ESSAYS IN PUBLIC FINANCE

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

vorgelegt von
Carina Neisser

April 30, 2021

Dean
Prof. Dr. Joachim Lutz

Referees
Prof. Dr. Philipp Dörrenberg
Prof. Dr. Andreas Peichl

Day of oral examination
07.06.2021

Acknowledgments

I am very thankful for the support I received from my supervisors Philipp Dörrenberg and Andreas Peichl for guiding me throughout this thesis. You were always patient with me and interested in my ideas. I highly appreciate your constructive criticism and invaluable advice. I am also grateful for the encouragement, mentoring and support I received from Sebastian Siegloch and Holger Stichnoth. I acknowledge the time and energy you all have invested in me. You not only gave me the freedom to pursue my research ideas but you were always there for me when I needed you. I further thank Johannes Voget for completing my dissertation committee.

My projects have benefited a lot from the research environment in Mannheim. I am grateful for the stimulating and welcoming working environment my (former) colleagues and staff created at ZEW. I enjoyed sharing an office with Max Löffler, Martin Ungerer and Florian Buhlmann. Thank you for the inspiring discussions about research and life. I also would like to thank Maximilian Blömer, Paul Hufe, and Christian Wittneben. I appreciate that you always cheered me up in challenging situations. A big thank you goes to my co-authors Abel Brodeur and Nils Wehrhöfer for their continuous patience and constructive feedback. I learned a lot while working with you.

Visiting UC Berkeley has been a rewarding experience, both personally and academically. I am grateful to Emmanuel Saez for his time and comments on my projects. He made it possible for me to visit such an inspiring place. Faculty members, PhD students and seminar speakers gave valuable advice on my projects and research in general.

The final part of this thesis was written at the University of Cologne. I thank everyone who supported me during the final stage of my PhD. I thank Felix Bierbrauer, Anna Bindler, Barbara Boelmann (our weekly session helped me a lot!), Emanuel Hansen and Matthias Heinz.

Throughout my PhD, I have received support and advice from many people. I apologize for all the forgotten names. This thesis has benefited strongly from the contributions of various discussants and comments received at numerous research seminars, conferences and conversations. My work also benefited from the support of numerous RAs. A special thanks goes to Markus Becker and Lothar Groß. Without your help I would never have studied

Economics in the first place. But most importantly, I would like to thank my husband Florian for his unconditional support. His love and humour have been a great source of help to deal with all kinds of problems related to my PhD. Anna, thank your for putting a smile on my face. You made all the (research related) challenges disappear.

Carina Neisser
Cologne, April 2021

Preface

This thesis consists of three chapters. The first chapter covers topics in the fields of public economics. The second chapter covers transparency in economic research. Chapter three focuses on questions from the field of political economy and public economics.

In Chapter one, sole-authored (accepted at the Economic Journal), I systematically investigate the sensitivity of the elasticity of taxable income (ETI). Optimal tax policy requires to quantify the magnitude of behavioural responses to tax changes. The ETI summarizes different types of behavioural responses to income taxation such as real responses (e.g. labour supply adjustments), tax avoidance (e.g. (legally) claiming deductions or income shifting between tax bases) and illegal tax evasion behaviour. To examine the large variation found in the literature of taxable and broad income elasticities, I conduct a comprehensive meta-regression analysis using information from 61 studies containing 1,720 estimates. To account for the central role of deductions and to disentangle real and reporting responses by individuals, I allocate all reported elasticities to two subsamples: before (BD) and after deduction (AD) elasticities. I show that the the vast majority of estimates (90%) lies within an interval of -1 and 1, with a strong propensity to report estimates between 0 and 1 and both distributions reveal an excess mass between 0.7 and 1. The broader range of responses is reflected by larger AD elasticities. Within my sample AD elasticities exhibit a mean of 0.403, while BD elasticities have a mean of 0.287. My findings reveal that estimated elasticities are not immutable parameters. First, elasticities that account for deductions are not only larger by definition, but they are also more sensitive to the estimation technique. Second, I link estimated elasticities to inequality measures as well as tax system- and economy related characteristics. My study shows that AD elasticities are highly correlated with top income shares. Finally, selective reporting bias is prevalent, and the direction of bias depends on whether deductions are included in the tax base. Overall, my findings have important policy implications. An application of a simple formula to derive optimal revenue maximising top tax rates, lead to tax rate of 62.5% if I incorporate the mean AD elasticity of 0.403 found in the empirical literature. Using my derived stylized AD estimates ranging from 0.074 to 0.827, lead to tax rates between 44.63% to 90.01%. Hence, the large variation found in the literature translate into a broad range of optimal top tax rates.

In Chapter two, co-authored with Abel Brodeur, we investigate the relationship between methods of data collection (administrative versus survey, for instance), availability of replication material and statistical significance. The credibility revolution in empirical economics is characterized by a shift towards methods focusing on causal inference along with the availability of more and better data. A large literature documents the extent of p-hacking (i.e., manipulation or selective reporting of $p-$values) and publication bias (i.e., outcome and statistical significance of a study is related to the decision to publish) in economics and other disciplines (Andrews and Kasy (2019); Brodeur et al. (2016) among others), the question of whether specific methodologies for data collection suffer from more selective reporting has not received a great deal of attention. This is a key research question given the increasing accessibility (and use) of administrative and proprietary data in general, and the fact that publication bias and p-hacking issues cast doubt upon the credibility of published research in the eyes of policymakers and others. If policymakers and citizens are more likely to see studies finding a significant effect of a given policy, then this would lead to a misrepresentation of the policy's real effect (Blanco-Perez and Brodeur (2019) and Abadie (2020)). In addition, we test whether the availability of data and code sharing policies somehow affect selective reporting bias. Using the universe of hypothesis tests reported in journal articles employing experimental and quasi-experimental methods reported in 25 top economics journals, we find no evidence that the extent of selective reporting differs for admin and survey data and that papers providing replication material are less p-hack. The last result is of particular importance, since disclosure standards and open sharing of data and code have been been increasing over the last decade. Given that the proportion of articles receiving exemptions from data-sharing policies for admin data is larger than for other data types, the increasing use of administrative records may raise concerns about the reproducibility of research findings, and ultimately, research credibility. We show that the difficulty of providing replication material is not a key factor driving p-hacking in economics.

In Chapter three, jointly written with Nils Wehrhöfer, we study the effects of public disclosure laws on outside activities and earnings of German federal members of parliament (MPs). German politicians are allowed to carry out outside activities in addition to their political work but the execution of the mandate should be central to a politician's activity. Starting in 2005, German federal MPs were forced by law to publish their outside activities and earnings in a bracket system top-coded at 7,000€. Initially, the information was only privately disclosed. In 2007, private was replaced by public disclosure (also retroactively to 2005) and these information can be accessed online on webpages of the German Bundestag. However, the top-coding at 7,000€ was criticized because MPs could cover their well paid activities and voters could not differentiate between a moderate and high earning MP. In 2013, more brackets were introduced such that earnings above 250,000€ were now censored. This increased greatly

the information available to voters. First, we exploit both reforms to identify the causal effects of disclosure rules on politician's outside earnings using administrative tax return data. It allows us to observe pre-reform income as well as using unaffected state MPs as a control group. Our results indicate that for the first reform the top-censored nature of the reporting scheme has the consequence of raising outside earnings, while the second reform provides evidence that a higher degree of public disclosure leads to a decrease in outside earnings. Second, we explicitly distinguish between the effects of private versus public disclosure and find no effect of private disclosure. Third, to identify potential mechanisms behind our findings, we collected the published information on earnings and activities along with political and electoral variables. We show that social norms and electoral accountability matters.

Contents

The Elasticity of Taxable Income: A Meta-Regression Analysis

1.1 Introduction

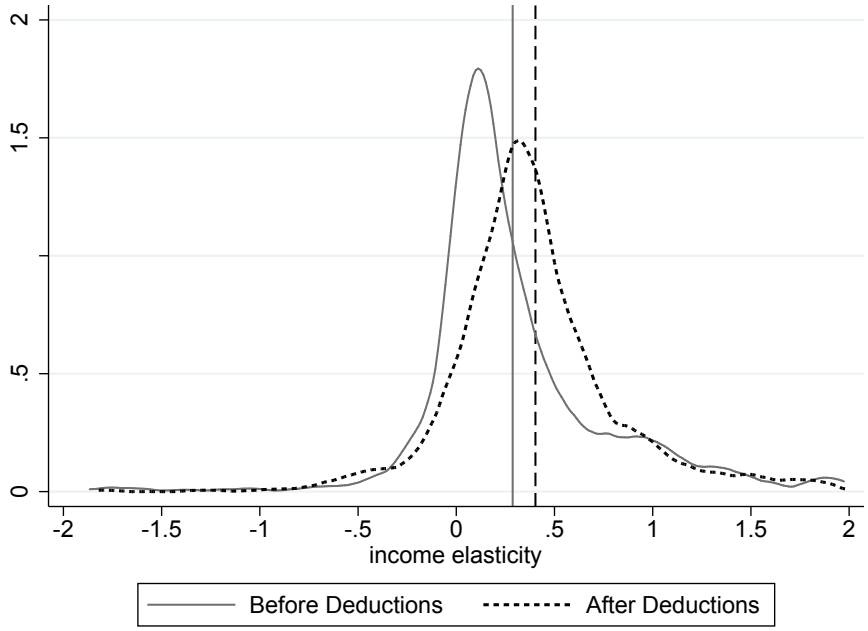
The design of tax and transfer policies requires to quantify the magnitude of behavioural responses to tax rate changes in order to determine optimal policy. Larger responses to taxation, for example, will lead to smaller revenue-maximizing tax rates for top income earners, conditional on the shape of the income distribution (Saez (2001); Saez et al. (2012)). The elasticity of taxable income (ETI) summarizes different types of behavioural responses to income taxation such as real responses (e.g. labour supply adjustments), tax avoidance (e.g. (legally) claiming deductions or income shifting between tax bases) and illegal tax evasion behaviour. It serves not only as a behavioural parameter in optimal taxation models (Mirrlees (1971); Diamond (1998); Saez (2001); Piketty and Saez (2013)) but also as sufficient statistic for dead-weight loss calculation (Feldstein (1999) or Chetty (2009)). Since Feldstein (1995), a large body of empirical work estimating taxable income responses has emerged. Despite the importance, there is little consensus on the magnitude of these elasticities to be used in economic policy analysis.

In this paper, I provide a comprehensive quantitative survey by applying meta-regression techniques. 'Elasticity of taxable income' is used as an umbrella term for all types of elasticities (e.g. adjusted gross and taxable income). In total, I collect 1,720 estimates extracted from 61 studies. I only consider Difference-in-Differences (DID) and Instrumental Variable (IV) approaches and do not cover bunching (e.g. Saez (2010)) or time series evidence (e.g. Mertens and Montiel Olea (2018)) as these estimates are conceptually different and therefore not comparable to each other.

To account for the central role of deductions and to disentangle real and reporting responses by individuals, I explicitly differentiate between behavioural responses that are based on income concepts with or without tax deductions and allocate all reported elasticities to two subsamples:

before (BD) and after deduction (AD) elasticities. In addition to real responses (e.g. changes in labour supply), many tax systems offer a wide range of deductions to legally avoid taxes. Figure 1.1 plots the distribution of elasticities of both subsamples. The vast majority of estimates (90%) lies within an interval of -1 and 1, with a strong propensity to report estimates between 0 and 1 and both distributions reveal an excess mass between 0.7 and 1. The broader range of responses is reflected by larger AD elasticities. Within my sample AD elasticities exhibit a mean of 0.403, while BD elasticities have a mean of 0.287.

Figure 1.1: Distribution of elasticities



Notes: The distribution of before deduction (BD) elasticities are displayed as a solid line and the corresponding vertical line highlights the mean of 0.287 (N=940). The distribution of after deduction (AD) elasticities are displayed with a dashed line and the corresponding mean of 0.403 is highlighted with the vertical dashed line (N=780). Both figures are based on an an Epanechnikov kernel with a bandwidth of 0.072.

Researchers that estimate the ETI face various empirical challenges. Most importantly, income and marginal taxes are jointly determined and potential solutions like Instrumental Variables (IV) approaches have been developed. In addition, different income growth rates across the population or reversion to the mean require solutions because it is hard to disentangle income growth driven by tax and non-tax effects. Most notably functions of past income are included in the regressions. While the choice of the specific regression specification depends on the underlying model, there is some discretion in the way that specific methods and controls

are implemented, which can partially affect chapters/etiresults.

I identify and assess different explanations for the pattern of estimates found in the empirical literature. More precisely, different categories for each study (e.g. empirical strategy or country) are recorded and differences between elasticity estimates are quantitatively examined. Importantly, my meta-analysis provides a replicable statistical framework for summarizing and assessing the full range of empirical evidence.[1] Although the ETI literature has been reviewed by Saez et al. (2012), I am not aware of any meta-regression analysis of taxable income elasticity estimates.

My results show that elasticities that account for deductions are not only larger by definition, but they are also more sensitive to the estimation technique. A calculation of stylized elasticity estimates documents a wide range of possible estimates. When accounting for the implemented estimation specification in primary studies, my regression chapters/etiresults show that average BD elasticities lie in the range of 0.053 to 0.120, while average AD elasticities vary from 0.074 to 0.827. Richer income control variables always lower estimated elasticities and the effect is more pronounced in the AD subsample. It remains unclear which income control is an appropriate choice to disentangle non-tax from tax-related income responses.

I link estimated elasticities to inequality measures as well as tax system- and economy related characteristics. More precisely, I add country and year specific characteristics to my collected data to to provide suggestive evidence that elasticities are related to contextual factors. Slemrod and Kopczuk (2002) and Kopczuk (2005) emphasise the fact that the ETI is considerably larger in tax systems with more deduction possibilities and can therefore be controlled by policy makers. Much of the evidence is based on self-employed and/or high-income taxpayers, given their larger range of opportunities to adjust their (taxable or gross) income (e.g. Kreiner et al. (2016); Le Maire and Schjerning (2013) or Harju and Matikka (2016)). Alvaredo et al. (2013) highlight the role of tax policy and its effects on income inequality. In addition, Kleven et al. (2011) and Kleven et al. (2016*a*) stress that third party information reporting (e.g. the exchange of information of employers or banks and tax authorities) influences the magnitude of behavioural responses.

My analysis provides evidence that estimated elasticities are not immutable parameters with respect to the empirical strategy but they are also linked to past as well as current (tax-)policy and that the underlying context matters when interpreting these elasticities. There is a positive correlation between inequality measures and estimated elasticities. In particular, AD elasticities are highly correlated with top income shares. A widening of the income distribution might

[1] See Christensen and Miguel (2018) for a review of research transparency and reproducibility in economics. Card and Krueger (1995) and Card et al. (2010, 2018) are three examples that look into the field of labour economics. Havránek (2015*a*) examines the literature on intertemporal substitution elasticities and Lichter et al. (2015) study labour demand elasticities.

be the result of past tax cuts for high-income earners. Such developments are insufficiently considered in the initial estimation of elasticities of taxable income, leading to an upward bias in resulting AD elasticities. Moreover, the level of third party information reporting within an economy is unrelated to elasticities that account for the deduction component, while it is negatively related to the magnitude of elasticities that do not consider deductions. Typically, deductions are not subject to third party information reporting. The degree of information exchange between tax authorities and firms or other institutions can be influenced by policy makers and thereby also influence the magnitude of estimated elasticities.

I focus on two types of selective reporting bias. The first is the so-called 'file drawer problem' (Rosenthal (1979*a*)). It describes the fact that many studies or chapters/etiresults have never been published because they do not reveal the expected sign, magnitude and/or significance. The second type of selection reporting bias arises, if researchers use well-known chapters/etiresults as a reference point and hence are inclined to report only chapters/etiresults that are in line with these findings. With respect to the ETI, researchers generally put more trust into estimates ranging from 0 to 1. With their seminal contribution, Gruber and Saez (2002) have further shaped this belief by providing a value of 0.4 as their main result.

Graphical evidence as well as regression chapters/etiresults confirm the prevalence of selective reporting bias in the literature of taxable income elasticities. In general, there is a tendency to report significant chapters/etiresults more often. The existence of 'p-hacking' is more pronounced among AD elasticities and among published articles compared to working papers. Since the publication of Chetty (2009), BD (e.g. gross income) elasticities have begun to receive more attention. This increased interest is reflected by a larger amount of 'p-hacking' within the BD-subsample for estimates published after 2009. In addition, I observe excess (distributional) mass around 0 to 0.4 and below 1. These anomalies in the distribution of estimates suggests that chapters/etiresults are more likely to get reported because they are in line with theory and existing evidence. In general, there is an upward reporting bias for BD elasticities. For AD elasticities, the reporting bias goes in both directions, while the downward bias appears to be more dominant.

The remainder of this paper is structured as follows. In Section 1.2, I explain the meta regression model and I describe the data collection process. In Section 1.3, I outline a basic framework to discuss empirical challenges in the literature on taxable income elasticities (1.3.1) and provide explanations of defined categories of heterogeneity (1.3.2) along with descriptive statistics (1.3.3). In Section 1.4, I provide and discuss the baseline results and correlations between contextual factors and elasticities. In Section 1.5, I highlight the prevalence of selective reporting bias. Section 3.6 concludes.

1.2 Meta-regression Framework and Data Collection

I follow standard meta-regression analysis techniques (e.g. Card et al. (2010, 2018)). The meta regression model is given by

$$\zeta_{is} = \zeta_0 + \beta X_i + \delta Z_s + \epsilon_{is}, \tag{1.1}$$

where ζ_{is} represents the i-th estimate estimate collected from study s. ζ_0 denotes the intercept, X_i and Z_s represent study and estimate-specific variables respectively, and ϵ_{is} is the sampling error. Since the variances of collected estimates are heteroscedastic, it is preferable to estimate the model using Weighted Least Squares (WLS) rather than through an OLS estimation. I use the inverse of the error term variance of an individual estimate $\mathrm{V}(\hat{\zeta_{is}}) = \sigma^2_{is}$ as analytic weights. Hence, I give observations with smaller variances a larger weight and greater influence on the estimates since precision can be seen as an indicator of quality.[2] Standard errors are clustered at the study level to control for study dependence in the estimates.

Data Collection. A comprehensive review and examination of the ETI literature provided the data for the meta-analysis.[3] As a first step, I searched Google Scholar and IDEAS RePEc using the following search terms: 'elasticity of taxable income', 'eti', 'taxable income', 'new tax responsiveness' and 'tax elasticity.' In addition, I relied on a survey by Saez et al. (2012) to identify relevant studies published prior to 2011 and I cross-checked these with the reference list of all previously identified papers. I checked only English- and German-speaking articles. The main search process lasted from 2015 to 2019 and I identified 203 potential studies.

In the second step, I applied certain exclusion criteria to determine the final sample of studies. I only coded studies that provide their own empirical estimates and rely on commonly used income concepts as described below. Based on this sample, I found 37 studies that were published. Additional working papers increased the number of articles to 61.[4] In the third step, I collected every estimate derived from a different specification (so-called multiple

[2]To test the robustness of the chapters/etiresults with respect to the underlying weights, I conduct various regressions (see (online) Appendix F): (1) a simple OLS, (2) Random effects meta-regression technique, (3) a WLS with weights that are based on the inverse of the share of observations per study in relation to the full sample and (4) a WLS with weights that account for the sample size of each study. Last, to check whether clustering in the meta-analysis produces misleading inferences, I apply a wild-cluster bootstrap procedure for improved inference with only a few clusters.

[3]The meta analysis follows reporting guidelines proposed by Stanley et al. (2013). A list of people who have coded and checked the data, a list of identified but not-included studies and estimates or a list of all included estimates plus sources is provided upon request.

[4]In the (online) Appendix A, I provide an overview of studies included in the sample. On the one hand, adding unpublished papers to the meta-sample might lower the quality of included estimates but, on the other hand, most working papers are more recent and use better datasets and improved estimation techniques. It should be noted that this meta study is only as good as the studies on which it is based and there might be variation among the studies that cannot be reflected by the coded variables.

sampling) so that they are different with regard to the defined categories of heterogeneity (e.g. income concept or sample restrictions). I collected all point estimates, corresponding standard errors, number of observations and type of control for heteroscedasticity and autocorrelation. Additional information on journal, year of publication, country and time period is coded. In the fourth step, I restricted the final dataset to estimates that provide a standard error or t-statistic. My sample consists of 1,720 observations. Finally, I collected all necessary study characteristics, which I will explain in the next section. Additional information on contextual factors such as tax system and economic characteristics as well as inequality measures are collected and merged with the dataset (see Table 1.1 for an overview).

1.3 Elasticity of Taxable Income

In this Section, I briefly explain the concept of taxable income elasticities. I outline the most standard regression specification and I state empirical challenges. For a detailed discussion, please refer to an excellent survey by Saez et al. (2012). I present various reasons why elasticity estimates differ and describe the coded characteristics along with a more in depth explanation in Section 1.3.2. In Section 1.3.3 I provide some descriptive statistics.

