those 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 some suggestions for future research.

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 do 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 behavior 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

[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.

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, Behavior, and Development Unit of the World Bank, the Behavioral Insights Team) conduct such experiments.

In a typical nudging experiment, the agents are randomized into one or several treatment arms, which are exposed to 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 behavior 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 behavior, the length of the time horizon over which this behavior is studied as well as the delivery method of the nudge can vary across experiments. While some experiments may study one type of behavior (e.g., probability to pay) over one time horizon after the delivery of the nudge using a certain method (e.g., digital letters), others may study multiple types of behaviors (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 centered 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 role 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, not only do we study the question of whether the nudging interventions 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,[3] Luttmer and Singhal (2014) summarize 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.

[2]We are aware of two other meta-studies of tax experiments by Alm and Malézieux (2020), Blackwell (2007), 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 probability increases tax compliance on the extensive margin, while audit probability and the tax rate influence tax compliance negatively on the intensive margin.

[3]The nudges we study are arguably the most common types of behavioral interventions in taxation, but governments can nudge in other ways too. For example, policies that publicly recognize 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.

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. Below we discuss the 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 maximizers who compare the benefits of tax evasion against the costs of detection and punishment when deciding to comply 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

[4] Recent 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).

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 behavior may depend on the simple behavioral 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 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 utilized 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).

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

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 behavior 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 behavior: "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).

[6] As 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.

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 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, informs us about the share of compliant taxpayers in the control group at the end of the intervention.

[7] Few 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.

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 Baskaran et al. (2016), Card et al. (2010, 2017), Heinemann et al. (2018), Klomp and De Haan (2010), 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 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 stylized 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

Table 1: Stylized summary of the framework

(1) Treatment nudge	(2) Treatment effect estimates Control group: No letter	(3) Reminder	(4) Implied compliance effects In addition to c (% of population)	(5) In relation to c (% 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.

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 do this 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, i.e., a binary variable on whether the time interval between the date on which 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 mix of 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

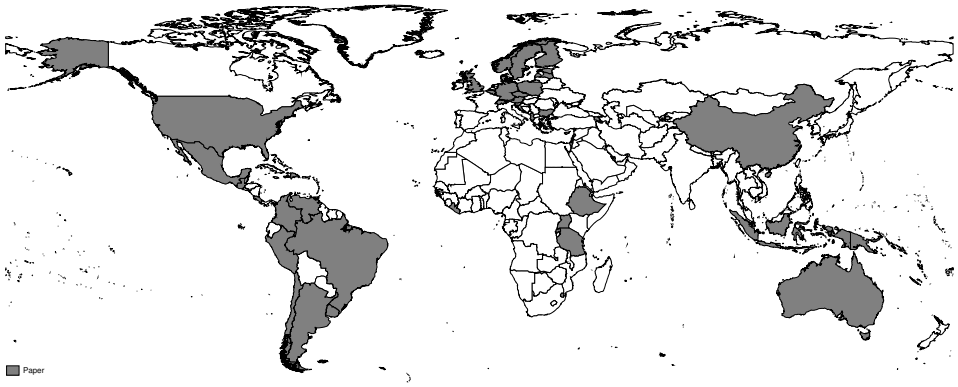
Literature search: We ran a literature search on a rolling basis throughout 2019 to 2023. First we searched for relevant papers using a defined combination of keywords in the main literature databases for the profession. Second, to identify ongoing work, we continued our search in the programs of the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

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

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

Figure 1: Country coverage of nudging experiments



Notes: The studies where these experiments come from are are listed in Table A1.

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 behavior of the taxpayer; and iv) the resulting study reports all of the relevant statistics necessary for our meta-analysis (e.g., effect sizes along with the standard errors) for at least one treatment effect estimate.

Final sample of studies: After applying the four filters to the list of papers collected from our extensive search we arrive at an overall sample of 71 studies. For analyzing the extensive margin of compliance we use 55 studies, while for the publication bias analysis and the robustness checks with t-values we use all 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 decided to perform our baseline analysis on the "main" estimates reported in studies, while using the full sample (i.e., main and non-main) of treatment effect estimates for robustness checks.[10]

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 utilizing 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 intention to treat effects (ITT) as the main estimate.[11] Fourth, when

[10] The only exception is that we use the full sample of treatment effect estimates in Section 6. Since in this section we are interested in the potential selection effects in reported estimates, we choose to use the full sample of estimates that papers report, rather than the main sample of estimates which is chosen by us.

[11] Studies in our sample either report ITT estimates only, or both ITT and ToT estimates. ITT estimates include every subject that is randomized according to randomized treatment assignment, disregarding non-compliance, protocol deviations, withdrawal, and anything that happens after randomization (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

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,[12] 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: After applying the filters defined above, we arrive at our baseline sample of 55 studies that contain main estimates measured at the extensive margin of compliance. Table A2 reports the number of observations with which each study contributes to the sample, along with several other important study-level characteristics. The number of estimates per study ranges between 1 estimate to a maximum of 28 estimates per study, and they are 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

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.

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

Table 2: Summary statistics

	(1) Extensive margin		(2) T-values	
Dependent variable				
Treatment effect (mean)	270	0.035	475	3.230
Control group				
Reminder vs no letter	270	123	475	188
Nudge type				
Deterrence	270	126	475	228
Public Good	270	39	475	78
Social norm	270	40	475	52
Moral Appeal	270	16	475	43
Other	270	49	475	74
Late-payer sample				
General Sample	270	115	475	255
Late	270	155	475	220
Response horizon				
Short run	270	138	475	226
Publication status				
Published	270	171	475	278
Year of publication (mean)	270	2015	475	2015
Delivery				
Letter	270	175	475	289
Digital	270	81	475	161
In Person	270	14	475	25
Tax type				
Income Tax	270	116	475	198
Corporate Tax	270	16	475	31
Property Tax	270	76	475	105
VAT	270	11	475	45
Other	270	33	475	72
Multiple	270	18	475	24
Taxpayer type				
Individual	270	169	475	244
Business	270	65	475	162
Individual and Business	270	36	475	69
Development level				
Low Income	270	27	475	46
Middle Income	270	108	475	195
High Income	270	135	475	234

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.