1.3.1 Empirical Challenges

The (taxable) income literature uses an extension of the traditional labour supply model. Individuals maximize a utility function $u(c, z)$, where z is income and c consumption. An elasticity of the income tax base measures the responsiveness of income to changes in the net-of-tax rate (NTR) - defined as one minus the marginal tax rate. This is the percentage change in income in response to a one percent increase in the NTR. An increase in the marginal tax rate reduces the NTR, which in turn reduces taxable income. Hence, the expected elasticity should be positive.[5]

Collected elasticity estimates are summarized such that they belong either to the before or after deductions subsample. Since an elasticity is a function of the definition of the tax base, the applied income concept determines the range of responses. These responses can take many forms, including changes in labour supply (participation and working hours), tax avoidance (changing the timing of income/transactions, changes in the extent of spending on tax deductible activities, e.g. donations, or even claiming questionable deductions) and tax evasion (understating income, claiming unjustified deductions). The distinction between whether or

[5]Information about estimated income effects is rarely available (e.g. Gruber and Saez (2002) or Bakos et al. (2010)) so, I ignore them and assume that compensated and uncompensated elasticities are equal.

not an income concept considers deductions is crucial. Real responses can be captured with a before-deduction elasticity while an after-deduction elasticity captures a broader range of responses, including avoidance behaviour. Tax evasion affects both types of elasticities. Ideally, we would like to observe a comparable and uniformly defined income across all studies. This is impossible even for conceptually equal income concepts like taxable income. The exact definition varies from country to country and, even within a country, over time. Researchers mainly use taxable, adjusted gross, or total income to capture behavioural responses towards taxation. Total income (= gross or broad income) is the sum of all income. Subtracting specific deductions (e.g. retirement plan contributions), yields adjusted gross income. Taxable income is calculated as adjusted gross income minus personal exemptions and itemized deductions.[6]

The most standard regression specification is derived as:

$$log\left(\frac{z_{it}}{z_{it-k}}\right) = \zeta log\left(\frac{1-\tau_{it}}{1-\tau_{it-k}}\right) + \delta f(z_{it-k}) + \theta X_{it-k} + \mu_t + \epsilon_{it}, \tag{1.2}$$

where i refers to the respective taxpayer and t is the underlying year. ζ is the parameter of interest, k is the chosen difference length and $t-k$ denotes the base-year. X_{it-k} is a vector of control variables. Time dummies μ_t control for any omitted variables in differences that are the same on average for all individuals. $f(z_{it-k})$ denotes the income control in order to capture non-tax related income trends. In equation (1.2) ζ represents the elasticity of the income tax base that measures the responsiveness of income to changes in the net-of-tax rate $(1-\tau)$.

Several conditions must hold in order to estimate behavioural responses correctly. First, only the marginal tax rate τ should change, while changes in the tax base z are kept constant. In reality, however, the underlying tax base often changes simultaneously with the tax rate itself. To rule out any tax legislation-induced tax base effects, the broadest definition and, therefore, an 'artificial' tax base across years is used. For the US, researchers mostly rely on the TAXSIM calculator developed at the NBER. In other cases, the constant tax base along with a tax simulation model is often constructed by the researcher himself. Building a tax simulation model requires a broad understanding of the underlying tax law as well as tax base changes across the years under study.[7]

[6] In the (online) Appendix B, Table 8 provides summary statistics by reported *income concept*. As a sensitivity check, I run the estimations on a subsample of the dataset and look only at taxable income elasticities (see Table 17). These chapters/etiresults remain unchanged compared to estimation chapters/etiresults that consider all AD estimates.

[7] In many studies, details of the tax simulation model are missing. For example, although capital gains are part of taxable income, only a few studies explicitly mention that they subtract capital gains. In addition, most researchers remain salient on whether or not they apply a constant tax base approach. Both things influence the definition of taxable income and therefore the chapters/etiresults.

Second, in a progressive tax system, the marginal tax rate τ and income z are jointly determined, and tax rates increase automatically if an individual faces a (non-tax related) positive income shock and potential income responses are (wrongly) captured by the ETI. Following Auten and Carroll (1999) and Gruber and Saez (2002), most studies use an instrument that is based on mechanical changes in tax rates that are induced by tax reforms. The idea is that this change in net-of-tax rates is free of any behavioural responses, representing only mechanical changes that can be used as an instrument for the NTR. To construct mechanical tax rate changes, one uses income from base year $t-k$ and assumes that it remains the same in year t. Applying tax rules for year t yields a mechanical (sometimes called predicted or synthetic) tax rate. More developed instruments try to account for the growing concern that this instrument is not sufficiently exogenous. For instance, Weber (2014) argues that mechanical tax rate changes mentioned above should be lagged in order to fulfill the exclusion restriction. Her approach makes it possible to deal with serially correlated transitory income shocks.

Third, different income growth rates across the population (e.g. larger income growth for high-income earners) and reversion to the mean further aggravate a 'clean' estimation. For example, when income changes are driven by temporary income shocks or different parts of the income distribution grow at a higher rate, it is hard to disentangle income growth driven by tax and non-tax effects. In the case of tax cuts for upper-income groups, secular changes in income (e.g. larger income growth at the top), lead to an upward bias and mean reversion might go in both directions depending on the type of income shock. These shocks influence the shape of an income distribution and they need to be incorporated in an empirical framework. While administrative tax datasets offer precise information about a taxpayer's income and deductions, socio-demographic information and therefore the amount of other control variables is limited. To capture non-tax related income growth, researchers use income controls $f(z_{it-k})$ and apply sample restrictions. The simplest income control is the log of base-year income $ln(z_{i,t-k})$ (Auten and Carroll, 1999). More sophisticated income controls like a spline of $ln(z_{i,t-k})$ are applied as well (Gruber and Saez, 2002). The same is true for sample restrictions. Since mean reversion is pronounced at the bottom of the income distribution, the income distribution is often restricted from below. For instance, typically taxpayers with an income below 10,000$ are excluded from the analysis.

Fourth, variation in marginal tax rates used for identification are assumed to be exogenous. This assumption is violated if tax changes are systematically correlated with other developments that affect economic measures such as GDP. For instance, a tax policy that reduces taxes because policy makers are anticipating a recession is clearly endogenous (Romer and Romer (2010)). If tax changes are correlated with other developments, this leads to a biased estimated of the effect of tax changes. Even previous tax changes can affect current elasticities. Finally, many tax

reforms do not target a single income group, and income groups may face different tax rate changes in magnitude throughout the income distribution. To disentangle any non-tax related income changes that systematically vary by income group from the effects of tax rate changes on income becomes even harder, if the extent of tax rate changes is correlated with income.

1.3.2 Categories of Heterogeneity in Estimated Elasticities

Now, I describe my coded characteristics in more detail. Many factors influence the size of an estimate. To assess the relevance of different explanations, I define various categories of heterogeneity: (1) income concept; (2) estimation techniques; (3) sample restrictions; (4) publication characteristics, including variation across countries and time; and (5) contextual factors. Dimension (1) to (4) are collected from primary studies while dimension (5) is based on external data sources. There are more dimensions of heterogeneity worth investigating, such as the role of income effects, restrictions on demographics (e.g. gender) or tax system-related characteristics (e.g. restricting the sample to individuals who are not eligible for the alternative minimum tax in the US) and even certain control variables such as education. However, a limited number of estimates account for these variables, which makes it impossible test for them. Table 1.1 provides an overview of all included characteristics and I describe each coded variable in greater detail in the (online) Appendix C.2.

Income Concept. I only distinguish whether or not the dependent variable considers deductions, and I allocate all reported income concepts to two subsamples: before (BD) and after deductions (AD). Kopczuk (2005) shows how the ETI varies with its tax base. While the AD elasticity is considerably larger in a tax system with more deduction possibilities, it can also be lower in a country with a high degree of third party information reporting (e.g. exchange of information between employer and tax authority) (Kleven and Schultz, 2014).

Estimation techniques. I define four distinctive features with respect to estimation techniques that influence the ETI: (a) regression technique; (b) income control; (c) difference length; and (d) weighting by income.

I categorize five *regression techniques*. Since income and marginal tax rates are jointly determined, almost all approaches follow an Instrumental Variable (IV) procedure. They essentially differ in the way they instrument for the net-of-tax rate $(1 - \tau)$. Following Gruber and Saez (2002) the most standard approach is defined as 'IV: mechanical tax rate changes.' The second estimation technique is called 'IV: (lagged) mechanical tax rate changes' (Weber, 2014). Different instruments have recently been developed. For instance, Burns and Ziliak

(2017) use a Wald-type grouping instrumental variables estimator. Instead of using a person-specific instrument, they construct a new instrument, which is the cohort-state-year mean of the synthetic tax rate. I summarize all other types of instruments in a third category (IV:other).[8] The earliest method, namely a basic Difference-in-Differences (DID) approach, uses a defined treatment and control group without any instruments and income controls (Feldstein (1995)). Difference-in-Differences (DID) with a dummy variable as an instrument represents another category. This is a conventional DID approach in which the NTR is instrumented by the interaction of the after-reform and treatment group dummy. This is similar to Feldstein's (1995) tabulated DID approach, but estimated in a regression framework that allows for additional control variables (Moffitt and Wilhelm (2000)).

I define five generations of *income control variables*. First, there is the use of no additional income control variables (none). Studies published prior to 2000 use no income controls and most studies estimate a specification with no income controls as a sensitivity check. The second generation covers studies that use only the log of base-year income control $ln(z_{i,t-k})$ (Auten and Carroll, 1999). Following Gruber and Saez (2002) researchers use more sophisticated income controls like a spline of log base-year income. A spline divides income groups into deciles to account for non linear income trends across these groups. Kopczuk (2005) argues that using only base-year income and some flexible function is not sufficient. He explicitly distinguishes between permanent and transitory income components and proposes two types of income control variables: the log of lag base-year income $ln(z_{i,t-k-1})$, which allows one to control for an individual's rank in the income distribution and therefore for the permanent income level; and transitory income trends are captured by using the deviation between log base-year and log lag base-year income $ln(z_{i,t-k}) - ln(z_{i,t-k-1})$. The last generation covers every *other* (non-standard) income control used in the literature, e.g. cohort-state-year income control as used in Burns and Ziliak (2017).

All studies apply a 'First Difference' estimation strategy with a varying *difference length* to eliminate the impact of unobservable time-invariant characteristics. An estimate is either based on a specification with a time window of 1-year, 2-year, 3-year or of 4 and more years. The chosen difference length $t - k$ has an effect on resulting estimates. Most estimations use a 3-year time window such that researchers relate income and marginal tax rates e.g. from 2001 to 2004. One might think that the longer the time window, the larger the behavioural response. However, the timing, announcement and implementation of underlying reform(s), individual speed of understanding, as well as an individual's ability to adjust their income have an effect on the size of behavioural adjustments. Since many tax reforms are phased-in over several years, an estimate is only a combination of short-, medium- and long-run responses (Weber

[8] All 'other' instruments are explained in greater detail in the (online) Appendix C.2.

(2014)).

Since weighted elasticity parameters reflect the relative contribution to total revenues, regression chapters/etiresults are sometimes *weighted by income* (Gruber and Saez (2002)).[9] If responses do not vary by income, weighting the estimates by income will not affect elasticity estimates. However, it seems reasonable to assume that behavioural responses are not homogeneous across the income distribution. Weighted chapters/etiresults account for the fact that high-income taxpayers tend to exhibit larger responses. Typically, these weights are censored at the top (e.g. at 1 $ million) and are not free of criticism, since income itself is endogenous (Weber (2014)). Individuals who face a temporary positive income shock will receive a larger weight. The weight is even larger if high-income earners are affected. Hence, resulting estimates are even more strongly distorted.

Sample restrictions. I coded whether *income cutoffs* are used and, if so, the corresponding threshold. These thresholds are re-calculated in US-Dollar. To account for mean reversion at the beginning and end of an individual's working life, researchers apply an *age cutoff* to limit the sample to the working population and to exclude pensioners.[10]

Publication characteristics and variations across countries and time. To account for potential differences, I control for whether or not an estimate is reported in a journal or in a working paper. Given the research process, I include different categories for *publication decade* ((1) <= 2000, (2) 2001-2010; (3) >2010) as controls. Publication decade does not necessarily coincide with the timing of a tax reform. To identify a potential development over time which is not directly related to any type of methodological progress but rather related to tax policy at a given time, I include *estimation/ data decade* as a control. For a particular estimate, I calculate the mean of the first and end years of the underlying data period ('mean year of observation') and assign the corresponding decade: 1980s, 1990s or 2000s. Countries are summarized in different *country groups* (1) USA, (2) Scandinavia (Denmark, Norway, Sweden), and (3) other countries (Canada, Finland, France, Germany, Hungary, Netherlands, New Zealand, Poland, Spain).

Contextual Variables. Inequality measures and economic characteristics shape behavioural responses to taxation. To account for income inequality within an economy, I include the *Gini coefficient* (disposable income, post taxes and transfers). In addition, I consider a measure of the share of pre-tax national income that is held by the *top 1%* and *top 10%* as

[9] Similar to missing details regarding whether or not capital gains are included in the income concept, it often remains unclear by what type of income estimates are weighted.

[10] In the (online) Appendix D, I provide estimation chapters/etiresults that account for sample restrictions with respect to *marital status* and *employment type*.

contextual variables in my regression. An increase in inequality might be the result of past tax cuts for high-income tax payers. Hence, larger estimates might not be the result of larger responses, but rather of a widening in the income distribution that is captured by estimated elasticities.

Aspects of a given tax system as well as the underlying business cycle are related to behavioural responses to taxation. Kleven and Schultz (2014) find that behavioural elasticities are larger when estimated from large tax reform episodes and a more salient tax reform is more likely to overcome optimization frictions. Therefore, I account for the *introduction of a top tax bracket*. Since such a reform is more salient and the affected tax group is the most responsive one, this might lead to higher estimates.[11] Hargaden (2020) provides evidence of a weaker behavioural response during a recession and therefore highlights the role of business cycle fluctuations. To account for a given economic situation, I add the respective *unemployment rate* as a contextual variable in my regression.

Third party information reporting (e.g. the exchange of information of employers or banks and tax authorities) plays a key role in tax compliance and a country's overall tax take. Kleven et al. (2011) find that the overall tax evasion rate is very small in Scandinavia because almost all income is subject to third party information reporting. I include two variables as a proxy to check for its influence. First, the *fraction of self-employed* workers within a country. Traditionally, self-employed taxpayers provide most of the necessary information to tax authorities themselves. I expect a positive relationship between elasticities and the share of self-employed workers within an economy. As a second measure, I include the share of *modern taxes per GDP* to proxy for the share of tax revenue that are exposed to third-party information reporting compared to the overall tax take. Kleven et al. (2016*a*) distinguish between what they call traditional and modern taxes. Unlike traditional taxes, which rely on self-reported information, modern taxes rely on third-party information.[12] I expect a negative correlation between reported elasticities and modern taxes to GDP ratio.

1.3.3 Descriptive Statistics

Table 1.1 provides an overview of the collected information to explain differences in elasticity estimates. As already mentioned, I divide the meta-sample in two subsamples depending

[11] Tax reforms are necessary to generate variation that can be exploited. A reform does not happen in a single year, nor is it easy to tell exactly which income group is affected. Moreover, most estimates are based on a data period with more than one single change in tax law. This makes it difficult to account for other tax reform characteristics in the meta analysis.

[12] Modern taxes are defined as personal and corporate income taxes, value-added taxes, payroll taxes, and social security contributions, whereas traditional taxes are all other taxes (e.g. inheritance tax). Modern taxes play a crucial role in the economic development of a country and there is a strong positive correlation between GDP per capita and modern taxes to GDP.

on whether the underlying income concept accounts for deductions. The before deductions subsample consists of 940 observations collected from 46 studies and the after deduction subsample of 780 observations from 41 studies. Around 60% of the estimates refer to a regression technique that uses mechanical tax rate changes as an instrument. One third of estimates use the log of base year income (Auten and Carroll, 1999) as an income control. Most estimates either use a difference length of three years or consider a short time window of one year. 40-50% of all primary estimates are weighted by income. Almost half of the estimates apply an age cutoff and the vast majority of estimates use an income cutoff.

Table 1.1: Descriptive Statistics: Categories of Heterogeneity

	Before Deductions (BD) (N=940)		After Deductions (AD) (N=780)	
	Mean	**Std. Dev.**	**Mean**	**Std. Dev.**
Estimation Techniques				
Regression technique				
IV: mechanical tax rate changes	0.651	0.477	0.609	0.488
IV: (lagged) mechanical tax rate changes	0.041	0.20	0.165	0.372
IV: other	0.094	0.291	0.127	0.333
DID and IV	0.188	0.391	0.045	0.207
classic DID	0.026	0.158	0.054	0.226
Income Control				
Auten Carroll (1999)	0.286	0.452	0.226	0.418
none	0.206	0.405	0.224	0.417
Gruber Saez (2002) spline	0.181	0.385	0.176	0.381
Kopczuk (2005) type	0.249	0.433	0.353	0.478
other	0.078	0.268	0.022	0.146
Difference Length				
3 years	0.395	0.489	0.512	0.500
1 year	0.366	0.482	0.287	0.453
2 years	0.074	0.263	0.128	0.335
4+ years	0.165	0.371	0.073	0.260
Weighted by Income	0.484	0.500	0.405	0.491
Sample Restrictions				
Age Cutoff	0.564	0.496	0.523	0.5
Income Cutoff				
0-10k	0.255	0.436	0.236	0.425
none	0.127	0.333	0.199	0.399
10k-12k	0.249	0.433	0.292	0.455
12-31k	0.191	0.394	0.114	0.318
$>$ 31k	0.178	0.382	0.159	0.366
Variations across Countries and Time				
Country Group				
USA	0.494	0.500	0.532	0.499
Scandinavia	0.184	0.388	0.099	0.298
other countries	0.322	0.468	0.369	0.483
Mean year in study data	1994.524	7.819	1995.976	8.849
Estimation decade				
$<$ 1999	0.286	0.452	0.288	0.453
1990 - 2000	0.394	0.489	0.262	0.440
$>$ 2000	0.320	0.467	0.450	0.498
Publication Characteristics				
Publication decade				
2001-2010	0.367	0.482	0.414	0.493
$<=$ 2000	0.063	0.243	0.033	0.180
$>$ 2011	0.570	0.495	0.553	0.498
Published Type				
published	0.672	0.470	0.671	0.470
working paper	0.328	0.470	0.329	0.470
Mean Year of Publication	2011.169	0	2011.169	0
Contextual Variables				
Gini	30.908	5.178	31.684	4.445
top 10% inc. share	0.333	0.059	0.341	0.061
top 1% inc. share	0.109	0.034	0.114	0.037
intro top bracket	0.278	0.448	0.218	0.413
unemployment rate	6.917	2.874	7.023	1.638
fraction self-employed	10.934	3.833	11.056	3.6
share of modern taxes	26.688	9.195	25.449	8.388

Notes: I present descriptive chapters/etiresults separately for two subsamples: before (BD) and after deductions (AD). The sample covers only observations with a given standard error or t-statistic. Reference categories are given in italics. More details can be found in the (online) Appendix C.2. For a given estimate, contextual variables are merged via country and/or mean year of observation.

1.4 Meta-Regression Results

In Section 1.4.1, I separately present the chapters/etiresults for before (BD) and after deduction (AD) elasticities. In addition, I present some stylized elasticity estimates. These estimates are intended to facilitate the interpretation of my chapters/etiresults and to summarize chapters/etiresults that correspond to the two most commonly applied approaches in the literature. In Section 1.4.2, contextual characteristics will be analysed separately and I show that both BD and AD elasticities are correlated with tax system- and economy related characteristics.[13]

1.4.1 Baseline chapters/etiresults

I run specification (1) on the *before* and *after* deduction subsample separately and present the chapters/etiresults in Table 1.2 and 1.3. I define the most commonly used characteristic as a reference category (written in bold) and omit this feature such that reported coefficients need to be interpreted as a deviation from a particular characteristic to the corresponding reference category. I gradually add the defined characteristics. In column (1) and (2) I only control for estimation technique, and in column (3) I account for sample restrictions. If 'no restriction' defines the base category, it means that a particular estimate is not restricted with respect to a certain characteristic. For instance, the baseline category for age restriction is 'no restriction.' Hence, estimates need to be interpreted in reference to other estimates that do not apply an age cutoff. chapters/etiresults on country group coefficients are presented in column (4) and (5), with column (4) accounting for (estimation) decade, column (5) controlling for (publication) decade. Column (6) presents the most comprehensive specification.[14] Baseline chapters/etiresults do not account for contextual factors. The reference specification in column (1) is defined as a specification that uses mechanical tax rate changes as an instrument, log base-year income control and a three-year difference length. For example, it refers to the most standard approach used by Kleven and Schultz (2014) in their baseline specifications. On average, such a specification yields a BD elasticity of 0.073 and an AD elasticity of 0.445. As expected, estimates that allow for deduction responses mostly reveal a larger constant and, therefore, are statistically more elastic to marginal tax rate changes compared to chapters/etiresults obtained based on the before (BD) subsample. Next, I present chapters/etiresults obtained for both subsamples by category of heterogeneity.

[13]To verify the robustness of the baseline chapters/etiresults, I apply various estimation techniques and further limit the dataset along certain dimensions (see tables 17 and 18 in the (online) Appendix F).

[14]Multicollinearity might be a problem in the regressions resulting in standard errors that are too large. This makes it difficult to isolate the influence of a single variable from overall influence. Therefore, I check if the variance inflation index is below 10 such that the presented chapters/etiresults are reliable within every estimation. Except for column (6) in Table 1.2 and Table 1.3 this condition holds.

Estimation techniques. My chapters/etiresults show that AD elasticities are more sensitive with respect to different aspects of the underlying estimation technique compared to BD elasticities. Starting with the choice of *income control*, most studies follow Auten and Carroll (1999) and include log base-year income as an explanatory variable. Compared to this approach a regression that does not consider any income control leads to lower and often negative BD elasticities (a fact that is already noted in most primary studies). My chapters/etiresults generalize this finding and quantify an average decrease by 0.2 in BD elasticities. This result is quite robust even in the most sophisticated specification in column (6). All other kinds of income control variables (in most cases more sophisticated ones) lower elasticities in both but in particular in the AD subsample. The success of these controls depends on the extent of year-to-year mean reversion and the stability of the underlying income distribution. However, there is a potential risk that they absorb too much identifying variation (see Saez et al. (2012) for a discussion). It is worth highlighting that Kopczuk-type income controls lower AD elasticities (on average) by 0.371 compared to a log base-year income control while other types of income controls (mostly splines) also decrease AD elasticities but at a lower rate.

The chapters/etiresults suggest that the chosen *difference length* has different effects on BD and AD elasticities. In the BD subsample, all specifications with a two-year time window have a marginally lower elasticity compared to specifications based on three-year differences, while there are no statistically significant differences in difference length among AD elasticities. It is reasonable to assume that the result represents different responses. Whereas BD estimates mainly reflect labour supply responses that are not easily to adjust, exploiting tax deductions is an easy way to change an individual's income in response to tax rate changes.

There is no statistical significant difference across BD elasticities if they are *weighted by income* or not, whereas weighted AD elasticities are significantly lower compared to unweighted ones. These chapters/etiresults are unexpected, in particular the finding for the AD subsample. If high taxpayers exhibit larger behavioural responses, weighting by income should result in higher estimates. As noted in Weber (2014), weighting by income is a controversial model choice, because income itself is endogenous and it further lead to distorted chapters/etiresults. Moreover, the chapters/etiresults obtained in primary studies are mixed. For instance, Gruber and Saez (2002) find that an unweighted gross elasticity is substantially lower to the weighted elasticity, while a weighted ETI is very similar to the unweighted ETI. Giertz (2010) on the other hand finds that unweighted ETI estimates are smaller than income-weighted estimates.

Sample Restrictions. An *age cutoff* restricts income and employment fluctuations at the beginning and end of a person's working life. Such a cutoff has contrasting effects on elasticities

depending on the subsample. Estimates in the BD subsample are lowered when a primary study restricts its data to a certain age, while I observe a positive effect on AD elasticities. *Income cutoffs* have no effect on estimated BD elasticities. This is in stark contrast to findings for the AD subsample where an income cutoff and its value matters greatly. This is an interesting finding since it is unclear whether or not a certain cutoff (and its level) helps or impairs identification.

Variations across time and countries. Column (4) and (5) take into account *country group*. While column (4) controls for *estimation decade*, column (5) shows the chapters/etiresults for *publication decade*. In both subsamples (publication) decade has a significantly larger effect on resulting estimates than (estimation (or data)) decade. Estimates published prior to 2001 are always larger than those published at a later date - even when controlling for various aspects of estimation technique. (Estimation) decade only influences BD estimates. For instance, those BD estimates that rely on a dataset that cover the 1980s are always larger than those in later years. Most of the other findings of Tables (1.2) and (1.3) discussed before prevail. Column (6) shows the chapters/etiresults of the most comprehensive specification that accounts for all the defined categories of heterogeneity. Unfortunately, multicollinearity seems to influence the chapters/etiresults to the extent that the precision of some coefficients vanishes.

Table 1.2: WLS before deductions baseline

Dependent Variable: Elasticity BEFORE deductions	(1)	(2)	(3)	(4)	(5)	(6)
Estimation Technique:						
Reg. Technique (omitted: IV: mechanical tax rate changes)						
IV: (lagged) mechanical tax rate changes	0.060*	0.061*	0.054*	0.025**	0.025*	0.026
	(0.031)	(0.033)	(0.029)	(0.012)	(0.015)	(0.017)
IV-other	0.075	0.076	0.081*	0.074	0.107*	0.094
	(0.056)	(0.053)	(0.044)	(0.054)	(0.056)	(0.062)
DID-IV	0.298***	0.269***	0.224**	0.257***	0.313***	0.247***
	(0.053)	(0.066)	(0.105)	(0.056)	(0.075)	(0.074)
DID-classic	0.332***	0.304***	0.068	0.187***	0.149**	0.091
	(0.059)	(0.072)	(0.132)	(0.063)	(0.065)	(0.068)
Income Control (omitted: Auten Carroll)						
none	-0.213***	-0.216***	-0.212***	-0.209***	-0.207***	-0.209***
	(0.024)	(0.023)	(0.025)	(0.028)	(0.029)	(0.028)
Gruber Saez Spline	-0.020***	-0.015***	-0.021***	-0.013*	-0.016***	-0.013*
	(0.005)	(0.005)	(0.007)	(0.007)	(0.005)	(0.007)
Kopczuk	-0.017**	-0.014**	-0.014**	-0.015**	-0.010**	-0.012*
	(0.007)	(0.007)	(0.005)	(0.007)	(0.005)	(0.006)
other	-0.034**	-0.070*	-0.029**	-0.020*	-0.009	-0.033*
	(0.017)	(0.039)	(0.013)	(0.012)	(0.009)	(0.019)
Difference Length (omitted: 3-years)						
1 year	0.060	0.054	0.033	0.034	0.012	0.003
	(0.063)	(0.057)	(0.045)	(0.050)	(0.040)	(0.032)
2 years	-0.013	-0.013	-0.030*	-0.035***	-0.035***	-0.041**
	(0.021)	(0.021)	(0.016)	(0.011)	(0.008)	(0.016)
4 years and more	0.082*	0.085*	0.068**	0.009	0.026	0.027
	(0.042)	(0.043)	(0.030)	(0.019)	(0.021)	(0.022)
Weighting by Income (omitted: no restriction)						
Weighting by Income applied		-0.041				-0.046
		(0.025)				(0.040)
Sample Restrictions:						
Age Cutoff applied (omitted: no restriction)						
Age Cutoff applied			-0.282**		-0.267	-0.259
			(0.122)		(0.174)	(0.168)
Income Cutoff applied (omitted: 0-10k)						
none			0.018		-0.020*	-0.023
			(0.021)		(0.010)	(0.024)
10k-12k			0.024		-0.015**	-0.015
			(0.016)		(0.007)	(0.011)
12k-31k			0.009		0.007	0.014
			(0.007)		(0.008)	(0.011)
>31k			0.021		-0.005	-0.004
			(0.017)		(0.012)	(0.013)
Variation across countries and time:						
Country Group (omitted: USA)						
Scandinavia				-0.135***	0.239*	0.176
				(0.042)	(0.123)	(0.143)
other countries				-0.020	0.343***	0.300**
				(0.051)	(0.126)	(0.127)
(Publication) Decade (omitted: 2001-2010)						
prior to 2001					0.426**	0.388**
					(0.207)	(0.191)
after 2010					-0.205***	-0.130
					(0.073)	(0.104)
(Estimation) Decade (omitted: 1980s)						
1990s				-0.048***		-0.049***
				(0.002)		(0.002)
2000s				-0.031***		-0.060
				(0.010)		(0.037)
Constant	0.073***	0.110***	0.351***	0.239***	0.296***	0.359***
	(0.007)	(0.028)	(0.123)	(0.041)	(0.054)	(0.059)
Observations	940	940	940	940	940	940
Adjusted R^2	0.566	0.575	0.615	0.637	0.655	0.680

Notes: Columns (1) to (6) estimated using WLS with the inverse of an estimate's variance as analytical weights. Reported coefficients need to be interpret as a deviation from the reference category (in bold). Baseline chapters/etiresults do not account for contextual factors. Standard errors (in parentheses) are clustered at the study level. Significance levels are * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.3: WLS after deductions baseline

Dependent Variable: Elasticity AFTER deductions	(1)	(2)	(3)	(4)	(5)	(6)
Estimation Technique:						
Reg. Technique (omitted: IV: mechanical tax rate changes)						
IV: (lagged) mechanical tax rate changes	0.409***	0.362***	0.420***	0.333***	0.207***	0.141**
	(0.088)	(0.127)	(0.061)	(0.079)	(0.074)	(0.069)
IV-other	-0.265*	-0.253	-0.246**	0.069	0.197	0.009
	(0.145)	(0.155)	(0.118)	(0.275)	(0.218)	(0.193)
DID-IV	-0.590**	-0.615***	-0.702**	-0.379	-0.289	-0.393
	(0.224)	(0.219)	(0.281)	(0.273)	(0.475)	(0.468)
DID-classic	-0.188	-0.200	-0.189	-0.061	-0.178	-0.166
	(0.372)	(0.334)	(0.363)	(0.402)	(0.305)	(0.281)
Income Control (omitted: Auten Carroll)						
none	0.108	0.074	0.045	0.084	-0.249	-0.237
	(0.078)	(0.084)	(0.089)	(0.087)	(0.159)	(0.163)
Gruber Saez Spline	-0.100	-0.007	-0.137**	-0.110	-0.119	-0.048
	(0.068)	(0.029)	(0.068)	(0.069)	(0.088)	(0.119)
Kopczuk	-0.371***	-0.231***	-0.387***	-0.229**	0.025	0.126
	(0.043)	(0.074)	(0.075)	(0.091)	(0.104)	(0.076)
other	-0.195**	-0.134	-0.331**	-0.066	0.048	0.082
	(0.075)	(0.115)	(0.132)	(0.114)	(0.124)	(0.161)
Difference Length (omitted: 3-years)						
1 year	-0.048	-0.079	0.073	-0.066	0.119	0.103
	(0.106)	(0.131)	(0.074)	(0.121)	(0.090)	(0.105)
2 years	0.033	0.045	0.019	-0.058	0.057	0.053
	(0.086)	(0.081)	(0.119)	(0.078)	(0.105)	(0.109)
4 years and more	0.285	0.187	0.182	0.139	-0.362	-0.399*
	(0.191)	(0.200)	(0.212)	(0.204)	(0.242)	(0.235)
Weighting by Income (omitted: no restriction)						
Weighting by Income applied		-0.195**				-0.208
		(0.091)				(0.160)
Sample Restrictions:						
Age Cutoff applied (omitted: no restriction)						
Age Cutoff applied			0.252**		0.140	0.187
			(0.113)		(0.124)	(0.138)
Income Cutoff applied (omitted: 0-10k)						
none			0.154***		0.254***	0.337***
			(0.054)		(0.087)	(0.058)
10k-12k			0.109		0.353	0.514**
			(0.090)		(0.236)	(0.224)
12k-31k			0.111*		0.068	0.093*
			(0.063)		(0.059)	(0.050)
>31k			0.468		0.518	0.625**
			(0.424)		(0.353)	(0.306)
Variation across countries and time:						
Country Group (omitted: USA)						
Scandinavia				-0.111	0.410	0.477*
				(0.089)	(0.305)	(0.279)
other countries				0.237	0.632**	0.608*
				(0.215)	(0.304)	(0.312)
(Publication) Decade (omitted: 2001-2010)						
prior to 2001					1.164*	1.203*
					(0.662)	(0.607)
after 2010					-0.500***	-0.417*
					(0.173)	(0.221)
(Estimation) Decade (omitted: 1980s)						
1990s				-0.018		-0.043
				(0.060)		(0.040)
2000s				-0.185		0.030
				(0.242)		(0.131)
Constant	0.445***	0.496***	0.208***	0.420***	-0.019	-0.082
	(0.040)	(0.059)	(0.066)	(0.103)	(0.272)	(0.243)
Observations	780	780	780	780	780	780
Adjusted R^2	0.405	0.423	0.479	0.425	0.621	0.633

Notes: Columns (1) to (6) estimated using WLS with the inverse of an estimate's variance as analytical weights. Reported coefficients need to be interpret as a deviation from the reference category (in bold). Baseline chapters/etiresults do not account for contextual factors. Standard errors (in parentheses) are clustered at the study level. Significance levels are * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Summary. To highlight the sensitivity of both types of elasticities with respect to the estimation technique, I calculate some stylized elasticity estimates. In Table 1.4 I present average BD and AD elasticity estimates for the most commonly used specifications. The upper part of the table considers an approach that uses mechanical tax rate changes as an instrument and a difference length of three years (= (basic) Gruber Saez approach). The lower part considers an approach that uses (lagged) mechanical tax rate changes and a difference length of two years (= (basic) Weber approach). Both parts show the chapters/etiresults for various income controls. Compared to BD elasticities the magnitude of AD elasticities is not only larger by definition but AD elasticities are also more sensitive with respect to aspects of the chosen estimation technique. Average BD elasticities lie in the range of 0.053 to 0.120, while average AD elasticities vary from 0.074 to 0.887. Richer (or more sophisticated) income controls always lower elasticities and the effect is more pronounced in the AD subsample.

Table 1.4: Stylized elasticity estimates

An approach that uses the following characteristics leads to:	BD	AD
IV: mechanical tax rate changes Difference Length of 3 years and the following income controls:		
Auten Carroll	0.073	0.445
Gruber Saez Spline	0.053	0.345
Kopczuk	0.056	0.074
IV: (lagged) mechanical tax rate changes Difference Length of 2 years and the following income controls:		
Auten Carroll	0.120	0.887
Gruber Saez Spline	0.100	0.787
Kopczuk	0.103	0.516

Notes: Stylized elasticity estimates are based on chapters/etiresults presented in column 1 in tables 1.2 and 1.3. For instance, a specification that uses (i) (lagged) mechanical tax rate changes, (ii) a difference length of 2 years and (iii) Kopczuk-type income controls provide an average AD elasticity of 0.516 = 0.445 + 0.409 + 0.033 - 0.371 (compare table 1.3). Column BD refers to before and column AD to after deduction elasticity estimates.

1.4.2 Contextual Factors

The following descriptive analysis shows how various contextual factors are associated with the size of elasticity estimates. The baseline specification involves controls for estimation technique, income controls and difference length (see column (1) from Tables 1.2 and 1.3). I use this

specification and gradually take into account contextual factors as defined in Section 1.3.2. The exercise shows that past as well as current (tax-) policy and the underlying context matters when interpreting elasticities. The relevant coefficients are displayed in Table 1.5.

Table 1.5: WLS: Contextual Factors

Dependent Variable: Elasticity:	Before Deduct.	After Deduct.
Additional Variables		
Gini Coefficient	0.008***	-0.002
	(0.002)	(0.014)
Top 10%	0.814*	3.563**
	(0.442)	(1.536)
Top 1%	0.330	7.709**
	(0.448)	(3.202)
Intro top bracket	-0.026	-0.086
	(0.094)	(0.117)
Unemployment Rates	-0.007	0.067*
	(0.004)	(0.039)
Fraction of self-employed	0.016***	-0.022
	(0.006)	(0.023)
Modern taxes (in 2005)	-0.010***	0.016
	(0.002)	(0.012)

Notes: Both columns are estimated using Weighted Least Squares with precision as weights. Standard errors (in parentheses) are clustered at the study level. The baseline specification only includes controls for estimation technique (regression technique, income control and difference length) (same as column (1) from Tables 1.2 and 1.3. I gradually add each contextual characteristic separately. For the first characteristic, I compare the first and last year of a data period. Remaining characteristics are merged via mean year of observation. For observations that are based on a classic DID approach, I do not have information of the share of self-employed people that corresponds to the respective mean year of observation. Full chapters/etiresults can be found in the (online) Appendix E (see Tables 15 and 16). Significance levels are * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

There is a positive correlation between *inequality measures* and elasticities. In particular, AD elasticities are highly correlated with top income shares. An increase in inequality might be the result of past tax cuts for high-income taxpayers. As Alvaredo et al. (2013) observe, there has been a widening of the income distribution and top tax rates have moved in the opposite direction from top pre-tax income shares. While top pre-tax income shares are rising, top tax rates are decreasing. Such widening in the income distribution affects estimated elasticities. It might be the case that income control variables do not fully account for such a development and this leads to an upward bias of AD elasticities. This confirms the fact that not only current but also past tax policy still has an effect on estimated elasticities and that the underlying context matters when interpreting elasticities.

Given that wealthier people tend to be more responsive, I expect a positive relationship between an *introduction of a top tax bracket* and behavioural responses. Contrary to my expectation,

the coefficient is insignificant and close to zero.[15] Business cycle effects are approximated by *unemployment rate* are weakly related to AD elasticities and I do not find any correlation with BD elasticities.

As shown by Kleven et al. (2016*a*), there is a close relationship between tax enforcement, tax compliance and third party information reporting. My regression chapters/etiresults show that the share of tax revenue that are exposed to third-party information reporting within a country (*modern taxes per GDP*) is negatively related to BD elasticities. Given that self-employed people have greater control over their income, there is a positive correlation between BD elasticities and the *fraction of self-employed* workers in an economy. Neither measure influences AD elasticities. This strengthens the fact that AD responses are mainly driven by avoidance behaviour. Most taxpayers respond via itemized deductions that are not subject to third party information reporting. The magnitude of estimated elasticities are affected by the degree of third party information reporting which can be influenced by policy makers.

1.5 Selective Reporting Bias

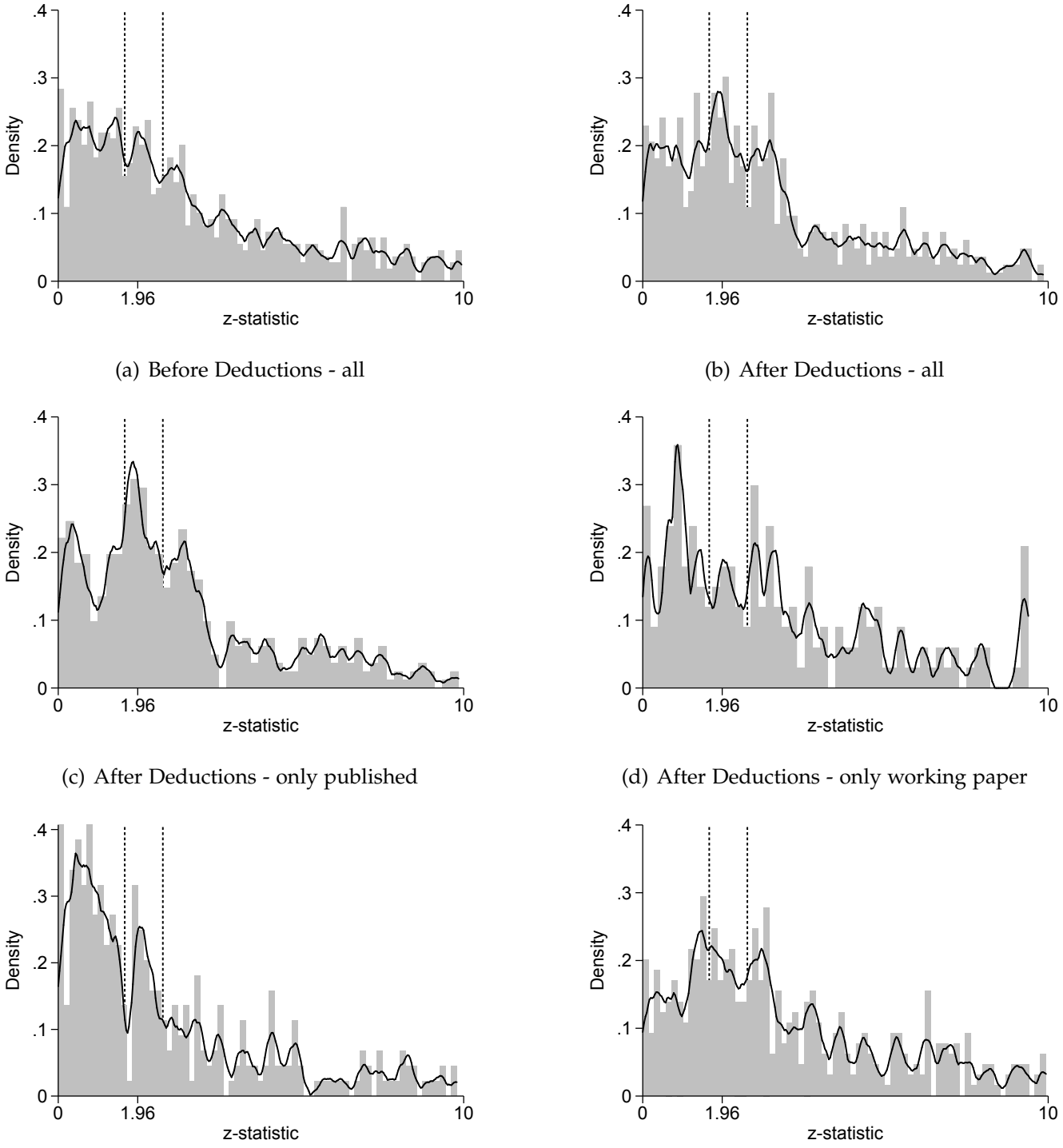
In the last part of my analysis, I check for the presence of a selective reporting bias. Publishing statistical chapters/etiresults that reject the hypothesis of no effect reflects a general desire. Moreover, researchers naturally want to publish chapters/etiresults that exhibit intuitive magnitudes. Publication or reporting selection bias has been identified in other areas of empirical work. Ashenfelter et al. (1999) review the literature on the rate of return on schooling investment and show reporting selection bias in favour of significant and positive returns to education. Card and Krueger (1995) find such biases in the minimum wage literature and Lichter et al. (2015) in the literature on labour demand elasticities. A study by Brodeur et al. (2016*a*) uses more than 50,000 tests published in three top economic journals and find that researchers are prone to choose more 'significant' specifications in order to increase the chance of publication. Moreover, they show that scientists use z-statistics of 1.64 or 1.96 as reference points.