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".[13]

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 28 main estimates as shown in Table A2. These treatment effect estimates are collected from Table 2 of the study, and are available for two outcome variables (ever-paid, paid-in-full), seven treatment effects that include five nudge types, 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 55 studies. These are the compliance effects measured at the extensive margin and represent only the estimates that we consider to be the main ones. Table A1 presents the number of estimates coming from each of the studies. The summary statistics of all the variables we use are

[13]Time horizons are approximated for several studies since the exact timing of nudge dispatch and treatment effect measurement are not discussed.

presented in Table 2. Note that this baseline sample excludes the reminder nudges, r, which are available only for the experiments that do not treat the control group with any communication. Column "reminders" of Table A2 shows the studies where these treatment effects come from. We consider these again for the robustness analysis of the reminder effect in Section 5.3 which increases the sample to 298. Among those robustness exercises, we also use a sample consisting of t-values, rather than just the extensive margin estimates, which are in total 475 as shown in Table A1. Additionally, we have two further robustness samples consisting of both main and non-main estimates which, as shown in Table A1, have 581 and 968 observations, respectively, for the extensive margin and t-value compliance measures.

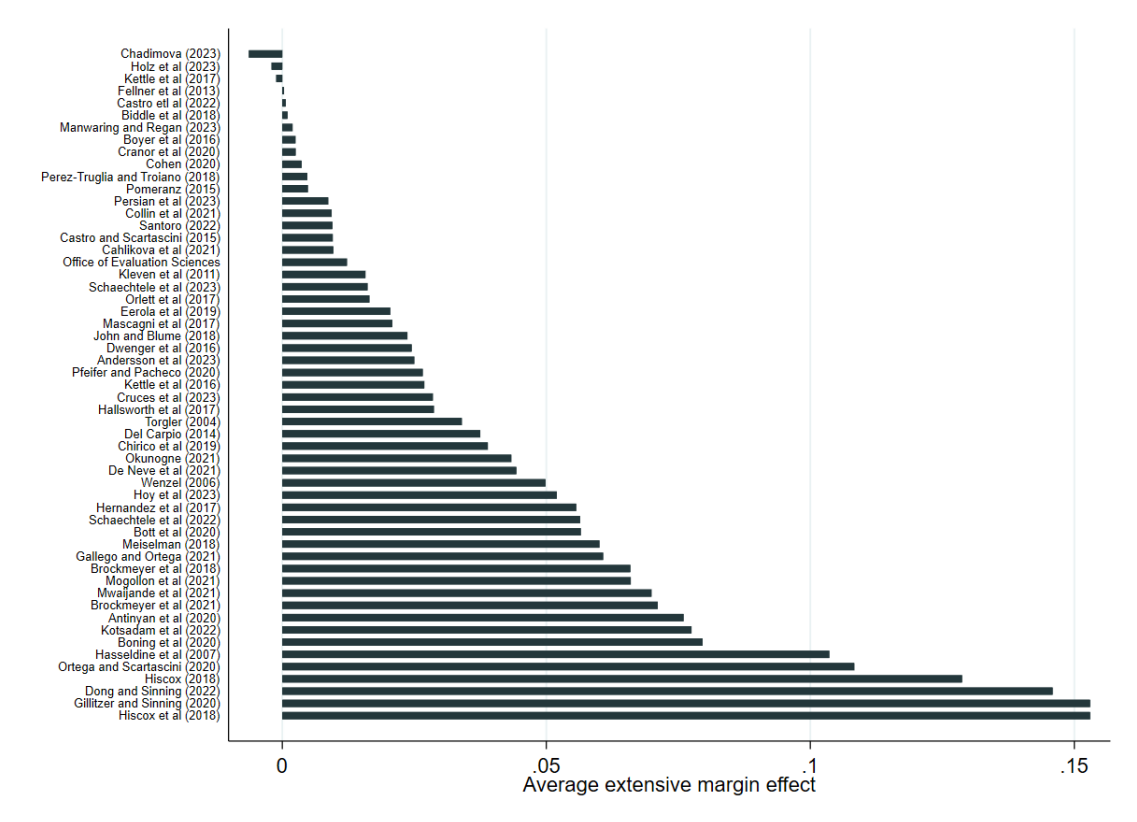
4 Empirical design

Baseline specification: We estimate the following equation:

$$\begin{aligned} ComplianceEffect_{i,p} &= \gamma[NonDet - vs - Reminder] + \\ &+ \alpha Reminder - vs - Control_p + \beta Det - vs - NonDet_{i,p} + \epsilon_{i,p} \end{aligned} \tag{1}$$

Dependent variable: $ComplianceEffect_{i,p}$ is the i^{th} estimate of extensive margin treatment effect from paper p. The distribution of this variable averaged for the 55 studies in our main sample is plotted in Figure 2. In all of our analyses, we winsorize the dependent variable at its 5th and 95th percentiles. This winsorization 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

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 specifications by paper in ascending order of average effect magnitude.

estimates.[14] 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. This data is available for a larger sample of 71 studies.[15]

[14] For example, Card et al. (2017) winsorize the data at the 10th and 90th percentiles. Table 6 provides a robustness exercise by running our baseline specifications on non-winsorized data.

[15] In instances where the standard errors of the estimates are either not reported or are reported as rounded to 0, we take advantage of the reported significance levels of these estimates (e.g., indicated by stars in the regression tables) and replace the missing values with the most conservative t-values that pass the respective significance threshold. When analyzing publication bias, values derived in such a way are excluded in order to avoid artificial bunching at critical significance thresholds.

Independent variables: $Reminder - vs - Control_p$ 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 roughly equal number of these two types of estimates. Consequently, α compares the treatment effect estimates across the two experimental designs fixing for the type of the nudge that was sent: if a non-deterrence nudge was sent $\hat{\alpha} = r + n - n = r$, and if a deterrence nudge was sent $\hat{\alpha} = r + d - d = r$. Since we do not have studies that utilize both types of experiments, the reminder effect is only identified from across study variation. In Section 5.3, we present an alternative 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 reminder nudge that is compared to that control group.

$Det - vs - NonDet_{i,p}$ 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 but not including reminder nudges. We have about as many deterrence as non-deterrence nudges in the sample. Consequently, β compares the treatment effect estimates of deterrence and non-deterrence nudges within the two experimental designs: if the control group is not treated with any communication $\hat{\beta} = (r + d) - (r + n) = d - n$, and if the control group receives a reminder letter $\hat{\beta} = d - n$. 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.