To start the analysis, I follow Brodeur et al. (2016*a*) and plot the distribution of z-statistics and, I then examine the relationship between standard errors and estimates and the distribution of elasticity estimates. Finally, I check statistically whether publication bias is prevalent.

Distribution of z-statistics An obvious type of bias is the excessive production and selection of significant chapters/etiresults. Given that z-statistic = beta coefficient/standard

[15] I ignore all other tax system-related issues (e.g. base broadening) that might have been occurring simultaneously.

Figure 1.2: Raw distribution of z-statistics



(a) Before Deductions - all

(b) After Deductions - all

(c) After Deductions - only published

(d) After Deductions - only working paper

(e) Before Deductions - prior to Chetty (2009)

(f) Before Deductions - after Chetty (2009)

Notes: All graphs plot the distribution of z-statistics. The significance level of 5% (1.96) and also the z-values for the 10% and 1% level of significance are highlighted. Subfigure (a) plots all estimates from the Before Deductions (BD) subsample and Subfigure (b) for the After Deductions (AD) subsample. Subfigures (c) and (d) split the AD subsample into estimates published in journals and estimates reported in working papers. Subfigures (e) and (f) split the BD subsample into estimates that are published prior to and after 2009.

error, there are three ways to receive significant values. First, to find a specification where standard errors are low enough. Second, to search for a specification where coefficients are large enough to offset 'large' standard errors. Or third, through a combination of these two things. Since research on behavioural responses to taxation relies on administrative datasets with a large number of observations, standard errors are generally small.

I plot the distribution of z-statistics for the two subsamples (see Figure 1.2).[16] Subfigure (a) shows the BD and Subfigure (b) the AD subsample. In accordance with Brodeur et al. (2016*a*), I observe a local maximum around 2 (= 5% significance) and also a valley before this. Moreover, I also observe a spike around 1.64 (= 10% significance) and around 3 (= 0.05-0.01% significance).[17] These simple graphs provide evidence consistent with the existence of 'p-hacking.' This pattern is more pronounced in the AD subsample because researchers usually use the elasticity of taxable income (and not necessarily the elasticity of broad income) when they apply optimal tax rate formulas.

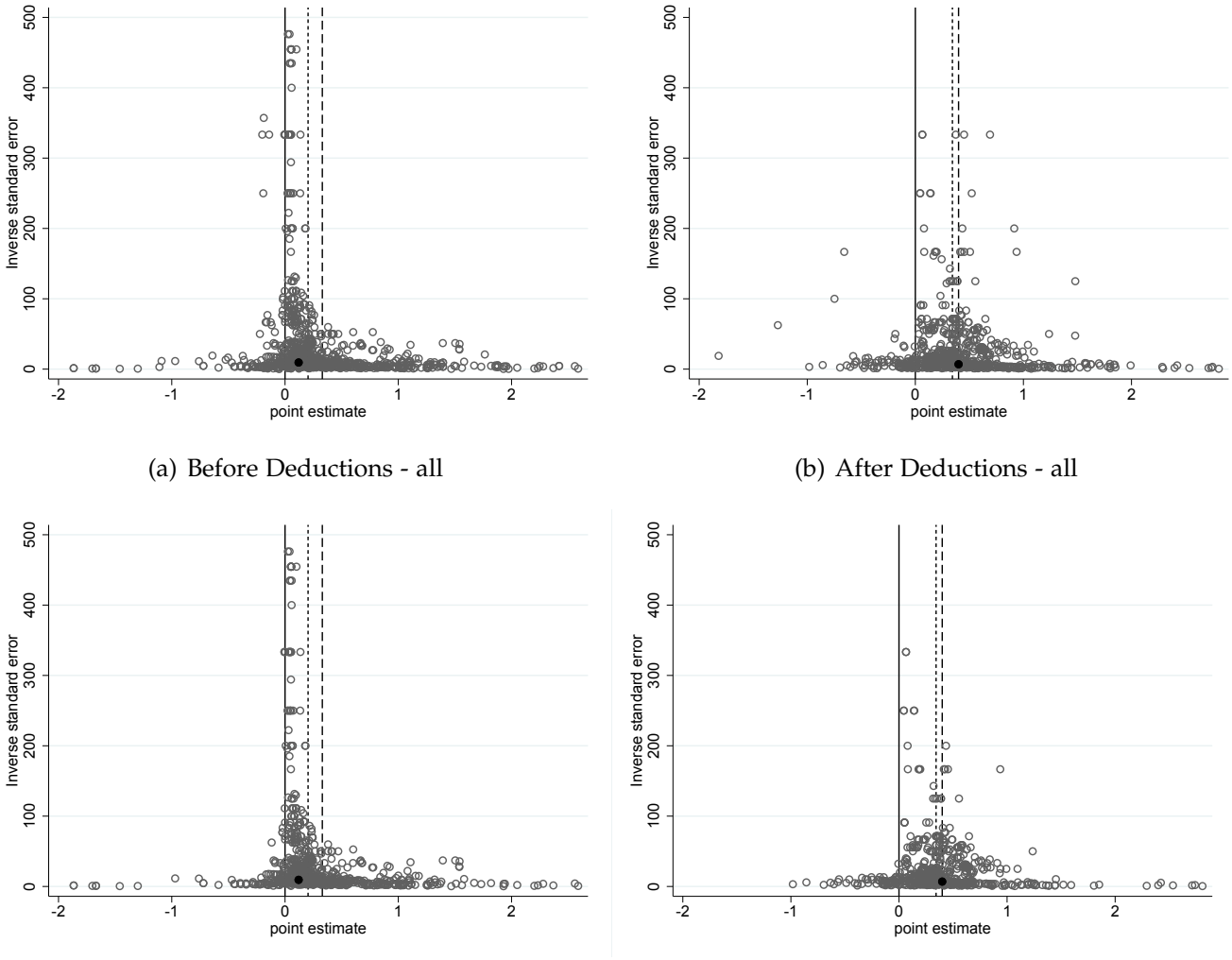
In Subfigure (c) and (d) I divide the AD subsample into estimates reported in journal articles and working papers. The maximum around 2 is even more pronounced for published AD elasticities. It is unclear whether a researcher chooses the most credible findings in the first place to increase the chances of publication and/or that referees/journals prefer significant estimates. Moreover, journal editors often require authors to streamline their papers prior to publication, leading them to limit the number of tables and figures in their paper. Therefore, it is unclear who chooses which estimates are published.

Chetty (2009) shows that the excess burden of taxation depends on a weighted average of taxable income and total earned income elasticities. Since the publication of his study in 2009, BD (e.g. gross income) elasticities have begun to receive more attention. Therefore, I divide the BD subsample into estimates reported prior to and after 2009. As seen in Subfigure (e), I observe a larger insignificant mass before 2009 and a huge spike at 1.96 (=5% significance level) and a missing mass before. After 2009 I observe a much smaller insignificant mass but still a spike at 1.64 (=10% significance), 1.96 (=5% significance) and now also around 3 (=0.05-0.01% significance level). The graphical evidence confirms that the share of significant BD elasticities has increased over time.

[16] I formally tested the equality across distributions. I applied a Kolmogorov-Smirnov test which tests whether different t-distributions are equal. More specifically, I test (i) whether the t-statistics of before and after deduction distribution elasticities differ, (ii) within the AD subsample, I check whether the distribution of t-statistics from published estimates and estimates collected from working papers differ, and last (iii) within the BD subsample, I check whether the distribution of t-statistics before and after 2009 differ. In all three cases, I am able to reject the hypothesis that this is the case.

[17] There are other peaks and valleys across the distributions. Unlike Brodeur et al. (2016*a*) I use considerably fewer observations, with the result that my graphs appear to be more 'bumpy.'

Figure 1.3: Funnel Plot



(a) Before Deductions - all

(b) After Deductions - all

(c) Before Deductions - only income control(s)

(d) After Deductions - only income control(s)

Notes: Funnel plots are presented separately for the before and after deductions subsamples. The short dashed line denotes the median and the dashed line the mean of the corresponding (full) subsample. In the dataset the median (mean) BD elasticity is 0.185 (0.287) and 0.353 (0.403) respectively for an elasticity that considers deductions. The base chapters/etiresults from Gruber and Saez (2002) are highlighted in black. They report coefficients of 0.4 with a standard error of 0.144 for the ETI and 0.12 with a standard error of 0.106 for the elasticity of broad (=gross) income. Subfigures (a) and (b) display all collected estimates. Subfigures (c) and (d) are based on a subset of estimates that rely on a specification with income control(s).

Relationship between estimate and standard error. In the second step, I follow Card and Krueger (1995) and analyse the relationship between an estimate and its standard error. I apply a standard procedure and use what is known as a funnel plot in order to analyse the correlation. Funnel plots are simple scatter plots of elasticity estimates on the horizontal axis and their precision (=inverse of standard error) on the vertical axis. The most precise estimates are close to the top of the funnel and as precision decreases, the dispersion of estimates increases.

The shape of the graph should look like an inverted funnel. In the absence of selective reporting bias, there should be no systematic relationship between estimates and standard errors. All imprecise estimates should have the same probability of being reported. The funnel should be symmetric with the estimates randomly distributed around the population elasticity. If the estimates are correlated with their standard errors, the funnel can take an asymmetric shape. This might happen when researchers select only significant estimates and/or estimates with a certain sign (e.g. omit negative values) such that their chapters/etiresults are consistent with theory.[18]

Figure 1.3 plots BD and AD elasticities separately along with their precision. I highlight the mean and median as well as estimates obtained by Gruber and Saez (2002). Subfigures (a) and (b) are based on the full sample of estimates, while I restrict the sample to estimates that rely on income controls and therefore explicitly account for non-tax related income growth in Subfigure (c) and (d). Subfigures for BD and AD reveal some noticeable differences. First, I observe a more pronounced missing mass on the negative side in the BD compared with the AD subsample. According to theory an increase in the marginal tax rate lowers the net of tax rate, which in turn should reduce taxable income in the simplest case with no income effects or frictions. If a researcher receives a negative value, this translates into a situation where the government can tax income by 100% while the people earn/work even more. Hence, it seems plausible that researchers tend to put more trust in positive chapters/etiresults to keep in line with theory. This behaviour causes a positive relationship between standard errors and estimates. AD elasticities allow a wider range of responses and it is also well-documented that running the exact same specification chapters/etiresults in a larger AD elasticity compared to an BD elasticity (Gruber and Saez (2002)). The chance of reporting negative values is therefore larger for an elasticity that does not consider deductions. This might explain why I observe a larger missing mass on the negative side in the BD subsample.

Within the AD subsample, it appears that researchers tend to report an estimate between 0 and 0.4 with a higher probability compared to estimates ranging from e.g. 0.4 to 0.8. I expect a negative relationship between standard errors and estimates and therefore a downward bias of AD estimates.

Distribution of estimates. Another kind of selection reporting bias arises, if researchers use well-known chapters/etiresults as a reference point and hence are inclined to report only chapters/etiresults that are in line with these findings. Piketty and Saez (2013) write in their

[18] As well as a graphical analysis, I formally checked for funnel asymmetry and conducted a so-called Funnel-asymmetry test as proposed by Egger et al. (1997). In all cases, I am able to reject the hypothesis of funnel symmetry. Besides selective reporting bias, there are other reasons why funnel asymmetry could arise (e.g. data irregularities or low methodological quality of some studies).

handbook chapter that an elasticity of 0.25 seems realistic (same as Chetty (2009)), 0.5 is high and 1 is extreme. As seen in Figure 1.1, there is a general tendency to report chapters/etiresults that lie within an interval of 0 and 1. I observe a considerable excess mass between 0.7 and 1. This indicates an aversion to report a value above 1. In their well-known and widely-cited survey, (Saez et al., 2012, p. 42) refer to their estimates and write '[...]. While there are no truly convincing estimates of the long-run elasticity, the best available estimates range from 0.12 to 0.4. [...]' and '[...] 0.25 corresponds to the mid-range of estimates found in the literature. [...]' With regard to the AD-funnel, there is a slight incline to report values between 0 and 0.4 (=mean of AD estimates in the dataset).

Regression Results. To statistically examine the presence of selective reporting bias, I take specification of column (1) of Tables 1.2 and 1.3 as the baseline specification (=WLS with estimation technique controls) and explicitly control for an estimate's standard error and other publication-related characteristics. Point estimates and respective standard errors should be independent according to random sampling theory (Card and Krueger, 1995; Stanley and Doucouliagos, 2010). For the sake of interpretation, I normalize the standard error.
Overall, my regression chapters/etiresults confirm what can already be seen in figures presented before. The funnel plot for BD estimates indicates selective reporting bias towards positive elasticities. This is confirmed in column (1). Published AD estimates suffer more from 'p-hacking' and I statistically show that selective reporting bias is even more pronounced in journals with a high impact factor among AD elasticities (see column (6)).[19] To account for the fact that larger datasets increase the change of yielding standard errors that are small enough to produce significant and trustworthy chapters/etiresults, I calculate the median of observations for each subsample and create a dummy variable if an estimate is based on a dataset that is smaller or larger compared to the median sample size of all other collected estimates. For BD elasticities, the relationship is significantly positive (see column (3)). In columns (4) and (8) I include a dummy variable indicating if an estimate was reported prior to Chetty (2009). Both aspects influence BD but not AD elasticities.

Summary. The graphical evidence and regression chapters/etiresults indicate an upward reporting bias among BD elasticities, while the reporting bias for AD elasticities goes in both directions with a downward bias appearing to be dominate. The distribution of elasticities an the funnel plot show that there is a tendency to report chapters/etiresults that lie within an interval of 0 and 1. In general, reference points related to statistical significance such as 1.96 matters for both types of elasticities and well-known chapters/etiresults are targeted.

[19] I downloaded the IDEAS RePEc simple impact factor (22.06.2016) and working papers receive a value of 0.

Table 1.6: Testing for Selective Reporting Bias

Dependent Variable:	BD	BD	BD	BD	AD	AD	AD	AD
Elasticity:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Standard Error	3.654***	4.084***	0.972	0.652	-0.030	-0.834***	-0.223	-0.360
	(0.719)	(0.845)	(0.812)	(0.988)	(0.203)	(0.294)	(0.354)	(0.530)
Journal impact factor		-0.012				0.030**		
		(0.008)				(0.014)		
Std.Error* Impact Factor		-0.051				0.084***		
		(0.035)				(0.022)		
Dummy if obs > median(obs)			0.771***				-0.066	
			(0.279)				(0.285)	
Std.Error*D if obs > median(obs)			4.375***				0.113	
			(1.142)				(0.540)	
Dummy reported prior to 2009				0.575**				-0.1122
				(0.267)				(0.304)
Std.Error*D reported prior to 2009				3.726***				0.217
				(1.322)				(0.614)
Constant	0.876***	0.982***	0.477***	0.460**	0.424**	-0.027	0.400**	0.416*
	(0.180)	(0.213)	(0.138)	(0.181)	(0.158)	(0.221)	(0.158)	(0.248)
Observations	940	940	940	940	780	780	780	780
Adjusted R^2	0.614	0.624	0.628	0.627	0.404	0.456	0.408	0.420

Notes: Columns (1) to (8) are estimated using Weighted Least Squares using precision as weights. I control for estimation technique (= regression technique, income control and difference length. Full chapters/etiresults can be found in the (online) Appendix G in Tables 19 and 20. Standard errors (in parentheses) are clustered at the study level. Significance levels are * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Included standard errors as explanatory variables are normalized such that they can be interpreted as a standard deviation.

In particular, I observe a larger missing mass for negative values in the BD subsample and I find that researchers report AD estimates ranging from 0 to 0.4 more often compared to chapters/etiresults that are located e.g. within 0.4 to 0.8. Among the AD subsample selective reporting bias is even more prevalent in journals with a high impact factor, while the year of publication matters for BD elasticities. Since 2009 have become more significant because of its increased interest.

1.6 Conclusion

This study applies meta-techniques to identify and to assess different explanations for the varying sizes of estimated elasticities. The magnitude of such estimates is of major importance for tax policy analysis. I differentiate between real responses (before deduction elasticities) and avoidance behaviour (after deduction elasticities) and use 1,720 estimates from 61 studies.

The paper consists of three parts. First, I conduct a meta-regression analysis and quantify the impact of various model choices. Compared to BD elasticities the magnitude of AD elasticities is not only larger by definition, but AD elasticities are also more sensitive with respect to the estimation technique. Second, my study points to correlations between reported estimates and tax system- and economy related characteristics, as well as inequality measures. Last, it shows that selective reporting bias is prevalent in the literature of taxable income elasticities. There

is an upward reporting bias among BD elasticities while the reporting bias for AD elasticities goes in both directions with a downward bias appearing to be dominate.

Several important conclusions can be drawn from this analysis. As already acknowledged in the literature, the ETI is not a structural parameter and this study shows that policy conclusions can be misleading. Reported estimates need to be interpreted within the context they are estimated in and researchers and policy makers need to be careful about what type and size of elasticity should be used for policy analysis (e.g. when calibrating an optimal tax model). An application of a simple formula to derive optimal revenue maximising top tax rates, lead to tax rate of 62.5% if I incorporate the mean AD elasticity of 0.403 found in the empirical literature. Using my derived stylized AD estimates ranging from 0.074 to 0.827, lead to tax rates between 44.63% to 90.01%.[20] To develop new (empirical) strategies that are robust to certain model choices, we need to raise the awareness that insignificant and even implausible estimates are meaningful. Instead of proving a single estimate, a range of estimates might help to shed light into the heterogeneity of behavioural responses across the income distribution and different socio-economic groups.

Finally, the literature on taxable income elasticities suffer from selective reporting bias. Unlike the literature on the effects of taxation on labour supply, which relies mostly on survey data, the ETI-literature predominately uses administrative tax-return data.[21] On the one hand, administrative tax-return data provides precise information about a tax unit's income situation that is needed for estimation but, on the other hand, a replication of existing findings is very difficult. Data access is often restricted to a small number of people and its utilisation is costly in various dimensions (e.g. lack of institutional knowledge and language barriers). Future researchers should be encouraged to provide as much information as possible to promote a comprehensive understanding of the obtained elasticities (Slemrod (2016)). Reporting standards or even a pre-analysis plan might reduce the problem of selective reporting bias (see Burlig (2018); Christensen and Miguel (2018)).

[20] Assume that the shape of the income distribution in the highest tax bracket is characterised by the Pareto parameter a and e is the elasticity of taxable income or the range of AD elasticities found in this study. Following Saez (2001) the revenue-maximising tax rate is defined as $t = \frac{1}{1+a*e}$. For instance, if $a = 1.5$ and $e = 0.074$, the resulting tax rate is equal to $t = \frac{1}{1+1.5*0.074} = 90.01\%$.

[21] There are some exceptions who either use survey or aggregated administrative data. Recently, Burns and Ziliak (2017) use the Current Population Survey for the US and find elasticities in the range of 0.4 and 0.55. Although deductions and exemptions are precisely measured in administrative tax records, survey data offers a larger set of demographic characteristics and information about the low end of the income distribution. Tax units who do not file a tax return are not available in the tax data and these tax units are in most cases poor households. Future work might consider survey data to (at least) estimate BD elasticities. Saez (2017) provides evidence that even simple tabulated tax data can provide valuable evidence and he points out to possible advantages of such data compared to microlevel data (e.g. simplicity and transparency).

P-Hacking, Data Type and Availability of Replication Material

2.1 Introduction

In the last two decades empirical microeconomics has experienced a 'credibility revolution' (see Angrist and Pischke (2010). Various improvements in empirical work has been made. The greater availability of administrative records has had a great impact on applied microeconomics research. Indeed, Chetty (2012) documents an important rise in the number of micro-data-based studies published in top economic journals between 1980 and 2010 that used administrative records data. The advantages of administrative records include, among others, providing highly reliable and representative data covering a large number of observations (see Künn (2015) for a discussion). However, administrative data comes with potential costs such as limited access because of privacy concerns. As Christensen and Miguel (2018) document the first 'top five' general interest economics journals explicitly require data and code to be submitted at the time of article publication. This allows to replicate and to verify the underlying results. In addition, it also expands scientific knowledge because other researchers are now able to build upon existing (empirical) research. The increasing fraction of papers relying on administrative data sources might explain the rising share of papers that received exemptions from the data-sharing policy (Vilhuber (2020)). Given the lack of replication files, it may hamper the ability of other researchers to duplicate the results of a prior study if the same procedures are followed.[1]

In this paper, we investigate the relationship between methods of data collection (administrative versus survey, for instance), availability of replication material and statistical significance. While a large literature documents the extent of p-hacking (i.e., manipulation or selective

[1] Availability of replication material is important for reproducibility of results, but also replicability (i.e., replicating prior results using the same codes but new data) and generalizability (i.e., extension of findings to other populations or settings). See Bollen et al. (2015) for definitions and a discussion of reproducibility and replicability.

reporting of *p*-values) and publication bias (i.e., outcome and statistical significance of a study are related to the decision to publish) in economics and other disciplines (Andrews and Kasy (2019); Brodeur et al. (2016*b*); Bruns et al. (2019); DellaVigna and Linos (2020); Doucouliagos and Stanley (2013); Franco et al. (2014); Furukawa (2019); Gerber and Malhotra (2008*a*); Gerber and Malhotra (2008*b*); Havránek (2015*b*); Havránek and Sokolova (2020); Rosenthal (1979*b*)), the question of whether specific methodologies for data collection suffer from more selective reporting has not received a great deal of attention. This is a key research question given the increasing accessibility (and use) of administrative and proprietary data in general, and the fact that publication bias and p-hacking issues cast doubt upon the credibility of published research in the eyes of policymakers and others. If policymakers and citizens are more likely to see studies finding a significant effect of a given policy, then this would lead to a misrepresentation of the policy's real effect (Blanco-Perez and Brodeur (2019)).