Finally, after setting the $Reminder - vs - Control$ and $Det - vs - NonDet$ dummies to be zero, it is straightforward to see that $\hat{\gamma} = n$, that is the intercept γ captures the effect of non-deterrence type nudges compared to a control group of

taxpayers receiving reminders. Note, again, that this is possible, because our baseline sample excludes reminder nudges, r. 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.

Coefficients of interest: Our three main effects of interest are as follows: i) The effect of reminders, r, is identified by $\hat{\alpha}$, ii) the effect of reminder and non-deterrence nudges, $r + n$, is identified by $\hat{\alpha} + \hat{\gamma}$, and iii) the effect of reminder and deterrence nudges, $r + d$, is identified by $\hat{\alpha} + \hat{\gamma} + \hat{\beta}$.

Statistical significance: To show the statistical significance of the effect of reminders compared to the control group, r, we test the null hypothesis $\hat{\alpha} = 0$. The difference between r and $r + n$ is tested by $\hat{\gamma} = 0$ (since this expression is the same as $\hat{\alpha} = \hat{\alpha} + \hat{\gamma}$). While the difference between r and $r + d$ is tested by performing a post-estimation test for $\hat{\beta} + \hat{\gamma} = 0$ (which is the same as $\hat{\alpha} = \hat{\alpha} + \hat{\gamma} + \hat{\beta}$). Finally, the differences between $r + n$ and $r + d$ is tested by $\hat{\beta} = 0$ (which is the same as $\hat{\alpha} + \hat{\gamma} = \hat{\alpha} + \hat{\gamma} + \hat{\beta}$).

Error term: $\epsilon_{i,p}$ is the error term. Since the 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. Columns (1) run the regression for the complete sample of 270 estimates, and column (2) repeats the analysis for the sub-sample of 218 estimates where the underlying compliance level, c, is available. 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, the intercept $\hat{\gamma}$ suggests that non-deterrence nudges increase the probability of compliance by 1.2 percentage points compared to the control group receiving reminder letters. And third, $\hat{\beta}$ suggests that deterrence nudges increase the probability of compliance by 1.9 percentage points compared to non-deterrence nudges. All of these effects are statistically distinguishable from zero at least at the 5% level. The results from the reduced sample plotted in column (2) are very similar to those from the complete sample of column (1).

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 25.0%. 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 about two-thirds of RCTs in our sample work with samples of late-payers where compliance is close to

Table 3: Baseline results

	(1) Main estimates	(2) Compliance observed
Reminder-vs-Control $\hat{\alpha}$	0.026***	0.027***
	(0.009)	(0.010)
NonDet-vs-Reminder $\hat{\gamma}$	0.012**	0.014**
	(0.005)	(0.006)
Det-vs-NonDet $\hat{\beta}$	0.019**	0.018**
	(0.007)	(0.008)
Observations	270	218
Adjusted R^2	0.162	0.160
Postestimation p-values:		
Deterrence + Reminder = 0	0.000	0.000

Regressions are estimated according to Equation 1. Column (2) restricts the sample to observations where the underlying compliance level, c, is observed.
Standard errors in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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 summarizes 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. Column (4) then shows the implied compliance effects in addition to the underlying compliance level in percent of population, while column (5) shows the implied compliance effects relative to the underlying compliance level. Compared to the share of compliant taxpayers of 25.0%, 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

Table 4: Effect magnitudes

(1) Treatment effect	(2) Estimator	(3) Estimated effects Table 3 col. 2 (percentage points)	(4) Implied compliance effects In addition to c (% of population)	(5) 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	$r + d = \hat{\alpha} + \hat{\gamma} + \hat{\beta}$	=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 implied compliance effects in addition and in relation to the underlying compliance level, c.

baseline. These three estimates are statistically distinguishable from zero as well as from each other.[16]

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 in the data and compare their effects to the control group of untreated taxpayers. The specification is as follows:

$$ComplianceEffect_{i,p} = \alpha_1[Reminder - vs - Control] + \\ + \beta_1 Det - vs - Reminder_{i,p} + \gamma_1 NonDet - vs - Reminder_{i,p} + \epsilon^1_{i,p} \quad (2)$$

where β_1 and γ_1 identify the effects of deterrence and non-deterrence nudges compared to reminder nudges, which is the omitted category, that is d and n, respectively.

[16] This follows Table 3, which rejects all of the following four null hypotheses: $\hat{\alpha} = 0$, $\hat{\beta} = 0$, $\hat{\gamma} = 0$, and $\hat{\beta} + \hat{\gamma} = 0$. Section 4 explains that these are the only hypotheses we need to test.

Consequently, the intercept α_1 identifies the effect of reminder nudges compared to the untreated control group, that is r. 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 very similar to the baseline results on the effects of reminders presented in Table 3.[17]

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:

$$ComplianceEffect_{i,p} = \beta_2 Det - vs - NonDet_{i,p} + \lambda_p + \epsilon^2_{i,p} \tag{3}$$

where λ_p is the new term capturing the study fixed effects. β_2 is the main coefficient of interest which, as before, shows the effect of deterrence nudges on tax compliance compared to that of non-deterrence nudges, but now identified within papers. 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 presented in Table 3.

[17] The effects of non-deterrence nudges compared to reminders of 1.1 percentage points is again similar in magnitude to the baseline results, however, in this sub-sample, the effect is statistically not distinguishable from zero. Note, that the effects of deterrence nudges here are compared to that of reminders, rather than to non-deterrence nudges as in the baseline. This 3.8 percentage point effect is somewhat larger, but generally consistent with the baseline magnitudes presented in Table 3 (i.e. 1.2+1.9=3.1).

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

	(1) Reminder	(2) Deterrence	(3) Non-Deterrence
Reminder-vs-Control	0.022***		
	(0.004)		
NonDet-vs-Reminder	0.011		
	(0.009)		
Det-vs-Reminder	0.038***		
	(0.009)		
Det-vs-NonDet		0.024***	
		(0.006)	
Non-deterrence effect of type $\hat{\gamma}^{\kappa}$			
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
Postestimation p-values:			
Public Good = Social Norm			0.175
Public Good = Moral Appeal			0.922
Public Good = Other			0.098
Social Norm = Moral Appeal			0.294
Social Norm = Other			0.283
Moral Appeal = Other			0.158

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. To do so, we augment Equation 1 and redefine $Deterrence_{i,p}$ as a categorical variable consisting of deterrence nudges as before, but, following our definitions of Section 2.3, divide non-deterrence ones into public good, social norm, moral appeal and other nudges, in addition to reminder nudges. 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.

$$ComplianceEffect_{i,p} = \gamma_2^{\kappa} \sum_{\kappa=0}^{4} NonDet - vs - Det_{i,p} + \lambda_p + \epsilon_{i,p}^2 \qquad (4)$$

Since we are interested in the effects of these types of non-deterrence nudges, here we take deterrence nudges, $\kappa = 0$, as the omitted category. Consequently, the effects of the four types of non-deterrence nudges are captured by γ_2^{κ}. As shown in Table 2 public good, social norm and moral appeal are about as numerous as deterrence nudges.[18] Additionally, a tenth of the nudges are grouped into the category of other nudges. 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

[18] Several interventions mix different types of non-deterrence nudges. In particular, 16, 12 and 16 messages mix, respectively, public good and social norm, social norm and moral appeal, and public good and moral appeal 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.

– 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 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 heteroskedastic.[19] As analytical weights we take the inverse of the squared standard error of the parameter estimates, which, unlike the OLS approach, yields to precision-weighed 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.

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-winsorized 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

[19] One 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.

Table 6: Results using alternative estimation methods

	(1) WLS (s.e.)	(2) Random Effects	(3) Non-winsorized	(4) Non-negative	(5) <90th %tile	(6) T-value	(7) Pr. to pay	(8) Pr. to file	(9) All specifications
Reminder	0.017*	0.023*	0.029***	0.020**	0.030**	1.451	0.029**	0.016	0.031***
vs Control	(0.009)	(0.013)	(0.010)	(0.009)	(0.012)	(0.934)	(0.012)	(0.026)	(0.011)
NonDet	0.010*	0.015	0.012**	0.021***	0.016**	1.528**	0.015**	0.013	0.010
vs Reminder	(0.005)	(0.010)	(0.006)	(0.005)	(0.007)	(0.690)	(0.007)	(0.021)	(0.006)
Det	0.014**	0.021*	0.017*	0.019**	0.013	1.719***	0.021*	0.023	0.018*
vs NonDet	(0.005)	(0.012)	(0.009)	(0.008)	(0.009)	(0.593)	(0.011)	(0.016)	(0.010)
Obs.	257	257	270	228	185	475	157	58	535
R^2	0.087		0.146	0.127	0.146	0.093	0.225	0.078	0.191

Regressions are estimated according to Equation 1. Column 1 estimates effects with weighted least squares, where weights are inverse of squared standard errors. Column 2 estimates effect with a Random Effects model. Column 3 takes raw estimates, that are not winsorized, as the outcome variable. Column 4 excludes negative estimates. 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 focuses on the subsample where the outcome is probability to pay. Column 8 focuses on probability to file. Column 9 includes non-main specification estimates. Standard errors in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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 use heterogenous measurements of compliance, 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 vari-

able of interest. The main benefit of this exercise is that t-values are available for all estimates increasing our baseline sample from 270 to 475 estimates. This approach is not uncommon in the field of meta-analysis, and is used by many applications in economics sometimes even as their primary outcome variable of interest, such as by Baskaran et al. (2016), Card et al. (2010, 2017), Heinemann et al. (2018), Klomp and De Haan (2010). 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 dissimilar 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 the full sample of treatment effect estimates. Last two columns of Table A2 show the list of papers and the available treatment effects per paper in this larger sample, and Table 2 presents the summary statistics of this sample. 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 that we have identified in explaining the heterogeneity in treatment effect estimates. To do so, we introduce vector $\mathbf{X_{i,p}}$ of these study characteristics to Equation 1 on top of the baseline regressors.

$$ComplianceEffects_{i,p} = \gamma[NonDet - vs - Reminder] +$$
$$+ \alpha Reminder - vs - Control_p + \beta Det - vs - NonDet_{i,p} + \delta \mathbf{X_{i,p}} + \epsilon_{\mathbf{i,p}} \quad (5)$$

The study characteristics enter the regression first one-by-one and then jointly.[20] These seven characteristics are defined in Section 2.6. Table 2 shows that in our sample about two-thirds 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 about three-fifth of the estimates come from papers that have been published already; that the mean study year is 2015; 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, corporate income taxpayers and fewer VAT taxpayers; that about a third of our sample interventions target businesses; 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] This 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.

Table 7: Role of study characteritsics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Reminder-vs-Control	0.028***	0.029***	0.026***	0.027***	0.025***	0.023***	0.030***	0.027***
	(0.008)	(0.009)	(0.008)	(0.009)	(0.008)	(0.009)	(0.009)	(0.008)
NonDet-vs-Reminder	0.024***	0.016**	3.621	0.012*	0.045	0.012**	0.019***	2.425
	(0.007)	(0.006)	(2.182)	(0.006)	(0.029)	(0.005)	(0.006)	(2.503)
Det-vs-NonDet	0.016*	0.022***	0.019***	0.017**	0.021***	0.020***	0.022***	0.019***
	(0.008)	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)	(0.008)	(0.006)
Late-payer sample (omitted: Late)								
General Sample	-0.027***							-0.017**
	(0.008)							(0.008)
Response Horizon (omitted: Short Run)								
Long Run		-0.014*						-0.012
		(0.009)						(0.007)
Publication Status (omitted: Published)								
Unpublished			0.001					0.004
			(0.010)					(0.009)
Year			-0.002					-0.001
			(0.001)					(0.001)
Delivery (omitted: Physical Letter)								
Digital Letter				-0.004				0.018*
				(0.009)				(0.009)
In Person				0.026*				0.057***
				(0.015)				(0.012)
Tax Type (omitted: VAT)								
Income Tax					-0.032			-0.008
					(0.031)			(0.027)
Corporate Tax					-0.064**			-0.035
					(0.030)			(0.026)
Property Tax					-0.034			-0.005
					(0.030)			(0.029)
Other					-0.035			-0.020
					(0.031)			(0.027)
Multiple					-0.017			-0.010
					(0.041)			(0.029)
Taxpayer Type (omitted: Individual)								
Business						0.012		0.024**
						(0.012)		(0.011)
Individual and Business						-0.008		-0.003
						(0.011)		(0.010)
Development Level (omitted: High Income)								
Low Income							-0.023*	-0.043***
							(0.013)	(0.015)
Middle Income							-0.019*	-0.022***
							(0.010)	(0.008)
Observations	270	270	270	270	270	270	270	270
Adjusted R^2	0.255	0.185	0.181	0.176	0.213	0.177	0.207	0.387

Regressions are estimated according to Equation 5.