The main hypotheses to be tested are: (1) the extent of p-hacking and publication bias in leading economics journals depend upon the methods of data collection, and (2) journal articles which provide data and/or code for replication suffer from less selective reporting.

To answer these research questions, we rely on the universe of hypothesis tests reported in journal articles using experimental and quasi-experimental methods published in 25 top economics journals for the years 2015 and 2018. We complement this data set, built by Brodeur et al. (2020), by collecting information about the data set used and distinguish between four different types of data: *administrative* data (admin), *survey*, *hand collected* and *other* data. We also collect data on whether the data *and* code or at least the code for replication for each journal article were made available on the website of the journal.

We find that the distribution of tests for admin and survey data exhibits a two-humped shape, with "missing" tests just before the 10% significance thresholds, and a "surplus" just after. There is a (local) maximum near the 5% significance threshold for both subsamples and the extent of misallocation is remarkably similar across these two data types. The extent of misallocation appears slightly smaller for hand collected data and slightly larger for other data types such as financial data.

We rely on three approaches to formally document the extent of p-hacking by data collection type. First, we follow Gerber and Malhotra (2008*b*) and apply a caliper test. This method focuses on discontinuities in the probability of a test statistic appearing just above or below a conventional statistical threshold. Our results suggest that the proportion of tests that are marginally significant in admin data articles is not significantly different than for survey data. We also provide weak (although not robust) evidence that hand collected data are less likely to report marginally significant estimates than admin and survey data.

One of the advantages of the caliper test is that we can control for journal fixed effects,

and authors and articles' characteristics.[2] This is potentially important in our context if users of a specific data type have characteristics that are related to specification search behavior.[3] Interestingly, we find that controlling for a large set of authors and articles' characteristics in our caliper analysis has no effect on our conclusions.

Second, we follow Brodeur et al. (2020) and quantify the excess number of test statistics in the range 1.65 to 2.58 by comparing the observed distribution of test statistics for each data type to a counterfactual distribution absent of selective reporting. The main advantage of this method is that we can directly measure the extent of misallocation without having to compare between data types. Our results suggest that the extent of misallocation is relatively small for all data types. Survey data appears to suffer from the largest amount of misallocation.

Last, we apply Andrews and Kasy (2019)'s measurement of publication bias. Their methodology provides the relative publication probability of statistically significant results in comparison to statistically insignificant results. For admin data, we find that a statistically significant result at the 10% and 5% levels is approximately 2.4 and 3.4 times more likely to be published than an insignificant result, respectively. The estimates are slightly larger for survey data. In contrast, statistically significant hand collected results (at the 10% and 5% levels) are only 1.7 and 2.1 times more likely to be published than an insignificant result, respectively.

We then turn to testing whether providing data and/or codes for replication is related to selective reporting. While the primary goal of mandatory data and code sharing policies is not to decrease selective reporting, but rather to increase reproducibility of empirical results, it is plausible that the availability of replication material could be related to selective reporting. One potential disadvantage of administrative (admin) data over other data type from a research transparency perspectives is thus the relative difficulty of data access for other researchers. In our sample, approximately 34% of tests are in journal articles which provide access to data and code for replication. This result is partly driven by a rise in the share of papers receiving exemptions from data-sharing policies at top journals.[4] More precisely, about 22%, 34% and 50% of journal articles using admin, survey and hand collected data provide direct access to data and code, suggesting large differences in data sharing.

Another way to promote research transparency is to provide at least the code that is used by

[2]We find that articles published by less experienced authors and with a higher share of women are significantly less likely to rely on admin data than other data types, while survey data are less likely to be published in one of the top 5 economic outlets.

[3]See Kapteyn and Ypma (2007) for a discussion of issues and problems with survey and admin data. The authors point out, for instance, that surveys are more costly and subject to nonresponse issues, while admin data may suffer from mismatching due to to imperfect linkage information from different sources.

[4]For example, Christensen and Miguel (2018) show that from 2005 to 2016 the proportion of journal articles published in the *American Economic Review* using data that received exemptions from the data-sharing policy has increased from 6 percent to 46 percent. See Vilhuber (2020) for a discussion of the importance and impact of restricted-access data.

the researchers. While the code itself does not allow a replication, it might improve transparency and replicability, and thereby reduce p-hacking. We observe that roughly half of tests are published in journal articles for which the underlying code was provided, with no striking differences by data type.

Using our different methods, we find no evidence that articles that provide data and/or codes for replication are significantly less likely to report marginally significant results. This result is robust to the inclusion of journal fixed effects and a large set of control variables. This finding is in line with our main result that the extent of selective reporting does not vary across data type, and potentially suggests that improved publication practices other than data and code availability policies are necessary to mitigate the problems of publication bias and p-hacking. For instance, Blanco-Perez and Brodeur (2020) provide evidence that an editorial statement on the importance of non-significant results addressed to authors and reviewers reduced the extent p-hacking and publication bias in the field of health economics.

Our results contribute to a growing literature on meta-analyses and research transparency by better informing the determinants of publication bias and p-hacking (Abadie (2020); Havránek et al. (2020); Ioannidis et al. (2017); Miguel et al. (2014); Stanley (2008); Stanley and Doucouliagos (2014); Swanson et al. (2020)).[5] Two relevant studies are Brodeur et al. (2020) and Vivalt (2019). They document differences in selective reporting by method, showing empirical evidence that randomized control trials and regression discontinuity designs in comparison to other non-experimental methods are less p-hacked. We contribute to this literature by testing other potential determinants of selective reporting in economics.

We also contribute to a literature on journal policies, and more generally editor and reviewer behavior.[6] In a recent literature review, Christensen and Miguel (2018) identify different approaches to address the credibility of research findings such as mandating greater data sharing and the use of pre-analysis plans.[7] Brodeur et al. (2016*b*) document for three top economics journals that data or programs availability does not mitigate p-hacking. We contribute to this literature by analyzing this relationship for a larger number of journals, and controlling for journal fixed effects and authors and articles' characteristics in our model.

Section 2.2 details the data collection. Section 2.3 documents differences in the users of admin, survey and hand collected data, and tests whether the likelihood of providing replication material is related to data types. In section 2.4, we present the distribution of test statistics by

[5]See Camerer et al. (2016), Chang and Li (Forthcoming), Hamermesh (2017) and Maniadis et al. (2017) among others for a discussion of replication in economics.

[6]See Card and DellaVigna (2020), Card et al. (2020) and Carrell et al. (2020) for two recent studies documenting how reviewers evaluate papers and whether editors follow reviewers' recommendations. See Blanco-Perez and Brodeur (2020), Feige (1975) and Höffler (2017) for comments on editorial policies.

[7]See Christensen et al. (2019) and McCullough et al. (2008) for a discussion of the benefits and limitations of data sharing.

data type. Section 2.5 presents our main findings. The last section concludes.

2.2 Data Collection

Our data come from Brodeur et al. (2020) and contain 21,440 test statistics from 684 articles published in 2015 and 2018 in 25 top economics journals.[8] The sample consists of journal articles using one of the following four methods: difference-in-differences (DID), instrumental variable (IV), sharp regression discontinuity design (RDD) and randomized control trial (RCT).[9] The data consists of coded z-statistics for all tests using the ratio of coefficients and, standard errors,[10] p-values transformed into equivalent z-statistics and t-statistics for each journal article. The sample was restricted to coefficients of interest from main results tables. Estimates from summary statistics, appendices, robustness checks, and placebo tests were excluded. Test statistics drawn from multiple specifications of the same hypothesis were collected. Of note, each article was independently coded by two of the authors to reduce concerns that only coefficients of interest were selected. All of the tests relate to two-tailed tests.

We augment the data by Brodeur et al. (2020) by adding information on data sets used and replication characteristics for each article. More specifically, we collect information on the method of data collection and the name of the data set. The median number of employed data sets per study is two. In case of multiple data sets within one journal article, we only used those articles that rely on solely one data type. Hence, we follow the most conservative approach and only consider those observations where the data type under study is clearly identifiable and unique for each article..[11] In Section 2.4 we show that the omission of journal articles using multiple types of data has no effect on our main conclusions.

Type of Data We collect information about the data set used in primary studies and distinguish between four different types of data: (a) admin, (b) survey, (c) hand collected and (d) other data.

Administrative (or register) data are generally collected by government agencies used for administrative purposes. Typical examples are social security or vital records. Compared to administrative data, *survey data* differ in terms of their purposes. Surveys are conducted to

[8] Top journals are identified using RePEc's Simple Impact Factor: `https://ideas.repec.org/top/top.journals.simple10.html`.

[9] Journal articles using matching, fuzzy RDD or Structural Equation Model are removed.

[10] We treat the ratios of coefficients and standard errors as if they were following an asymptotically standard normal distribution under the null hypothesis.

[11] This decision is also related to the fact that studies with multiple data types typically rely on multiple types for a given specification (e.g., dependent variable uses admin data while independent variables use survey data), making it tricky to code.

answer specific questions, while often targeting only subgroups of individuals. For example, a candidate survey conducted during an election. Two more prominent examples are the Current Population Survey (CPS) and the General Social Survey (GSS). These data are gathered by a third party and not by the researchers themselves. *Hand collected data* on the other hand, describe data sets that are manually collected by researchers. Such data might be an own implemented survey or experiment. We coded all remaining data sets as *other*. This involves data collected from financial data streams such as Bloomberg or Compustat but also statistical data like GDP or unemployment figures that are publicly available and provided by the OECD or World Bank. Appendix Table B.1 provides various examples of data sets by type of data.

Replication Characteristics We coded two characteristics to classify possibilities to replicate published findings: (a) direct access to the data and code and (b) provision of code. To specify how the respective data set(s) can be accessed, we coded *access* as a binary variable that indicates if the relevant data *and* code can be directly accessed on the journal's webpage.[12] Even if journals require the authors of a study to publish their code and data, authors face various restrictions. Often, a data set cannot be published due to confidentiality reasons (i.e., tax return data). To capture the availability of code, we coded *provision of code* as a binary variable that indicates if at least the underlying code (or data) can be accessed.[13] More precisely, we compare estimates that provide no replication material to those estimates that provide replication material. We collect this information on the journals' webpages and thereby ignore the voluntary provision of replication material (either code, data or both) on authors' webpages.

Descriptive Statistics In Table 2.1, we provide an overview about the type of data and replication characteristics. In total we identified 12,495 observations and 402 articles that rely on solely one data type. The largest share of articles collect the data themselves (29%), while approximately 28% employ *admin*, 20% *survey* and 23% use *other* data.

In our sample, 130 out of 402 journal articles provide direct access to the *data and code*. More specifically, we document that 22%, 34% and 50% of journal articles using admin, survey and hand collected data provide direct access to data and code. This result could be driven by a larger share of admin journal articles receiving exemptions from data-sharing policies at top journals and/or through composition effect in which admin data papers are more likely to be published in journals that do not have data and code availability policy. We provide empirical

[12]We only checked the journals' webpages for data and code availability. It is plausible that some researchers publish replication material on their personal webpage in the absence of a journal data and code repository.

[13]There are two papers (i.e., 27 observations) in our sample which provide data but not the code for replication. We code the variable *provision of code* to be equal to one for these 27 observations.

evidence throughout that exemptions from data-sharing policies is driving this result.[14]

In our sample, approximately half of the articles provide at least the codes for replication. In contrast to direct access to data and code, we do not observe large differences across data type for the provision of code.

Table 2.2 provide descriptive statistics for articles and authors' characteristics by type of data. We see that about 20% of tests in our sample come from articles published in the Top 5 journals in economics,[15] of which the majority are using admin or hand collected data. Approximately 21% of tests come from solo-authored articles, with a majority of those tests from admin and survey data. The average years of experience of authors (years since PhD completion) in our sample is 11.6 and about 30% of tests are written by authors affiliated with a top institution.[16]

In Appendix Table B.2, we provide an overview of journals and the prevalence of our four different types of data.[17] The journals with the largest number of tests in our sample are the *American Economic Journal: Applied Economics*, the *American Economic Review*, the *Journal of Development Economics*, the *Journal of Public Economics* and the *Journal of Finance*. There are large differences in data types by journal. In our sample, the share of articles using admin data is especially large for the *Journal of Public Economics*, the *Journal of Urban Economics* and the *Quarterly Journal of Economics*, and relatively low for journals such as the *Journal of Human Resources* and the *Journal of Development Economics*. The share of articles using hand collected data is relatively high for the *Journal of Development Economics*, the *Journal of the European Economic Association* and the *Review of Economic Studies*. The three journals with the highest share of articles using survey data are the *American Economic Journal: Macroeconomics*, the *Journal of Applied Econometrics*, the *Journal of Human Resources*. Last, the share of articles using other types of data is the largest for the *Journal of Economic Growth*, *Journal of Financial Economics* and *Journal of Financial Intermediation*.

Last, Appendix Table B.3 shows how often journals (i) provide direct access to the underlying data *and* code and (ii) provide at least the code. Each observation is a test statistic.

[14] See our caliper test analysis with journal fixed effects presented in Section 2.5 and Table 2.2 for journal composition by method of data collection.

[15] The Top 5 journals refer to the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics*, and the *Review of Economic Studies*.

[16] We follow Brodeur et al. (2020) and code as "top" institutions the following institutions/departments: Barcelona GSE, Boston University, Brown, Chicago, Columbia, Dartmouth, Harvard, MIT, Northwestern, NYU, Princeton, PSE, TSE, UC Berkeley, UCL, UCSD, UPenn, Stanford, and Yale. The choice of institutions is based on RePec's ranking of top institutions (`https://ideas.repec.org/top/top.econdept.html`).

[17] Appendix Table B.4 shows the distribution of different data type usage by estimation method. The most striking result is perhaps for hand collected data with over 90% of RCT tests using this data collection technique, while only 2 and 10 and 13% of RDD, DID and IV tests use hand collected data.

2.3 Data Type Characteristics

Before turning to our analysis of p-hacking across data types and availability of replication material, we document key differences in the users of admin, survey and hand collected data, and test whether the likelihood of providing replication material is related to data types.

2.3.1 Data Type and Articles and Authors' Characteristics

We first conduct an analysis in which we predict data type use with our set of articles' and authors' characteristics. Table 2.3 shows the results from probit regressions that consider some of these variables simultaneously. The equation is:

$$P(Data\ Type_{iaf}) = \Phi(\alpha + \beta_f + \gamma X_{ia}), \tag{2.1}$$

where $Data\ Type_{iaf}$ is a dummy variable for whether test i in journal article a in field f relies on a specific data type (e.g., admin). X_{ia} includes a dummy variable for whether the submission is solo-authored and the following author-level characteristics aggregated to the paper-level: average years since PhD, average years since PhD squared, average PhD institutional rank, average institutional rank, share of female authors, an indicator for whether at least one of the authors was an editor of an economics journal at the time of publication. In columns 1 and 2, the dependent variable is a dummy for whether the test is in an article which relies on admin data. This dummy takes the value of zero if the article relies on survey, hand collected or other data type. In columns 3 and 4, the dependent variable is a dummy for whether the test is in an article which relies on survey data, while the dependent variable in column 5 is a dummy for whether the test is an article using hand collected data. We include the articles' and authors' variables described above in all columns. We add to this list year and "Top 5" dummy variables. We also include field fixed effects in columns 2 and 4. We rely on probit models and report marginal effects.

In columns 1 and 2, we document that tests in articles published by less experienced authors and with a higher share of women are significantly less likely to rely on admin data than other data types. Of note, the relationship between admin data and experience is not robust to the inclusion of field fixed effects. The other predictors (articles and authors' characteristics) included in our analysis are not statistically significant at conventional levels. The fields which rely the most on admin data are urban and public economics.

For survey data (columns 3 and 4), we provide evidence that survey data are less likely to be used in 2018 and "Top 5" journals in comparison to other data types. We also find that the share of female authors is positively related to the use of survey data. Development economics

is the field that is the less likely to rely on survey data.

Last, column 5 reports estimates for hand collected data. We find that the share of female authors and authors who graduated from top institutions is positively related to the use of hand collected data.[18]

Overall, we find that our set of authors and articles' characteristics are relevant in explaining data type use. This finding suggests that the inclusion of these variables in our analysis is important for identifying the extent of selective reporting by data type as these characteristics may also be related to authors, reviewers and editors' behavior.

2.3.2 Data Type and Replication Material Availability

We now test whether journal articles relying on admin data are less likely to provide data and codes for replication. We estimate the following equation:

$$Pr(ReplicationMaterial_{iaf} = 1) = \Phi(\alpha + \beta_f + X'_{ia}\delta + \gamma Survey_{ia} + \lambda HandCollected_{ia} + \theta Other_{ia}) \quad (2.2)$$

where $Replication\ Material_{ijf}$ is a dummy variable for whether both data and codes (or at least codes) for replication are provided on the journal's website for test i in journal article a in field f. We rely on probit models and present standard errors clustered at the journal article-level. The variables of interest are $Survey_{ia}$, $Hand\ Collected_{ia}$ and $Other_{ia}$, which represent dummy variables for different data types. Administrative data is the reference category, which is omitted.

We include in our model the term X_{ia}, which is a vector including dummy variables for how results are reported (i.e., p-values, standard errors or t statistics), and two indicator variables for the year of publication and whether the article is published in a "Top 5" journal. We also include our set of authors and articles' characteristics and also estimation dummies (e.g. IV).

The results are presented in Table 2.4. In columns 1–3, the dependent variable is whether the data and code can be accessed directly on the journal's website. In our sample, 34% of tests are in journal articles that provided both direct data access and codes. For columns 4–6, the dependent variable is a dummy variable indicating that the authors provided (at least) the codes for replication. This is the case for half of our test statistics. In columns 1 and 4, we include only our variables of interest for data type. Columns 2 and 5 add to the model our articles' and authors' characteristics. In columns 3 and 6, we also include eight field fixed effects.

[18] We do not include field fixed effects for the hand collected data analysis since convergence is not achieved with probit or logit models.

We include dummy variables for the following fields: general interest, finance, macroeconomics, development, experimental, public and urban economics.

We find that survey and hand collected data are significantly more likely to provide data and codes than admin data. Our estimates (column 3) suggest that survey and hand collected data are about 13 and 34 percentage points more likely to provide data and codes than admin data. The estimates are statistically significant at the 1% level for hand collected data and borderline insignificant at the 10% level for survey data. In contrast, we do not find evidence that data type is related to the likelihood to provide at least the codes. Our estimates are small and statistically insignificant in all models. Given the fact that virtually all journal articles providing data also provided codes, our results suggest that articles using survey and hand collected data are significantly more likely to provide data and as likely to provide codes in comparison to admin data.

Interestingly, tests in journal articles published in one of the "Top 5" journals are significantly more likely to provide data and codes. This result is consistent with the fact that "Top 5" journals all had a mandated data sharing by 2016 (Christensen and Miguel (2018)). The estimates for most of the other control variables are not statistically significant, with the exception of a positive relationship between the share of authors who graduated at a top PhD institution and the provision of code.

2.4 Distribution of Test Statistics Across Data Types

Figure 2.1 illustrates the raw distribution of test statistics in our sample for $z \in [0, 10]$. Similar to Brodeur et al. (2020), we create z-curves by imposing an Epanechnikov kernel density (also of width 0.10). A kernel smooths the distribution, softening both valleys and peaks. Reference lines are provided at the conventional two-tailed significance levels. This figure plots two z-curves into a single panel. The first z-curve restricts the sample to journal articles that rely on one type of data, while the second curve does not impose this restriction and rely on the full sample. The distribution for both samples is extremely similar, and exhibits a two-humped shape with a first hump around 0.5 and a second hump between 1.65 and 2.5, suggesting misallocated z-statistics. About 55.7%, 48.3% and 34.2% of test statistics are significant at the 10, 5 and 1 percent levels.[19]

Figure 2.2 displays the raw distribution of z-statistics for each of the four data types. We only consider journal articles that rely on one type of data.[20] The shapes are striking with both

[19]For comparison, in the full sample that is used in Brodeur et al. (2020) 55.9%, 48.1% and 33.7% of test statistics are significant at the 10, 5 and 1 percent levels.