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

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 22 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 at the extensive margin. Second, the treatment effects are stronger in the 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, we ask whether the delivery method of the experiment matters for compliance. 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

[21] This 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 16.2% and in Table 7, column (8) it is 38.7%.

done by designing interventions that go beyond short horizons, tax bases and country borders, among other dimensions.

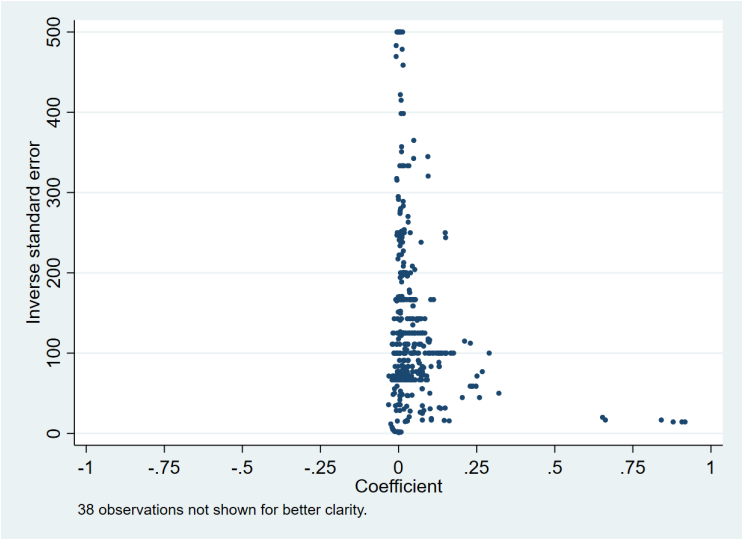
6 Publication selection bias

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 favorably, for example, due to the belief that non-significant effects are harder to publish (p-hacking).

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 effects as a measure of precision. We observe that the imprecisely estimated treatment effects, i.e., those at

Figure 3: Funnel plot

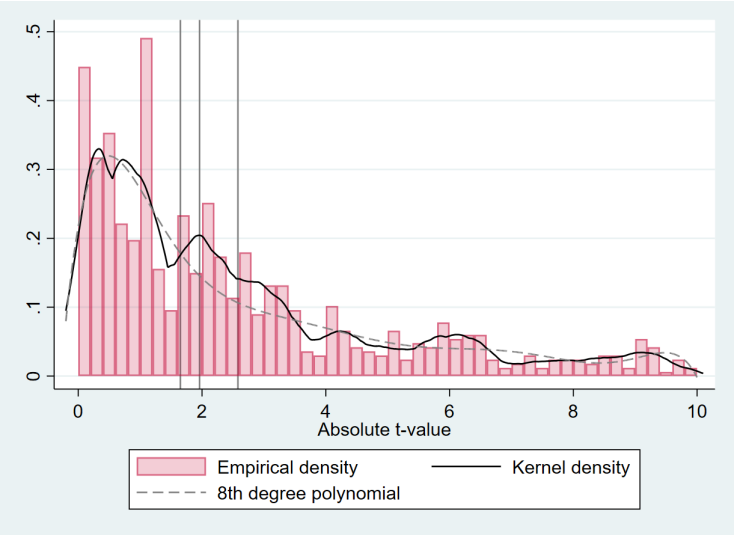


Notes: For visual clarity, the figure drops outliers larger than 500 on the y-axis, arriving at a sample of 495 estimates.

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

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 676 observations. Vertical lines denote critical values for two-sided significance tests at t-values of 1.645, 1.96 and 2.58.

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 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 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 randomization 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 centered 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 centering 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 Brazilian mayors about research into the positive tax compliance effects of reminder letters in an experimental setting 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 summarize 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. Our 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, or between periods in case these decisions are substitutes, we are not aware of any 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, and also allow studying the sources behind taxpayers' heterogeneous responses. 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 (e.g., sending letters) 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. Importantly, we do not have much knowledge of how interventions interact with the context they operate in. This is not surprising given that randomized 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 try to study whether 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. Technical report, 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.

Bérgolo, M. L., R. Ceni, G. Cruces, M. Giaccobasso, and R. Perez-Truglia (2017, July). Tax audits as scarecrows: Evidence from a large-scale field experiment. Working Paper Series 23631, National Bureau of Economic Research.

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. Technical report, 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.

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. 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. Technical report, Working Paper.

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.

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.

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. Pfrang, and J. Schmitz (2022). How to improve small firms' payroll tax compliance? evidence from a randomized field experiment.

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. Fooken, 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. Technical report, 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. Technical report, 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. *Bmj 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 J. Slemrod (2019). Missing miles: Evasion responses to car taxes. *Journal of Public Economics, forthcoming*.

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. Korczyc, 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), 000–000.

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.

Kettle, S., M. Hernandez, S. Ruda, and M. Sanders (2016). Behavioral interventions in tax compliance: Evidence from Guatemala. Policy research working papers 7690, The World Bank.

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.

Klomp, J. and J. De Haan (2010). Inflation and central bank independence: A meta-regression analysis. *Journal of Economic Surveys 24*(4), 593–621.

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? *mimeo*.

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.

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.

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. ICTD Working Paper 58, The International Centre for Tax and Development.

Meiselman, B. (2018). Ghostbusting in detroit: Evidence on nonfilers from a controlled field experiment. *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. *mimeo*.

Okunogbe, O. (2021). *Becoming legible to the state: the role of detection and enforcement capacity in tax compliance*. The World Bank.

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. Technical report, IRS, Working Paper.

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 Latinamerica.

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.

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. Technical report, 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.

Tortarolo, D., G. Cruces, and G. Vazquez-Bare (2023). Design of partial population experiments with an application to spillovers in tax compliance. Technical report, Institute for Fiscal Studies.

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.

Online Appendix

Appendix A: Sample of studies

Table A1: Sample of studies

No.	Paper	Country	Main specifications		All specifications	
			Extensive	T-value	Extensive	T-value
1	Andersson et al. (2023)	Sweden	5	5	10	10
2	Antinyan and Asatryan (2020)	Armenia	0	0	3	9
3	Antinyan et al. (2021)	China	0	0	2	2
4	Ariel (2012)	Israel	0	8	0	8
5	Bérgolo et al. (2017)	Uruguay	0	6	0	9
6	Biddle et al. (2018)	Australia	3	6	3	6
7	Blumenthal et al. (2001)	USA	0	4	0	4
8	Boning et al. (2020)	USA	2	6	8	24
9	Bott et al. (2020)	Norway	3	6	12	24
10	Boyer et al. (2016)	Germany	2	4	2	4
11	Brockmeyer et al. (2019)	Costa Rica	12	20	24	32
12	Brockmeyer et al. (2021)	Mexico	2	4	4	8
13	Cahlikova et al. (2021)	Czech Republic	12	12	12	12
14	Castro and Scartascini (2015)	Argentina	9	9	18	18
15	Scartascini and Castro (2019)	Argentina	0	3	0	9
16	Castro et al. (2022)	Peru	12	12	48	48
17	Chadimova (Chadimova)	Czech Republic	6	6	18	18
18	Chirico et al. (2019)	USA	24	36	42	63
19	Cohen (2020)	Uganda	2	4	6	12
20	Coleman (1996)	USA	0	8	0	8
21	Collin et al. (2021)	Tanzania	4	8	12	24
22	Cranor et al. (2020)	USA	6	6	15	15
23	Tortarolo et al. (2023)	Argentina	6	6	6	6
24	De Neve et al. (2021)	Belgium	3	3	5	5
25	Del Carpio (2013)	Peru	3	3	8	8
26	Doerrenberg and Schmitz (2017)	Slovenia	0	2	0	4
27	Doerrenberg et al. (2022)	Bulgaria	0	18	0	18
28	Dong and Sinning (2022)	Australia	2	2	4	4
29	Dwenger et al. (2016)	Germany	1	2	1	2
30	Schächtele et al. (2023)	Finland	3	9	8	24
31	Fellner et al. (2013)	Austria	3	3	3	3
32	Gallego and Ortega (2022)	Venezuela	3	6	6	12
33	Gillitzer and Sinning (2020)	Australia	3	6	6	12
34	Hallsworth et al. (2017)	UK	13	13	33	33
35	Harju et al. (2019)	Finland	0	2	0	2
36	Hasseldine et al. (2007)	UK	5	9	5	9
37	Hernandez et al. (2017)	Poland	9	27	18	54

38	Hiscox (2018)	Australia	4	6	4	6
39	Hiscox et al. (2018)	Australia	1	2	1	2
40	Holz et al. (2023)	Dominican Republic	5	10	5	10
41	Hoy et al. (2024)	Papua New Guinea	2	4	2	4
42	John and Blume (2018)	UK	3	2	3	2
43	Karver et al. (2022)	Albania	0	4	0	16
44	Kettle et al. (2016)	Guatemala	8	16	40	80
45	Kettle et al. (2017)	Guatemala	6	30	6	30
46	Kleven et al. (2011)	Denmark	2	4	2	4
47	Kotsadam et al. (2022)	Norway	2	4	2	4
48	Manwaring and Regan (2023)	Uganda	5	10	10	20
49	Mascagni et al. (2017)	Rwanda	8	10	24	31
50	Meiselman (2018)	USA	5	5	12	12
51	Mogollón et al. (2021)	Colombia	2	3	8	12
52	Mwaijande et al. (2021)	Tanzania	4	4	4	4
53	Office of Evaluation Sciences (2022)	USA	4	6	4	6
54	Okunogbe (2021)	Liberia	8	12	12	20
55	Orlett et al. (2017)	USA	0	0	8	7
56	Ortega and Sanguinetti (2013)	Venezuela	0	4	0	12
57	Ortega and Scartascini (2020)	Colombia	9	12	36	48
58	Perez-Truglia and Troiano (2018)	USA	4	0	4	0
59	Persian et al. (2023)	Indonesia	5	5	5	5
60	Pfeifer and Pacheco (2020)	Brazil	5	5	5	5
61	Pomeranz (2015)	Chile	2	4	2	4
62	Santoro (2024)	Eswatini	12	12	24	24
63	Saulitis and Chapkovski (2023)	Latvia	0	1	0	8
64	Schächtele et al. (2022)	Brazil	1	1	2	2
65	Schächtele et al. (2023)	Argentina	2	2	10	10
66	Shimeles et al. (2017)	Ethiopia	0	2	0	4
67	Slemrod et al. (2001)	USA	0	3	0	3
68	Torgler (2004)	Switzerland	1	1	2	2
69	Vainre et al. (2020)	Estonia	0	2	0	2
70	Wenzel and Taylor (2004)	Australia	0	3	0	4
71	Wenzel (2006)	Australia	2	2	2	2
			270	475	581	968