[20]See Appendix Figure B.1 for a similar figure relying not only on those estimates that rely on solely one data set but rather on the full sample as in Brodeur et al. (2020).

admin and survey data featuring a similar two-humped shape. For *survey* data, the distribution exhibits a local minimum around 1.5 and a maximum around 1.96, suggesting misallocated z-statistics. Approximately 62.7%, 54.5% and 38.4% of test statistics are significant at the 10, 5 and 1 percent levels. The distribution of tests for *admin* data also exhibits a local maximum around 1.96. About 58.4%, 51.4% and 37.2% of test statistics are significant at the 10, 5 and 1 percent levels. In contrast, *hand collected* data displays an almost monotonically falling curve with a much smaller local maximum around 1.96. 47.0%, 39.3% and 27.0% of hand collected tests are significant at the 10, 5 and 1 percent levels.

Last, the distribution of tests in journal articles categorized as *Other* has a maximum around 1.96 and seems to have the largest amount of misallocation. 67.3%, 60.4% and 43.7% of test statistics are significant at the 10, 5 and 1 percent levels. Appendix Figure B.2 splits the *Other* data type into two categories: financial data and non-financial data. Among the tests in *Other*, about 54% are in articles relying on financial data. This split into financial and non-financial data illustrate that the distribution of tests for both these subgroups is quite similar, with slightly less bunching at 1.96 for financial data.

One potential issue discussed in the literature is the overrepresentation of round values (e.g., coefficient of 0.02 and standard error of 0.01). We follow Brodeur et al. (2016*b*) and deal with this potential issue by randomly redrawing a number in the interval of potentially true numbers around each collected value using a uniform distribution. This de-rounding method has no impact on our conclusions. See Appendix Figures B.3 and B.4.

We investigate these patterns for different subsamples in Appendix Figures B.5-B.8. These figures illustrate decompositions by data types by journal ranking (Top 5 and non-Top 5), number of authors, institution rank and PhD institution rank. Of note, we find that the spike around 1.96 is more pronounced for journal articles with no authors who graduated from a top university, while the current affiliation does not appear to be related to the spike around the 5% threshold. The shape of the distributions is quite similar for solo- and multi-authored articles, with the exception of hand collected data where the spike is particularly striking for solo-authored articles.

Figures 2.3 and 2.4 illustrate the distribution of tests by replication characteristic. Figure 2.3 splits the sample by whether the article provides direct access to replication data *and* code, while Figure 2.4 splits the sample based on whether the journal article provides at least the code. (See Appendix Figures B.9 and B.10 for the de-rounded z-statistics.) All subfigures display a similar two-humped shape pattern, suggesting that data and code availability is not related to selective reporting.

This result is not necessarily surprising as we the main goal of data sharing is not to decrease selective reporting, but rather allowing other researchers to replicate prior studies' findings.

Moreover, replications of results using the same data in economics is quite rare, making it unclear whether authors of published studies believe other researchers will replicate their results (Hamermesh (2017); Mueller-Langer et al. (2019)).

We also investigate whether the relationship between the provision of data and codes and selective reporting varies by data type. Appendix Figures B.11-B.14 illustrate this relationship for our four data types. Overall, we do not see much difference in the distribution of tests for these subsamples. The amount of misallocation appears slightly larger for admin data for tests in articles that do not provide data and/or codes than for the sample of articles that provide replication material. We observe the opposite pattern for survey and hand collected data.

2.5 Main Results: P-Hacking by Data Type and Availability of Replication Material

In this section, we formally document the extent of p-hacking by data type and availability of replication material. We first describe and rely upon the caliper test method, which consists of comparing test statistics close to arbitrary significance thresholds. We then describe our results using the excess test statistics method and the methodology developed by Andrews and Kasy (2019). Last, we explore the role of the review process in mitigating/exacerbating the extent of selective reporting.

2.5.1 Caliper Test: Method

The caliper test compares the number of test statistics in a narrow range above and below a statistical significance threshold. We focus throughout on the 5% and 10% significance thresholds, but provide similar estimates for the 1% threshold. For the 5% threshold:

$$R_{-,h} = [1.96 - h, 1.96], R_{+,h} = [1.96, 1.96 + h] \tag{2.3}$$

for a bandwidth parameter h.

More precisely, we estimate the following equation:

$$Pr(Significant_{ij} = 1) = \Phi(\alpha + \beta_j + X'_{ij}\delta + \gamma Survey_{ij} + \lambda HandCollected_{ij} + \theta Other_{ij} + \mu Access_{ij}) \tag{2.4}$$

where $Significant_{ij}$ is a dummy variable for whether test i in journal i is statistically significant at the 10%, 5% or 1%-level. We rely on probit models throughout and present standard errors

clustered at the journal article-level.[21] We restrict the sample to $z \in [1.46, 2.46]$ for the 5% statistical significance in our baseline analysis. We also check the robustness of our results to smaller bandwidths. The variables of interest are $Survey_{ij}$, $Hand\ Collected_{ij}$ and $Other_{ij}$, which represent dummy variables for different data types. We define $Admin_{ij}$ as a reference category and omit this feature such that the reported coefficients on various types of data need to be interpreted as a deviation to our reference category. Moreover, we check for the influence of access to the data and code used in primary studies. We also test for differences in whether or not researchers at least disclose their code on the journals' webpages. This specification allows for the possibility that researchers provided access to the code but due to confidentiality reasons cannot disclose the data itself.

The main advantage of using caliper test instead of a graphical examination of the distribution of tests is that we can control for authors' and articles' characteristics. We thus include the vector X_{ij}, which includes our set of articles' and authors' characteristics. We also include journal fixed effects in some models.

2.5.2 Caliper Test: Results

We show our main results of equation 2.4 for the 5% and 10% significance thresholds in Tables 2.5 and 2.6.[22] We restrict the sample to $z \in [1.46, 2.46]$ for the 5% threshold and $z \in [1.15, 2.15]$ for the 10% threshold. Our sample consists of about 2,900 observations from 124 journal articles. Our variables of interest are dummy variables for data types. The coefficients presented are increases in the probability of statistical significance relative to the baseline category (admin).

In column 1, we do not include any fixed effects or control variables and find that survey data and other data are not significantly more or less likely to be marginally statistically significant than admin data. The point estimates for survey and other data are very small in magnitude. For hand collected data, the point estimates are negative in both tables, but statistically significant at conventional levels only for the 5% threshold. The estimate suggests that tests that are hand collected are 8 percentage points less likely to be marginally statistically significant at the 5% level than an admin or survey data estimate.

In column 2, we control for articles' and authors' characteristics. This allows us to check if authors (or articles) that are more/less likely to p-hack are also more/less likely to use specific data types. Column 3 adds to the model journal fixed effects. Our conclusions are unchanged for survey and other data. For hand collected tests, the point estimates are now smaller and

[21] We present bootstrapped standard errors, clustered by article in Appendix Table B.5.

[22] Appendix Table B.6 shows our estimates for the 1% significance level. The estimates for hand collected and survey data are small and statistically insignificant, suggesting that selective reporting is not meaningfully related to data type for those statistical thresholds. Similarly, we find no evidence that tests in journal articles that provide access to data and/or code for replication are less likely to marginally reject the null hypothesis at the 1% level.

marginally insignificant in column 3 for the 5% significance threshold, while the sign flips (becomes positive) for the 10% significance threshold.

In column 4, we further control for methods (i.e., DID, IV, RCT and RDD) since Brodeur et al. (2020) present evidence that DID and IV tests are significantly more likely to be marginally statistically significant than RCT and RDD. The inclusion of method dummies in the model has no effect on the survey and other coefficients but makes the estimate for hand collected data more negative and marginally statistically significant in Table 2.5.[23]

In columns 5 and 6, we test whether journal articles providing data and/or codes are less likely to report (marginally) statistically significant results. The inclusion of these variables also serve to check whether their inclusion in the model affects the relationship between selective reporting and data types. Recall that studies using hand collected data in our sample are more likely to report data and codes than admin data. Column 5 relies on our binary indicator for direct access to data and code, while column 6 relies on a binary indicator for whether at least the codes can be accessed. We find no evidence that access to data and/or code for replication is related to the extent of selective reporting. This result is robust to the exclusion of articles and authors' characteristics and journal fixed effects from the model (Appendix Tables B.7-B.9). Furthermore, the inclusion of these indicators has no effect on the size or significance of our data type variables.

We test the robustness of our caliper test results to the weighting scheme (Appendix Tables B.10-B.12), smaller bandwidths (Appendix Tables B.13-B.15) and coding of main tests (Appendix Tables B.16 and B.17) and de-rounding (Appendix Tables B.18 and B.19).[24] For the weighting scheme, we use the inverse of the number of tests presented in the same article to weight observations. This robustness check is important given that some articles have many more tests than others and could be driving our findings. We confirm our previous results that the extent of selective reporting is similar for admin, survey and other data, and that tests in articles providing data and codes are not significantly less likely to report marginally significant estimates. In contrast, our finding that hand collected data suffers from less p-hacking is not robust to the weighting scheme. The point estimates for the 5% threshold are now smaller and statistically insignificant when we include our set of controls and fixed effects in the model (Appendix Table B.10).

We show in Appendix Appendix Tables B.13-B.15 that the magnitude and significance of

[23]Interestingly, the estimates for the DID and IV dummy variables are now much smaller in magnitude and statistically insignificant (baseline RCT) in comparison to a similar model excluding data type dummies, suggesting that the addition of data type fixed effects entirely explain this relationship. Recall that the majority of hand collected data in our sample are from RCT studies (87%). Of note, DID and IV tests remain significantly different than RDD tests.

[24]We also show that the point estimates and conclusions are robust to the use of a logit model instead of a probit and to bootstrapped standard errors, clustered by article. See Appendix Tables B.5 and B.20.

our estimates for data types and replication material remain the same when using smaller bandwidths (e.g., $z \in [1.61, 2.31]$ for the 5% threshold). Similarly, our conclusions are unchanged if we drop test statistics for which there was initial disagreement between the researchers during the data collection as to whether they were 'main' tests of an article (Appendix Tables B.16 and B.17). Last, de-rounding our z-statistics has no effect on our main conclusions (Appendix Tables B.18 and B.19).

Overall, our findings suggest that the proportion of tests that are marginally significant in articles relying on survey data is not significantly different than in articles using admin data. We find weak evidence that hand collected tests are less likely to be marginally statistically significant than other data types, but this result is not robust to a battery of specification checks. Last, we find no evidence that tests in articles providing data and codes are less likely to be marginally statistically significant.

2.5.3 Excess Test Statistics: Method

We now turn to our second method in which we compare the distribution of tests for each subsample to a counterfactual distribution. The main advantage of this method over caliper tests is that we are able to document the extent of selective reporting for each data type without comparing it to a baseline (i.e., admin data). However, additional assumptions about the counterfactual distribution are necessary.

We follow Brodeur et al. (2020) and calibrate a different counterfactual *t*-distribution for each data type. More precisely, we calibrate a non-central input distribution by data type assuming that the observed test statistic distribution above $z = 5$ should not suffer from publication bias and selective reporting by the authors. This assumption is based on previous studies which do not find any bunching around or past the 1% statistical threshold (e.g., Brodeur et al. (2016*b*)) and on the lack of incentives to engage in specification searching in that range. We thus produce a non-central *t*-distribution for each data type by calibrating the degrees of freedom and non-centrality parameter leading to very similar observed and counterfactual distributions for $z > 5$. In other words, we calculate the non-centrality parameter that minimizes the difference in the observed and expected distributions above $z = 5$.[25]

[25]We focus only on positive integers for the degrees of freedom and optimize in steps of 1. The non-centrality parameter of the t-distribution is positive and real valued and optimize in steps of 0.01. We choose the best of the 10 optimized *t*-distributions by degree of freedom.

2.5.4 Excess Test Statistics: Results

Figure 2.5 presents the calibrated input distribution and the observed distribution by data type. We also report the excess test statistics for the non-significant and significant regions in Appendix Table B.21.[26] Let us first note that our fitting above $z = 5$ has succeeded as the calibrated t and the observed distribution are closely matched for all data types. For instance, hand collected tests have 9.3% of its mass in the tail and our algorithm produces a t-distribution also with a mass of 9.3%.

The difference between the observed and expected distributions is negative for survey, hand collected and other data, suggesting selective reporting. In contrast, the difference for admin data is positive for the non-significant region, i.e., a dearth of significant admin tests. Interestingly, these 'missing' tests are found just above the 1% significance threshold. This result provides suggestive evidence that admin tests are more likely than expected to have very large p-values.

We find that the mass difference between observed and expected between the 10% and 1% significance thresholds is very small for admin data and relatively small for hand collected data in comparison to survey (and to some extent other) data. Our estimate for the region $[1.96 < z < 2.58)$ suggests that survey data has an excess of about 2.2 percent of its total mass, or 16 percent more statistically significant test statistics than expected.

Overall, our findings using this second method confirm our previous results that admin tests do not suffer from more selective reporting. On the contrary, we provide some evidence that survey tests appear to have an excess of marginally significant tests.

2.5.5 Andrews and Kasy's Measurement of Publication Bias

We now turn to our third method in which we apply Andrews and Kasy (2019)'s measurement of publication bias. This method is particularly attracting in our setting given that it allows to compute the relative publication probability of statistically significant results in comparison to insignificant results for different subsamples. In other words, it provides estimates for the extent of publication bias, i.e., publication probability is related to statistical significance, by data type.

This method involves applying a step function at the different significance threshold to the conditional probability of publication, assuming that the underlying effect sizes follow a generalized t-distribution and that effect size estimates with smaller standard errors do not relate to different estimates.

[26] See Appendix Table B.22 for a direct comparison of the extent of selective reporting for admin data and the other data types.

Table 2.7 presents the estimates for the relative publication probability of a statistical region as compared to $Z > 2.58$. We also report the relative publication probability of a statistically insignificant test. For admin data, we find that a statistically significant result at the 10% and 5% levels is approximately 2.4 and 3.4 times more likely to be published than an insignificant result, respectively. For survey data, significant results at the 10% and 5% level are 2.8 and 3.6 times more likely to be published than an insignificant result, respectively. Last, statistically significant results for hand collected data are also more likely to be published than insignificant results, but by a smaller amount than admin and survey data, i.e., 1.7 and 2.1 times at the same significance thresholds.

2.5.6 Further Subsample Analyses

In this subsection, we test different channels through which different data types might produce differing patterns of selective reporting in the published literature. We also provide additional subsample analyses by estimation method.

Public and Private Administrative Data

We start by investigating whether it is easier to manipulate test statistics for public versus 'private' admin data. We code administrative data obtained from private companies (e.g., AXA) as non public admin data. Public admin data include governmental data such as administrative tax records. Our sample consists of 3,212 admin data tests, of which 89.8% of these tests are coded as public. It is arguably harder to share data and codes (and harder to obtain for other researchers) for non-public admin data than it is for public admin data. Among those tests that provide direct access to data and code, only 11.4% use admin data. If we split the admin sample into two categories public and non-public admin data, we see that almost none of the non-public admin data can be access directly.

Appendix Figure B.15 shows the distribution of tests for public and non-public admin data, respectively. The two subfigures are quite similar potentially suggesting that the type of admin data is not related to selective reporting.

By Estimation Method

We also investigate whether the distribution of tests is related to data type for each method separately. Recall that Brodeur et al. (2020) show that tests that use IV as a method reveal the largest misallocation of tests. Appendix Figures B.16, B.17, B.18 and B.19 illustrate the distribution of tests by data type for difference-in-differences, instrumental variables, randomized control trials and regression discontinuity design, respectively. We only report subfigures by

method if the sample is large enough, i.e., at least 200 observations for a given data type and method.[27] For IV tests, we find that the extent of p-hacking appears to be larger for survey and other data types than for admin and hand collected data.[28]

2.5.7 Role of the Review Process

We investigate the role of the reviewing process in mitigating/exacerbating the extent of selective reporting by data type in this subsection. For this exercise, we directly compare the distribution of test statistics in our sample of published articles to the distribution of tests in the corresponding working papers for each data type. The objective of this exercise is to document whether journal editors and reviewers require or propose changes that would lead to an increase/decrease prevalence of marginally significant tests.

In order to document the impact of the reviewing process, we only rely on working papers released before the date of submission to the journal. For the 11 journals for which we do not have the date of submission, we rely only working papers released at least two years prior to publication. For those with multiple working papers, we chose the working paper closest to the date of submission (or the two-year threshold). Our final sample of working papers comprises 133 articles/working papers.[29]

Our data collection methodology for the working papers is the same as for the published version. While some working papers include additional main tests/tables, or rely on different clustering or weighting techniques, we find that the distribution of tests is remarkably similar between the working paper and published version for all data types. Appendix Figure B.23 illustrates the distribution of test statistics in the working paper versus published version for each data type. The curves for the working paper and published version are mostly on top of each other for all data types. We formally test whether there are changes in reporting of marginally significant results due to the reviewing process in Appendix Tables B.25, B.26 and B.27. This table reports caliper test for the 5%, 10% and 1% significance thresholds, respectively. The variable of interest is a dummy for whether a test comes from the working paper or the published version. In column 1 the estimated effect of the publication process is very small, negative, and statistically insignificant. This leads us to believe the editorial process does

[27] Similarly, we explore whether the extent of p-hacking is lower for hand collected data for each method separately. Appendix Figures B.20, B.21 and B.22 present histograms for DID, IV and RCT respectively in which we split the sample in two: hand collected and non-hand collected tests. We find some evidence that the two humped shape is more pronounced for non-hand collected tests than for hand collected tests for DID, IV and RCT.

[28] For our sample of DID tests, we find that the spike at about 1.96 is the smallest for hand collected data and the largest for other data. The shape of the distributions is quite similar for admin and survey data. There are no clear differences across data types for RDD tests.

[29] The likelihood to find a valid working paper is not statistically related to data type. See Appendix Table B.24.

not change the extent of p-hacking. Columns 2–5 restrict the sample to admin, survey, hand collected and other data type, respectively.

2.6 Conclusion

Demands for and use of administrative data to analyze different aspects of our lives, our society and our economy continue to grow. Given that the proportion of articles receiving exemptions from data-sharing policies for admin data is larger than for other data types, the increasing use of administrative records may raise concerns about the reproducibility of research findings, and ultimately, research credibility. In this paper, we documented one unexplored facet of the link between methodologies for data collection and research transparency; the extent of p-hacking across data types. Our analysis points to small between-data types differences, with no significant differences between admin and survey data. These results are key from the point of view of policymakers and researchers who are interested in knowing to what extent they should be skeptical about the credibility of the published literature using specific data types.

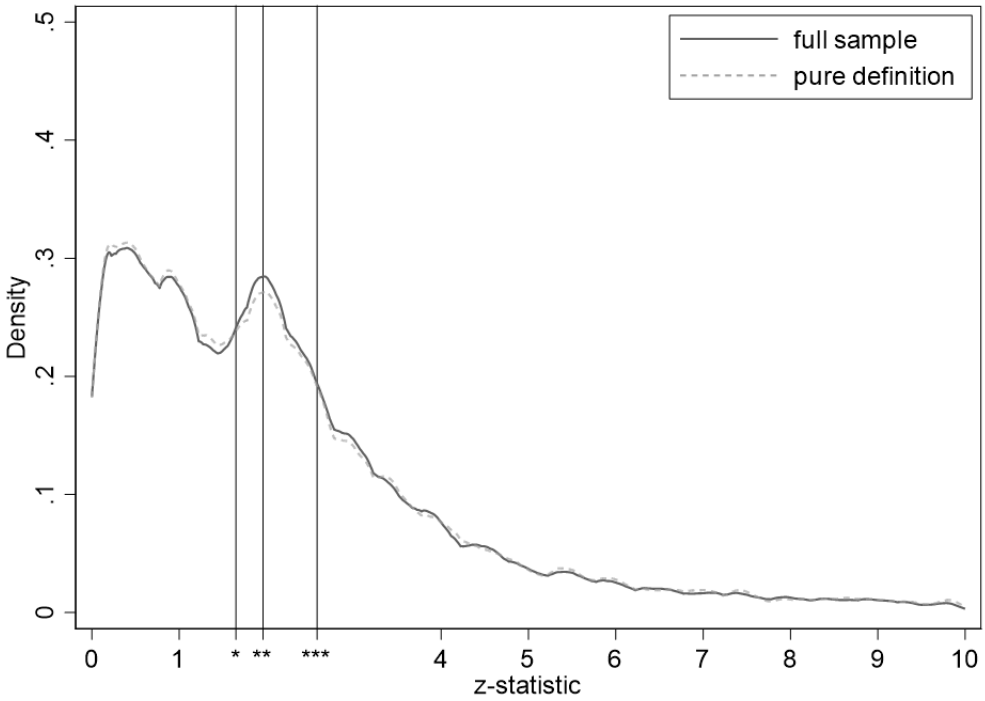
While our results are indicative that admin and survey suffer equally from p-hacking, they also suggest that the the difficulty of providing replication material is not a key factor driving p-hacking in economics.

To conclude, we briefly discuss some of the limitations of our study. First, our study deals with journal articles from top economics journals, and thus our findings might not generalize to lower-ranked journals. Second, we are unable to say much about the file drawer problem, and the possibility that studies using specific data types are more likely to end up unpublished. This would be an issue if unpublished studies that are more/less p-hack are more/less likely to use a specific methodology for data collection. Last, our results should not be viewed as indicating that archives for replication material are not useful.[30] They instead suggest that the benefits of such archives are more limited than previously thought, but still extremely valuable by allowing, for instance, other researchers to replicate research findings.