Table A2: Study characteristics: Extensive margin main specifications

Paper	Estimates	Reminders	Avg. estimate	Full compliance	Compliance measure	Comparison group	No. nudge types	Country	Published	Avg. compliance	Time horizons	Share late payer	Delivery methods	Taxes	Tax payer
Andersson et al. (2023)	5	0	0.025		PTP	Baseline letter	2	Sweden	1	0.66	1; 3	1.00	D	Income tax	Ind.
Antinyan et al. (2021)	1	1	0.076	Full	PTP	No letter	1	China	0	0.04	2	1.00	D	Property tax	Ind.
Biddle et al. (2018)	3	0	0.001	Partial	Other	Baseline letter	3	Australia	1		3	0.00	L	VAT	Bus.
Boning et al. (2020)	2	0	0.080	Partial	PTP	No letter	1	USA	1	0.58	3	0.00	L; P	Other	Bus.
Bott et al. (2020)	3	0	0.057	Partial	PTR	Baseline letter	3	Norway	1	0.20	.75	0.00	L	Income tax	Ind.
Boyer et al. (2016)	2	0	0.002	Partial	PTP	Baseline letter	2	Germany	1	0.02	4.6	0.00	L	Other	Ind.
Brockmeyer et al. (2019)	12	0	0.066	Partial	PTR; PTF; PTP	No letter	1	Costa Rica	1	0.03	3.75	1.00	D	Income tax	Bus.
Brockmeyer et al. (2021)	2	0	0.071	Partial	PTP	No letter	2	Mexico	0	0.06	1.3	1.00	L	Property tax	Ind.
Cahlikova et al. (2021)	12	0	0.010	Full	PTR	Baseline letter	4	Czech Republic	0	0.18	1.3; 3; 3.8	1.00	L	Other	Ind.
Castro and Scartascini (2015)	9	0	0.010	Full	PTP	Baseline letter	3	Argentina	1		.33; 1; 2	0.00	L	Property tax	Ind.
Scartascini and Castro (2019)	12	0	0.001	Partial	PTP	Baseline letter	3	Peru	1	0.04	.5; 5.5; 12.5	0.00	D	Income tax	Ind.
Chadimova (Chadimova)	6	0	-0.006	Full	PTR	Baseline letter	1	Czech Republic	0	0.18	1; 2	1.00	L	Other	Ind.
Chirico et al. (2019)	28	4	0.039	Full; Partial	PTP	No letter	5	USA	1	0.37	1; 3	1.00	L	Property tax	Ind.
Cohen (2020)	3	1	0.004	Partial	PTP	No letter	3	Uganda	0	0.05	3.1	0.00	D	Income tax	Bus.
Collin et al. (2021)	6	2	0.009	Partial	PTP	No letter	3	Tanzania	0	0.09	.33; 1.33	0.00	D	Property tax	Ind.
Cranor et al. (2020)	6	0	0.003	Full; Partial	PTP	Baseline letter	2	USA	1	0.19	1	1.00	L	Income tax	Ind.
Tortarolo et al. (2023)	6	0	0.029		PTP	No letter	1	Argentina	0	0.20	.15; 1	0.50	L	Property tax	Bus.
De Neve et al. (2021)	3	0	0.044	Partial	PTF; PTP	Baseline letter	1	Belgium	1	0.50	.46; .7; 2	0.67	L	Income tax	Ind.
Del Carpio (2013)	4	1	0.038	Partial	PTP	No letter	3	Peru	0	0.29	1	1.00	L	Property tax	Ind.
Dong and Sinning (2022)	2	0	0.146	Partial	PTR	No letter	2	Australia	1	0.04	1.5	1.00	L	Income tax	Ind.
Dwenger et al. (2016)	1	0	0.024	Partial	PTP	Baseline letter	1	Germany	1	0.21	5	0.00	L	Other	Ind.
Schächtele et al. (2023)	4	1	0.020	Partial	PTR	No letter	2	Finland	0	0.75	1	0.00	L	Income tax	Ind.
Fellner et al. (2013)	3	0	0.000	Partial	Other	Baseline letter	3	Austria	1	0.07	1.6	0.00	L	Other	Ind.
Gallego and Ortega (2022)	3	0	0.061	Full	PTP	No letter	1	Venezuela	1	0.02	1.5	1.00	D	Other	Ind.; Bus.
Gillitzer and Sinning (2020)	3	0	0.153	Partial	PTP	No letter	1	Australia	1	0.50	.79; 1.05; 1.28	1.00	L	VAT	Bus.
Hallsworth et al. (2017)	14	1	0.029	Partial	PTP	Baseline letter	4	UK	1	0.34	.75	1.00	L	Income tax	Ind.
Hasseldine et al. (2007)	5	0	0.104	Partial	PTR	No letter	3	UK	1	0.40		0.00	L	Income tax	Ind.
Hernandez et al. (2017)	9	0	0.056	Partial	PTP	Baseline letter	5	Poland	0	0.40	1	1.00	L	Income tax	Ind.
Hiscox (2018)	4	0	0.129		PTF; PTP	No letter	2	Australia	0	0.29	.5; .7	1.00	D	Multiple	Bus.
Hiscox et al. (2018)	1	0	0.153	Partial	Other	No letter	1	Australia	0	0.01	2	0.00	L	Income tax	Ind.
Holz et al. (2023)	5	0	-0.002	Partial	PTF	Baseline letter	2	Dominican Republic	1	0.51	4	0.00	D	Corporate tax	Bus.
Hoy et al. (2024)	2	0	0.052	Partial	PTF	Baseline letter	1	Papua New Guinea	1	0.21	.467	1.00	D	VAT	Bus.
John and Blume (2018)	3	0	0.024	Full	PTP	Baseline letter	2	UK	1	0.43	1	0.00	L	Property tax	Ind.
Kettle et al. (2016)	10	2	0.027	Partial	PTF; PTP	No letter	2	Guatemala	0		2.5	1.00	L	Income tax	Ind.; Bus.
Kettle et al. (2017)	6	0	-0.001	Partial	PTF	No letter	3	Guatemala	1		0	0.00	D	Multiple	Ind.; Bus.
Kleven et al. (2011)	2	0	0.016	Partial	Other	No letter	1	Denmark	1	0.14	1	0.00	L	Income tax	Ind.
Kotsadam et al. (2022)	2	0	0.078	Partial	Other	No letter	1	Norway	0		4; 10	0.00	D	Income tax	Ind.
Manwaring and Regan (2023)	5	0	0.002	Partial	PTP	Baseline letter	4	Uganda	0	0.04	1.4	0.00	D	Property tax	Ind.; Bus.
Mascagni et al. (2017)	12	4	0.021	Partial	PTF	No letter	3	Rwanda	0		2	0.00	D; P	Income tax	Bus.
Meiselman (2018)	6	1	0.060	Partial	PTF	No letter	3	USA	1	0.00	2.4	1.00	L	Income tax	Ind.
Mogollón et al. (2021)	2	0	0.066	Full; Partial	PTP	No letter	1	Colombia	1	0.05	2	1.00	L	Multiple	Ind.; Bus.
Mwaijande et al. (2021)	4	0	0.070	Partial	PTP	No letter	1	Tanzania	0	0.04	.47	0.00	D	Property tax	Ind.; Bus.
Office of Evaluation Sciences (2022)	4	0	0.012	Partial	Other	No letter	1	USA	0	0.29	4	0.00	L	Income tax	Ind.
Okunogbe (2021)	8	0	0.043		PTR; PTP; Other	Baseline letter	2	Liberia	0	0.08	6	1.00	P	Property tax	Ind.; Bus.
Orlett et al. (2017)	8	8	0.016	Partial	PTF	No letter	1	USA	0	0.60	1.5	1.00	L	Income tax	Ind.
Ortega and Scartascini (2020)	9	0	0.108	Full; Partial	PTP	No letter	1	Colombia	1		3; 4	1.00	L; D; P	Other	Bus.
Perez-Truglia and Troiano (2018)	4	0	0.005	Partial	Other	Baseline letter	2	USA	1		1.15; 2.5	1.00	L	Income tax	Ind.
Persian et al. (2023)	5	0	0.009	Partial	PTF	No letter	3	Indonesia	1	0.66	9	0.00	D	Income tax	Ind.
Pfeifer and Pacheco (2020)	5	0	0.027		PTP	Baseline letter	4	Brazil	0	0.49	1	1.00	L	Property tax	Ind.
Pomeranz (2015)	2	0	0.005	Partial	PTP	No letter	2	Chile	1		4	0.00	L	VAT	Bus.
Santoro (2024)	12	0	0.009	Partial	PTF	No letter	4	Eswatini	1	0.16	3	0.00	L	Income tax	Ind.; Bus.
Schächtele et al. (2022)	1	0	0.056		PTR	No letter	1	Brazil	1	0.01	1	1.00	D	Property tax	Ind.
Schächtele et al. (2023)	2	0	0.016	Full	PTP	Baseline letter	2	Argentina	1	0.77	2	0.00	L	Property tax	Ind.
Torgler (2004)	1	0	0.034	Partial	PTP	No letter	1	Switzerland	1	0.89	2	0.00	L	Income tax	Ind.
Wenzel (2006)	2	0	0.050	Partial	PTF	Baseline letter	2	Australia	1	0.46	1.15	1.00	L	VAT	Bus.

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

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

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

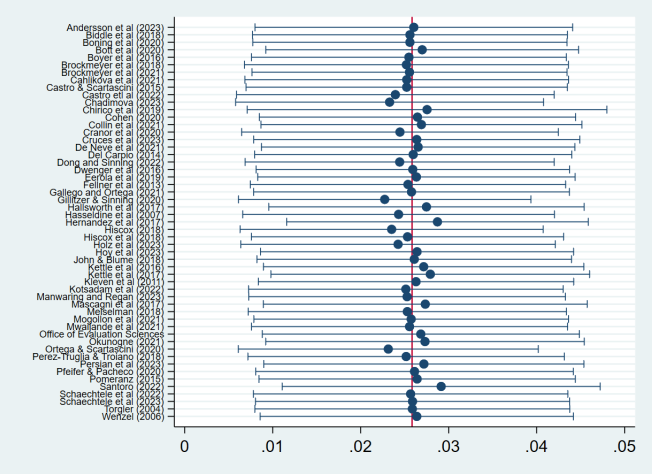
Table B2: P-hacking in normalized t-values around common significance thresholds and around 5% significance threshold

	Normalized 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: Randomization 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	0.007	-0.062	-0.084	0.071	-0.077	-0.160
p-value	0.903	0.392	0.348	0.502	0.403	0.103
Reminder vs no letter	-0.087	-0.154	-0.179	-0.066	-0.101	-0.052
p-value	0.119	0.141	0.229	0.559	0.414	0.669
Published vs working paper	0.056	0.161	0.179	-0.113	-0.015	0.118
p-value	0.481	0.117	0.125	0.263	0.861	0.278
Main vs other estimate	-0.112	-0.219	-0.165	0.019	0.054	-0.121
p-value	0.029	0.008	0.097	0.855	0.547	0.195
New vs old paper	-0.197	-0.176	-0.193	-0.037	-0.043	0.041
p-value	0.001	0.040	0.106	0.742	0.696	0.691

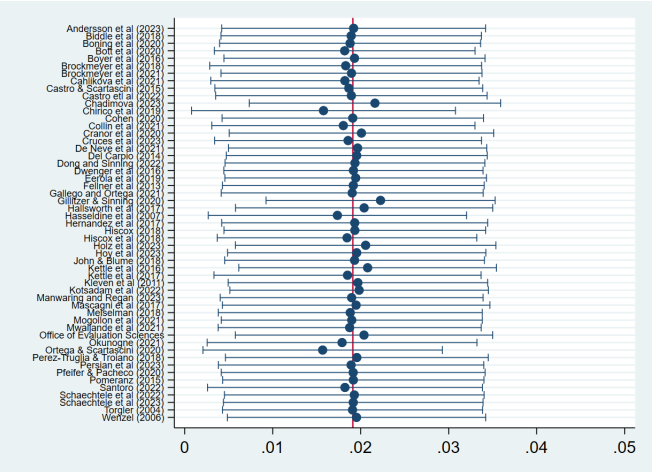
The table shows two exercises in detecting p-hacking for four different bandwidths around normalized significance thresholds in columns 1-3 and around the 5% significance threshold in columns 4-6 in the extensive margin sample. Panel A tests whether the empirical distribution corresponds to a binomial distribution with equal probability below and above the 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 paper level.

Figure B1: Jackknife exercise

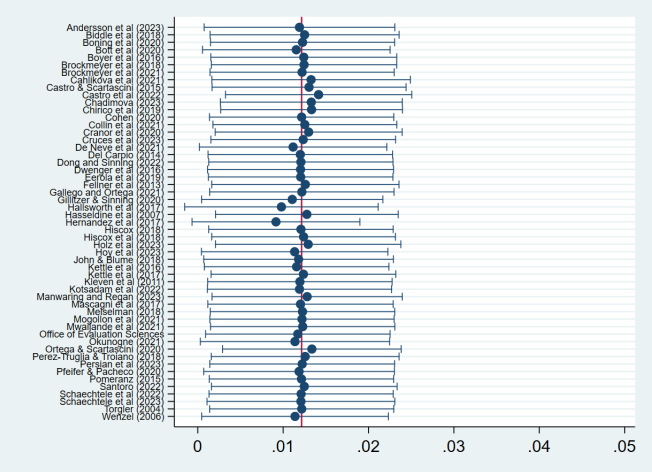
(a) Reminder-vs-Control, $\hat{\alpha}$



(b) Det-vs-NonDet, $\hat{\beta}$



(c) NonDet-vs-Reminder, $\hat{\gamma}$



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.

Download ZEW Discussion Papers:

https://www.zew.de/en/publications/zew-discussion-papers

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

IMPRINT

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de