2.7 Figures and Tables

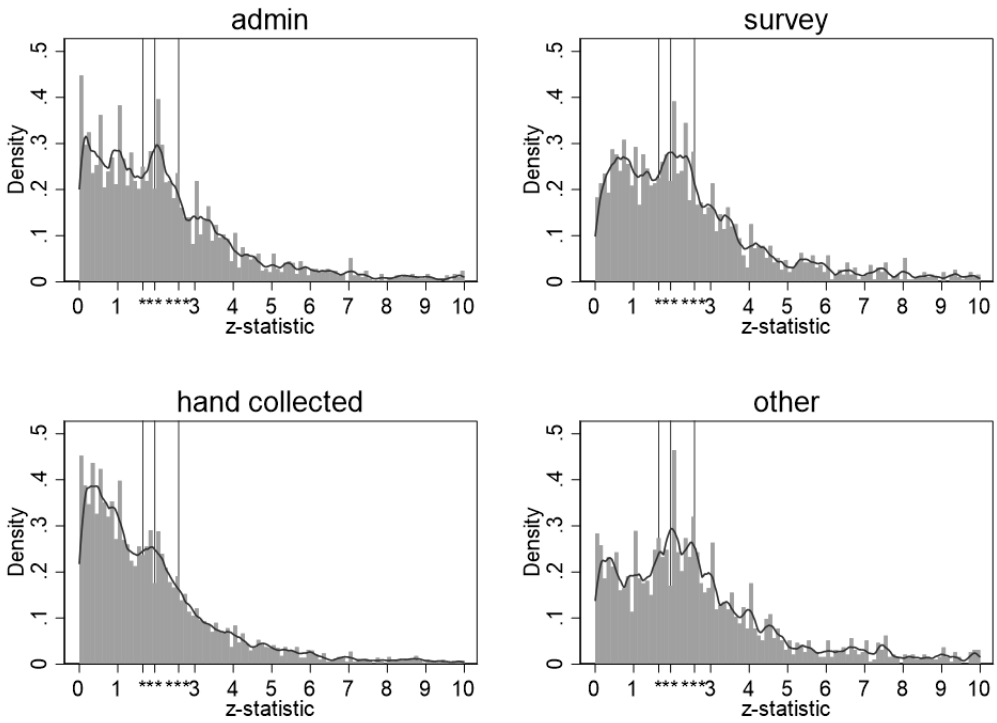
[30]It is worth mentioning the recent increase in formal restricted-access data environments, which facilitate access to admin data for a large number of researchers. Examples of such environments include the U.S. Federal Statistical Research Data Center and the German IAB FDZ.

Figure 2.1: *z*-statistics for all Estimates vs *z*-statistics for those Estimates that rely solely on one Data Type



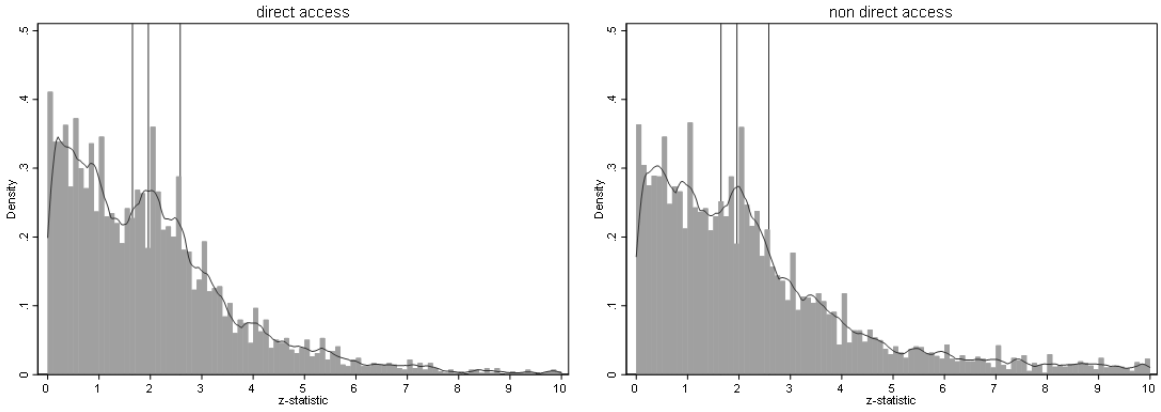
Notes: This figure displays two distributions. First, the solid line plots *z*-statistics for the sample used in Brodeur et al. (2020) (N=21,440) and second, the dashed line plots *z*-statistics for the sub-sample of estimates that rely solely on one data type (N=12,495). Both figures are based on an Epanechnikov kernel wih a bandwidth of 0.1. Estimates are not weighted.

Figure 2.2: z-statistics by Method of Data Collection



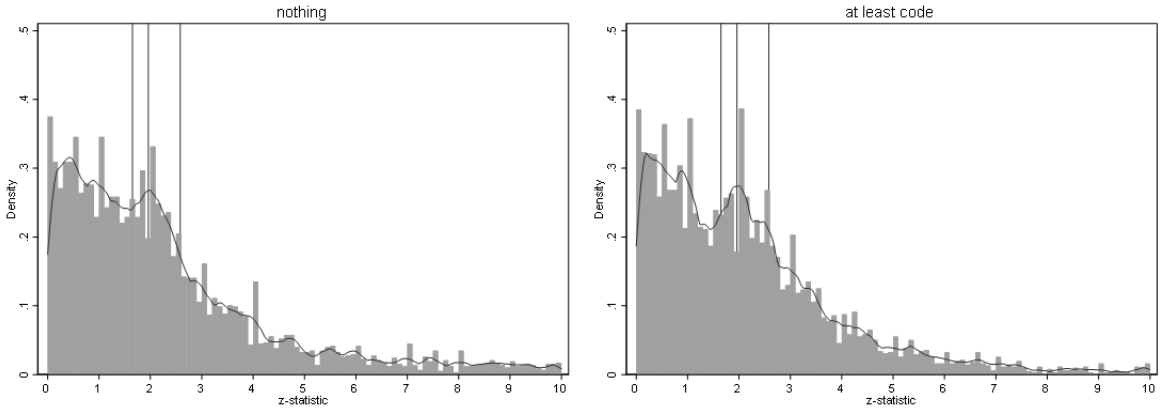
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by Method of Data Collection: *admin*, *survey*, *hand collected* and *other*. We only consider those observations that rely solely on one data type within each primary study (N=12,495). Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure 2.3: z-statistics by Accessibility of Replication Material: Data *and* Code



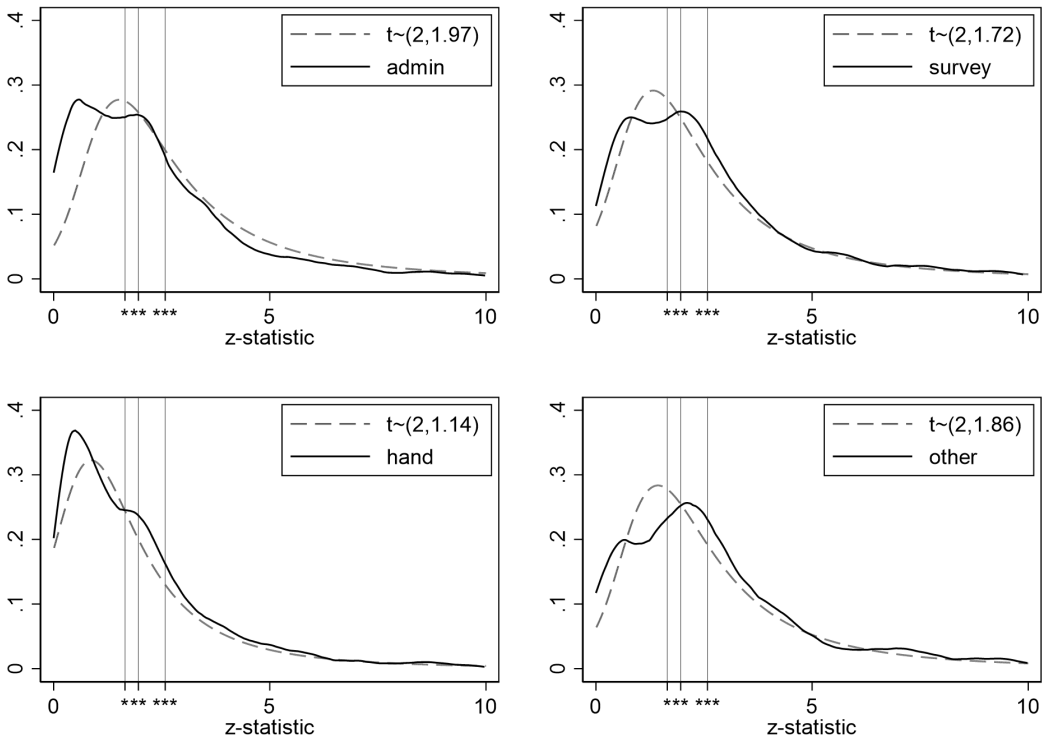
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. The left figure restricts the sample to estimates that provide direct access to data and code. The right figure restricts the sample to estimates that do not provide both data and code. We only consider those observations that rely solely on one data type within each primary study (N=12,495). Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure 2.4: z-statistics by Availability of Replication Material: At Least Code



Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. The left figure restricts the sample to estimates that do not provide any replication material (i.e., data and/or code). The right figure restricts the sample to estimates that at least provide code for replication. We only consider those observations that rely solely on one data type within each primary study (N=12,495). Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure 2.5: Excess Test Statistics by Method of Data Collection



Notes: This figure presents the calibrated input distributions with the observed distributions. We optimize for each type of data a student *t*-distribution with 2 degrees of freedom. The optimal non-centrality parameter varies across data types. We only consider those observations that rely solely on one data type within each primary study (N=12,495). See Section 2.5 for more details.

Table 2.1: Summary Statistics: Method of Data Collection and Replication Characteristics

Pure Sample					
	admin	survey	hand	other	Total
Total Articles	113	80	117	92	402
Articles in %	28.11	19.90	29.10	22.89	100
Total Tests	3212	1999	5265	2019	12495
Tests in %	25.71	16.00	42.14	16.16	100
Direct Access to Data and Code					
	admin	survey	hand	other	Total
Total Articles	25	27	58	20	130
Articles in %	19.23	20.77	44.62	15.38	100
Total Tests	486	558	2795	420	4259
Tests in %	11.41	13.10	65.63	9.86	100
Provision of (at least) Code					
	admin	survey	hand	other	Total
Total Articles	63	41	65	31	200
Articles in %	31.50	20.50	32.50	15.50	100
Total Tests	1664	867	2964	860	6355
Tests in %	26.18	13.64	46.64	13.53	100

Notes: The first part of this table provides an overview of the distribution of total tests and total articles by method of data collection. The second part shows the distribution of tests and articles by type of data that provide direct access to data *and* code. The percentage shares by method of data collection relates to all tests (or articles) that provide direct access. For example, 19.23% of all articles that provide direct access to data and code are admin data. The third part shows the distribution of tests and articles by method of data collection that provide at least the code. The percentage shares by method of data collection relates to all tests (or articles) that provide at least the code. For example, 26.18% of all tests that provide at least code use admin data. All numbers are based on a subsample (= Pure Sample) of estimates used in Brodeur et al. (2020) that rely solely one one datatype within each primary study.

Table 2.2: Summary Statistics: Method of Data Collection and Article and Author Characteristics

	Method of Data Collection				Total
	admin	survey	hand collected	other	
	(1)	(2)	(3)	(4)	(5)
Top 5	0.28	0.06	0.25	0.08	0.20
	(0.45)	(0.23)	(0.43)	(0.28)	(0.40)
Editor present	0.55	0.51	0.74	0.74	0.65
	(0.50)	(0.50)	(0.44)	(0.44)	(0.48)
Solo-authored	0.30	0.31	0.12	0.16	0.21
	(0.46)	(0.46)	(0.33)	(0.37)	(0.40)
Average	10.03	10.86	12.96	11.27	11.60
experience	(35.36)	(5.62)	(6.42)	(5.46)	(18.71)
Female authors	0.23	0.41	0.39	0.20	0.32
	(0.35)	(0.39)	(0.31)	(0.26)	(0.34)
Top institutions	0.33	0.19	0.36	0.18	0.30
	(0.37)	(0.30)	(0.35)	(0.31)	(0.35)
Top PhD	0.37	0.27	0.50	0.35	0.40
institutions	(0.41)	(0.36)	(0.39)	(0.40)	(0.40)
Test statistics	3212	1999	5265	2019	12495

Notes: Each observation is a test. The Top 5 journals in economics are the American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics, and Review of Economic Studies. Average experience is the mean of years since PhD for an article's authors. Share of female authors, share of authors affiliated with top institutions, and share of authors who completed a PhD at a top institution.

Table 2.3: Probit Regressions, Prediction of Data Type Use

	admin		survey		hand
	(1)	(2)	(3)	(4)	(5)
Year=2018	0.036	0.030	-0.091	-0.071	-0.003
	(0.054)	(0.047)	(0.044)	(0.041)	(0.075)
Top 5	0.108	0.184	-0.134	-0.270	0.059
	(0.073)	(0.150)	(0.062)	(0.128)	(0.089)
Experience	-0.023	-0.008	0.013	0.019	0.020
	(0.014)	(0.013)	(0.012)	(0.011)	(0.020)
Experience^2	0.038	0.007	-0.040	-0.049	-0.019
	(0.041)	(0.038)	(0.038)	(0.035)	(0.052)
Top Institution	0.131	0.096	-0.071	-0.036	0.051
	(0.088)	(0.075)	(0.071)	(0.065)	(0.112)
PhD Top Institution	-0.093	-0.056	-0.089	-0.090	0.189
	(0.085)	(0.068)	(0.058)	(0.056)	(0.103)
Sole-authored	0.066	0.084	0.037	0.065	-0.106
	(0.086)	(0.075)	(0.066)	(0.058)	(0.119)
Female authors	-0.176	-0.174	0.139	0.145	0.174
	(0.092)	(0.073)	(0.065)	(0.060)	(0.096)
Editor present	0.001	-0.049	-0.061	-0.079	-0.042
	(0.074)	(0.069)	(0.058)	(0.054)	(0.095)
Field: Finance		0.074		-0.224	
		(0.155)		(0.143)	
Field: General Interest		0.034		-0.102	
		(0.143)		(0.120)	
Field: Development		-0.222		-0.290	
		(0.158)		(0.128)	
Field: Labor		0.072		-0.028	
		(0.168)		(0.135)	
Field: Public		0.217		-0.139	
		(0.144)		(0.122)	
Field: Urban		0.612		-0.212	
		(0.181)		(0.162)	
Field: Macro Growth				-0.042	
				(0.159)	
Observations	12,495	12,302	12,495	12,422	12,472

Notes: This table reports marginal effects from probit regressions (equation (2.1)). The dependent variable in column (1)-(2) is a dummy for use of *admin* data; column (3)-(4) is a dummy for use of *survey* data; and column (5)-(6) is a dummy for use of *hand collected* data. The respective reference category are all other tests that do not use the respective dataset. For example, column (1) and (2) compare those estimates that rely on *admin* data to those estimates that *do not use admin* data. All regressions control for a set of authors and articles' characteristics, while columns (2), (4) and (6) add to the model field fixed effects. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table 2.4: Probit Regressions, Prediction of Provision of Data and/or Code

Provision of ...	Data *and* Code			at least Code		
	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.128	0.188	0.128	-0.084	0.032	-0.229
	(0.082)	(0.081)	(0.082)	(0.108)	(0.089)	(0.351)
hand collected	0.380	0.273	0.341	0.045	-0.060	-0.079
	(0.081)	(0.083)	(0.079)	(0.098)	(0.101)	(0.372)
other	0.057	0.162	0.197	-0.092	0.086	0.682
	(0.076)	(0.083)	(0.111)	(0.114)	(0.086)	(0.344)
Estimation Method: (omitted RCT)						
DID		-0.069	-0.004		-0.100	-0.108
		(0.072)	(0.072)		(0.091)	(0.335)
IV		-0.134	-0.121		-0.158	-0.478
		(0.070)	(0.076)		(0.086)	(0.332)
RDD		-0.184	-0.192		-0.196	-0.618
		(0.102)	(0.101)		(0.117)	(0.461)
Controls						
Year=2018		-0.113	-0.091		-0.142	-0.323
		(0.058)	(0.061)		(0.062)	(0.212)
Top 5		0.257	1.476		0.454	6.766
		(0.066)	(0.134)		(0.075)	(0.426)
Experience		0.007	0.010		0.005	0.032
		(0.005)	(0.005)		(0.006)	(0.022)
Experience^2		-0.000	-0.001		-0.000	-0.002
		(0.000)	(0.000)		(0.000)	(0.001)
Top Institution		0.035	0.066		0.015	0.271
		(0.088)	(0.090)		(0.103)	(0.356)
PhD Top Institution		-0.033	-0.037		0.145	0.551
		(0.081)	(0.082)		(0.093)	(0.307)
Other Controls						
Reporting Method		Y	Y		Y	Y
Solo Authored		Y	Y		Y	Y
Share Female Authors		Y	Y		Y	Y
Editor		Y	Y		Y	Y
Observations	12,495	12,472	10,622	12,495	12,472	11,563

Notes: This table reports marginal effects from probit regressions (equation (2.2)). The dependent variable in column (1)-(3) is a dummy for whether the dataset can be directly accessed and column (4)-(6) uses as a dependent variable a dummy for whether at least the code is available on webpages of the journals. The omitted category is *admin*. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table 2.5: Caliper Test, Significant at the 5 percent level: Unweighted Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	-0.001	-0.001	0.029	0.016	0.018	0.016
	(0.038)	(0.036)	(0.040)	(0.041)	(0.041)	(0.041)
hand collected	-0.076	-0.062	-0.052	-0.087	-0.083	-0.087
	(0.027)	(0.027)	(0.031)	(0.051)	(0.050)	(0.051)
other	0.025	0.019	0.016	0.007	0.009	0.009
	(0.045)	(0.047)	(0.053)	(0.052)	(0.054)	(0.052)
Estimation Method: (omitted RCT)						
DID				-0.013	-0.015	-0.016
				(0.042)	(0.043)	(0.042)
IV				-0.032	-0.034	-0.035
				(0.051)	(0.052)	(0.050)
RDD				-0.106	-0.108	-0.108
				(0.057)	(0.058)	(0.056)
Controls						
Top 5		0.029	0.117	0.103	0.110	0.121
		(0.029)	(0.059)	(0.059)	(0.065)	(0.072)
Year=2018		0.009	0.011	0.012	0.012	0.013
		(0.027)	(0.027)	(0.026)	(0.026)	(0.026)
Experience		-0.004	-0.005	-0.006	-0.006	-0.006
		(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Experience^2		0.011	0.016	0.019	0.019	0.019
		(0.021)	(0.023)	(0.023)	(0.023)	(0.023)
Top Institution		-0.037	-0.021	-0.020	-0.020	-0.022
		(0.038)	(0.042)	(0.041)	(0.041)	(0.041)
PhD Top Institution		-0.001	-0.023	-0.029	-0.030	-0.028
		(0.038)	(0.040)	(0.038)	(0.037)	(0.038)
Replication Characteristics						
Direct Access to Data & Code					-0.011	
					(0.034)	
Provision of (at least) Code						-0.025
						(0.043)
Other Controls						
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Journal FE			Y	Y	Y	Y
Observations	2,904	2,904	2,904	2,904	2,904	2,904
Window	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table 2.6: Caliper Test, Significant at the 10 percent level: Unweighted Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.024	0.025	0.022	0.010	0.014	0.010
	(0.038)	(0.035)	(0.037)	(0.039)	(0.039)	(0.039)
hand collected	-0.044	-0.013	0.010	-0.035	-0.028	-0.035
	(0.031)	(0.030)	(0.033)	(0.054)	(0.054)	(0.054)
other	0.039	0.049	0.034	0.030	0.034	0.029
	(0.037)	(0.038)	(0.043)	(0.042)	(0.042)	(0.042)
Estimation Method: (omitted RCT)						
DID				-0.025	-0.029	-0.023
				(0.045)	(0.045)	(0.045)
IV				-0.057	-0.062	-0.055
				(0.045)	(0.045)	(0.045)
RDD				-0.108	-0.112	-0.107
				(0.053)	(0.053)	(0.053)
Controls						
Top 5		0.071	0.102	0.087	0.098	0.079
		(0.032)	(0.087)	(0.087)	(0.087)	(0.085)
Year=2018		-0.016	-0.012	-0.010	-0.011	-0.010
		(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
Experience		0.007	0.008	0.006	0.006	0.006
		(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
Experience^2		-0.036	-0.033	-0.027	-0.027	-0.027
		(0.019)	(0.020)	(0.020)	(0.020)	(0.020)
Top Institution		-0.058	-0.049	-0.047	-0.048	-0.046
		(0.036)	(0.035)	(0.036)	(0.036)	(0.036)
PhD Top Institution		-0.059	-0.063	-0.069	-0.070	-0.070
		(0.029)	(0.033)	(0.032)	(0.032)	(0.032)
Replication Characteristics						
Direct Access to Data & Code					-0.023	
					(0.032)	
Provision of (at least) Code						0.013
						(0.040)
Other Controls						
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Journal FE			Y	Y	Y	Y
Observations	2,933	2,933	2,926	2,926	2,926	2,926
Window	[1.65±0.50	[1.65±0.50	[1.65±0.50	[1.65±0.50	[1.65±0.50	[1.65±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table 2.7: Relative Publication Probabilities by Method of Data Collection

	Method of Data Collection			
	admin	survey	hand collected	other
	(1)	(2)	(3)	(4)
Panel A				
$\beta_{[0<Z<1.96]}$	0.363	0.293	0.534	0.190
	0.016	0.018	0.024	0.011
Location	0.001	0.012	0.023	0.002
	0.000	0.002	0.002	0.001
Scale	0.000	0.008	0.013	0.002
	0.000	0.001	0.001	0.001
Degrees of freedom	1.157	1.876	1.805	1.647
	0.032	0.066	0.045	0.049
Panel B				
$\beta_{[0<Z<1.65]}$	0.345	0.238	0.500	0.150
	0.018	0.017	0.027	0.010
$\beta_{[1.65<Z<1.96]}$	0.820	0.656	0.845	0.434
	0.064	0.062	0.057	0.043
$\beta_{[1.96<Z<2.58]}$	1.182	0.862	1.041	0.732
	0.084	0.077	0.066	0.063
Location	0.001	0.011	0.021	0.001
	0.000	0.001	0.002	0.001
Scale	0.000	0.008	0.012	0.002
	0.000	0.001	0.001	0.001
Degrees of freedom	1.136	1.945	1.813	1.741
	0.034	0.072	0.049	0.056

Notes: In Panel A, $\beta_{[0<Z<1.96]}$ is the relative publication probability of a statistically insignificant test. For example, if a statistically significant test using admin data has a 50 percent chance of being published, then a statistically insignificant one has a 50% x 36.4% = 18.2% chance of being published. Panel B represents the relative publication probability of statistical significance regions as compared to the most significant test statistics ($Z > 2.58$). The table presents the results of applying he publication bias model presented in Andrews and Kasy (2019). The model assumes that the underlying effect sizes follow a generalized $t-$distribution, as elsewhere in this manuscript. We reported the fitted location and scale parameters, as well as the degrees of freedom. We restricted the sample to estimates that rely solely on one type of data.

The Effects of Public Disclosure by Politicians

3.1 Introduction

As in many other countries, German politicians are legally permitted to carry out outside activities in addition to their political work. Politicians engaging in activities other than their work in parliament remains a very controversial topic. On the one hand, there is doubt on whether elected representatives devote all their energy to their political duties and it also raises concerns of potential conflicts of interests (Akcigit et al., 2018). On the other hand, banning politicians from engaging in outside activities might negatively influence the selection of politicians (Gagliarducci et al., 2010; Fisman et al., 2021). A central concern of democratic countries is the degree to which voters can hold members of parliament (MPs) accountable (Djankov et al., 2010). Accountability heavily relies on availability of information about both parliamentary and non-parliamentary actions. One potential policy to inform voters on politician's outside activities are public disclosure laws.[1] If voters observe undesirable behavior, they can vote them out of office. This political pressure could cause politicians to change their behaviour. Despite being widely used, there is little causal evidence on the effects of public disclosure laws on outside activities and earnings. This is due to several reasons. First, it is hard to obtain high-quality data, especially before the introduction of disclosure rules, as politicians outside earnings are unobservable before the implementation of disclosure laws. Second, even the published (and thereby disclosed) data is often coarse and might be misreported. Finally, one has to find a suitable control group to establish a counterfactual scenario.

In this paper, we aim to fill this gap and identify the causal effect of public disclosure of outside activities and associated earnings on politician's outside earnings. We overcome the

[1] According to Djankov et al. (2010), 109 countries around the world have some form of a disclosure law, roughly half of those make disclosed information public. They find suggestive evidence that *public* disclosure is associated with better government and perceived corruption.

existing problems by exploiting (i) two policy changes with respect to public disclosure laws in Germany and (ii) high quality administrative tax return data giving rise to a difference-in-differences setup with German federal MPs as our treatment and state MPs as our control group. We exploit two reforms that differ in the degree of disclosure intensity. First, we use the introduction of a public disclosure law for federal Members of Parliament (MPs) in Germany as a source of exogenous variation. In 2005, a law was passed that requires MPs to publish their outside activities and levels of outside earnings on the website of the German Parliament *Bundestag* that are freely accessible to voters. Initially, disclosure was only private because a group of MPs filed a law suit against such public disclosure rules. In July 2007, the German constitutional court narrowly rejected the law suit, such that disclosure became public. Each activity is assigned an income bracket such that outside earnings were reported in a bracket system top-coded at 7,000€. The fact that information was top-coded was heavily debated in media and parliament and it raised concerns that voters were not adequately informed.[2] In 2013, our second reform under study, more brackets were introduced such that only earnings above 250,000€ were censored. This greatly increased disclosure obligations for MPs and the information available to voters.

We use administrative tax return data for 2001 to 2014 allowing us to observe politicians' outside earnings at a very precise level.[3] Our main outcome is the total amount of outside earnings. Another important feature of the tax data is that it allows us to use state MPs as a control group. Since state MPs were not subject to any disclosure rules during our sample period, we can use them estimate a difference-in-difference model. German state and federal MPs are highly comparable. Both are full-time politicians, they are elected in a similar way and due to the decentralized nature of the German government structure, both groups face a high degree of responsibility. This comparability is underlined by the absence of any differential trend between treatment and control group prior to the reform.

To examine who responded to disclosure of outside activities and earnings, we use (i) different income categories as outcome variables and (ii) run quantile regressions to check for heterogeneous responses along the earnings distribution. On the one hand, voters perceive sources of outside earnings differently (Campbell and Cowley, 2015). On the other hand, the literature on behavioral responses towards taxation shows that the self-employed can more easily adjust their labor supply and also the reporting of their income (Saez et al., 2012).

[2] During the campaign in the run-up of the 2013 federal elections, politician's outside activities were a much discussed issue because of large outside earnings of the candidate for the chancellorship, Peer Steinbrück.

[3] In general, tax data has very little amount of socio-demographic information and researchers face strict confidentiality rules. Importantly, we do not observe any names and we are not allowed to link any external data set to the tax data. Therefore, we cannot make statements about variables like party affiliation when using our tax data.

Therefore, we check for different effects between income from wages and salaries and income from self-employment and businesses. In addition, we use income from renting, an unaffected income category, as a placebo outcome. Given the differences in the bracket structure across both reforms, we expect heterogeneous responses across the earnings distribution. Since voters cannot distinguish between a moderate and a high earning MP, the first reform might induce MPs to cover their high earnings behind level 3 such that they earn larger amounts than 7,000€ , while the second reform and the associated changes in the bracket structure might discourage MPs to report activities with high levels of outside earnings.

On average, 89% of all federal MPs report an activity and 38% disclose positive outside earnings. The most disclosed remunerated activities belong to working as a lawyer (10%), in management and consulting (10%) or giving speeches (8%). Around 40% of all MPs hold a function in enterprises, either as being a member of the advisory or supervisory board. Using tax return data, we observe that the distribution of outside earnings is highly unequal following a pareto distribution.

Our results show that the introduction of public disclosure in 2007 *increased* total outside earnings by 15.3%. The amount of MPs having positive outside earnings also increased by 4.5 percentage points. Quantile regressions show that the effect is mainly driven by the upper end of the earnings distribution. This points to the problem of the conservative top-coding of activities at 7,000€. We show that the increase is mostly driven by income from self-employment and business income, which would be consistent with increased tax compliance as these incomes are self-reported and the public visibility of their incomes might have increased incentives to report income truthfully. However, the timing of the effect suggests that this mechanism is unlikely. We do not see any increase in earnings in two years of private disclosure even though MPs should have anticipated that there a significant chance of their disclosed activities becoming public retroactively. Other possible explanations for the increase include, for example, changing social norms regarding outside incomes, i.e. making outside incomes more normal and therefore, more acceptable.

The tightening of the disclosure law reform provides evidence that disclosure rules lead to a *lowering* of outside earnings. The introduction of seven new brackets allows to distinguish between medium and high-earning MPs. This leads to a reduction in outside incomes of 9.6%. This decrease is mainly driven by reductions in income from wages and salaries consistent with MPs working less for firms other than their own. Quantile regressions show that this decrease is particularly pronounced at the top of the distribution. This is consistent with top-earners being treated most intensely since the new brackets affected them the most.

We also make use of self-collected data on published earnings from webpages of the German Bundestag which we combine with rich data on demographic and political variables. First, we

examine the relationship between tighter disclosure rules and electoral accountability. Directly elected MPs had significantly lower outside earnings when compared to the runner-up in their election district, who joined via the party list, after, but not before the reform. Similarly, MPs with an unsafe rank on the party list had lower outside earnings than MPs with a very safe rank after the reform, while we could not find a difference before. Although income figures are imprecisely measured, it allows us to uncover the relationship between outside earnings and activities, the influence of party affiliation or (previous) occupation. In addition, we use out self-collected data and show descriptively that party affiliation and gender are highly correlated with the amount of outside earnings.

We contribute to several strands of the literature. To the best of our knowledge, this is the first paper examining public disclosure rules for politicians with administrative tax return data for a western democracy. More specifically, we test if individuals change their earnings and thereby the amount of outside activities in response to a mandatory disclosure of these activities along with the respective earnings. Most related, Slemrod et al. (2020) and Malik (2020) exploit an unexpected release of tax records of Pakistani politicians. In contrast to our study, their focus lies on tax evasion in a developing country. While Malik (2020) consider only MPs and provide strong evidence that the pressure to decrease tax evasion was highest for competitively and directly elected legislators, Slemrod et al. (2020) focus on the universe of tax filers and find a 9% increase in the tax paid by individuals that are exposed to public disclosure.

Second, our study contributes to a broader question of how a change in third party information requirements affects income reporting behavior and how public disclosure of income affects the (reported) income itself (Kleven et al., 2016*b*). The effects of income disclosure have been studied among others for the general population (Bø et al., 2015; Slemrod et al., 2020), CEOs (Mas, 2016), and public employees (Mas, 2017). Both Slemrod et al. (2020) and Bø et al. (2015) find that income disclosure leads to higher levels of tax compliance driven by shifting social norms and concern for reputation. Dwenger and Treber (2018) explicitly study whether public shaming increases tax compliance through social pressure. They exploit the introduction of a naming-and-shaming policy in Slovenia to show that taxpayers reduce their tax debt to avoid shaming. Perez-Truglia and Troiano (2018) run a field experiment to study shaming by sending different letters to tax delinquents in the US. They find that increasing the visibility of the delinquency status increases compliance by individuals who owe less than 2,500$, while the effect on individuals with larger debt is negligible.[4]

Lastly, we contribute to the moonlighting literature, which investigates the relationship between politicians' outside earnings and parliamentary activity, quality and corruption. This

[4] See Bursztyn and Jensen (2017) for a survey of the literature on social pressure and shaming effects.

literature shows that allowing moonlighting has ambiguous effects. On the one hand, it might attract more competent politicians, on the other hand these politicians are also more likely to shirk in office (Gagliarducci et al., 2010). Furthermore, politicians connected to private firms might hinder the process of creative destruction and thereby lower productivity (Akcigit et al., 2018). There are also two studies investigating moonlighting of German MPs. Arnold et al. (2014) show descriptively that (reported) outside earnings are not correlated with absence rates and speeches, but negatively correlated with oral contributions and group activities. Becker et al. (2009) find that politicians report less outside income if they face stronger political competition. However, no existing study examines the effect of disclosure rules in a casual manner. Furthermore, we are the first who use administrative tax records to evaluate public disclosure rules affecting politicians.

The remainder of this paper is structured as follows. In Section 3.2, we describe the institutional context and provide more details about the introduction of disclosure rules in 2007, the tightening of these rules in 2013 and briefly describe the German voting system. We describe our different data sources and provide descriptive statistics in Section 3.3. Section 3.4 outlines our empirical strategy for both reforms. In Section 3.5, we present our results both for the introduction and the tightening of the disclosure rules. Last, Section 3.6 concludes.

3.2 Institutional Context

3.2.1 Introduction of Disclosure Rules

Historical background In Germany, both federal and state member of parliament are legally permitted to carry out outside activities besides their political mandate, e.g. lawyers might continue to work within their profession. However, it is clearly stated in §44a of the Members of the Bundestag Act (*Abgeordnetengesetz*) that "the exercise of the mandate of a Member of the Bundestag shall be central to his or her activity". In late 2004, payments to federal MPs by large companies such as Siemens of Volkswagen became the focus of public attention. Subsequently, the German federal parliament passed a law in August 2005 that obliged MPs of the German Bundestag to publicly disclose their outside activities and associated earnings. The purpose of the disclosed information was to "indicate combinations of interests with implications for the exercise of the said mandate". The law was controversial and some MPs filed a lawsuit against it arguing that it would violate their privacy rights and the obligation to public disclosure makes it less attractive to run for office for citizens from certain occupations such as for example entrepreneurs.

Table 3.1: General disclosure requirements

(A) Outside Activities	
remunerated activity during the term of the mandate	e.g. speech
functions in enterprises	e.g. supervisory board
functions in public corporations and institutions	e.g. board of trustees
functions in clubs, associations and foundations	e.g. development aid agency or foundations
shareholdings in private corporations or partnerships	e.g. law firm
(B1) Outside Earnings (EP 16 and 17)	
level 0	income up to 1,000€
level 1	income between 1,000€ and 3,500€
level 2	income between 3,500€ and 7,000€
level 3	income over 7,000€
(B2) Outside Earnings (EP 18)	
level 0	income up to 1,000€
level 1	income between 1,000€ and 3,500€
level 2	income between 3,500€ and 7,000€
level 3	income between 7,000€ and 15,000€
level 4	income between 15,000€ and 30,000€
level 5	income between 30,000€ and 50,000€
level 6	income between 50,000€ and 75,000€
level 7	income between 75,000€ and 100,000€
level 8	income between 100,000€ and 150,000€
level 9	income between 150,000€ and 250,000€
level 10	income over 250,000€
(C) Frequency and Time Frame	
once, monthly or yearly	starting and ending date
(D) Source	
company's name and location	

Notes: We ignore the information on donations. The name of lawyer's clients are not revealed due to existence of lawyer-client-confidentiality. Shareholdings in private corporations only need to be reported if a MP holds more than 25% and no information about received outside earnings needs to be provided (no information about level, frequency and time frame of the activity). For more details we refer to 'Code of Conduct for Members of the German Bundestag'. Reported earnings and activities are published on webpages of the German Bundestag and in *Amtliches Handbuch*.

Private and public disclosure Until the final decision of the Federal Supreme Court, the President of the German Bundestag (*Bundestagspräsident*) decided that outside activities and earnings would have to be privately disclosed to the administration of the Bundestag, but would not be publicly disclosed. In July 2007, the lawsuit was narrowly defeated by a tied court and MPs were forced to publish their sources and levels of outside earnings on web pages of the German Bundestag. To conclude, starting in 2005 federal MPs privately disclose their information and from 2007 (retroactively to 2005 and onwards) all information was publicly disclosed.

Outside activities and associated earnings are published on webpages of the German Bun-

destag. Table 3.1 summarizes the disclosure rules.[5] Disclosure obligations involve publication of (i) each outside activity, (ii) corresponding outside earnings per activity, (iii) its frequency and (iv) its source. Disclosed earnings are determined by the gross amounts paid, including expenses, compensations and the value of benefits in kind, while deductions are not included. Therefore, the amount of earnings from an activity is therefore not necessarily equal to earnings that are taxes. Not all kinds of outside earnings need to be disclosed, for example stock options or shareholdings in private corporations, if they are lower than 25%, are exempt. In addition, activities with associated earnings of less than 1,000€ also need not be reported.

The amount of outside earnings are published in income levels. Earnings below 1,000€ are classified as level 0, those between 1,000€ and 3,500€ were referred to as level 1, outside earnings between 3,500€ and 7,000€ were called level 2, while level 3 described outside earnings of above 7,000€. In addition, the law required MPs to assign the respective source to each outside activity. Appendix Figure C.1 shows a screenshot of the webpage of an MP. Top-coding at 7,000€ was criticized since MPs might cover their well-paid activities and declare it as level 3. Nevertheless, various watchdog organizations and the media made extensive use of the published data in subsequent years.

The enforcement of the law works as follows. Every MP has to submit all outside activities and associated income levels, time frame and frequency, and its source to the President of the German Bundestag within three months. These data are then published on the individual websites of the respective MP that are administered by the German Bundestag. If a MP misreports or does not report at all, the violation will be made public and a fine has to be paid. Sanctions include cuts in their enumeration of up to 50%. In addition, considerable cost of reputation is added to the monetary fine, since these cases are widely discussed in the media.[6]

3.2.2 Tightening of Disclosure Rules

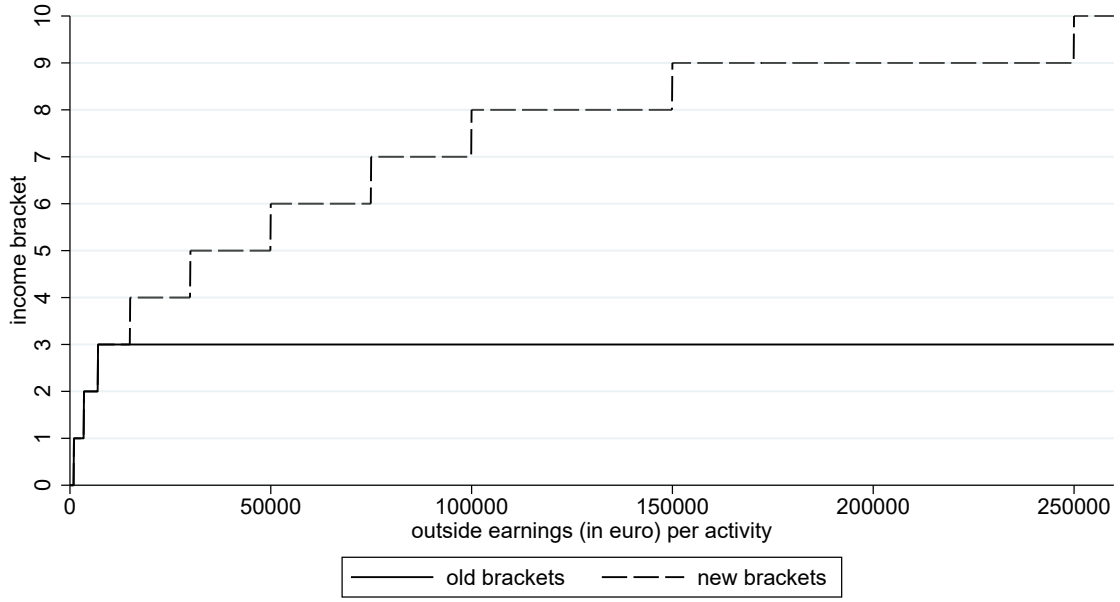
Historical background In 2012, the former German Minister of Finance Peer Steinbrück was nominated as candidate for chancellor for the upcoming federal election. Subsequently, it was pointed out by the media that he was the highest-earning member of parliament by giving a large number of highly-paid speeches.[7] Since most of his outside activities were top-censored, i.e. above 7,000€, his outside earnings were not appropriately reflected in the reporting scheme. This created a prolonged public debate about possible reforms of the reporting requirements

[5] The interested reader can find an English version of the Code of Conduct for Members of the German Bundestag online (Bundestag, 2013).

[6] This has already happened twice, most notably to the former minister of the interior, Otto Schily, in 2008. As an attorney, he argued that the rule would violate his client's privacy rights. In the end he had to pay a 22,000€ fine.

[7] There were even cases of him missing votes in parliament when giving a paid speech (Spiegel, 2010).

Figure 3.1: Visualization of both reforms and the underlying bracket structure



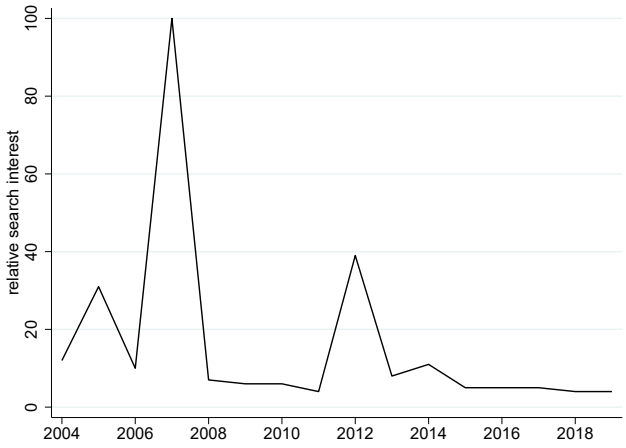
Notes: This figures visualizes the bracket structure of both reforms. The solid line refers to the first reform, where every activity that is remunerated with more than 7,000€ is is categorized as level 3. The dashed graph shows the bracket structure under the second reform and thereby the increase in disclosure of outside earnings to voters.

throughout 2012 with Google searches spiking (see Figure 3.2). Using a digitized database of all parliamentary speeches, we also show that the use of the phrase "outside earnings" in speeches by federal MPs spikes in 2012 (see Figure 3.2). Following this debate, the federal parliament passed a stricter version of the disclosure law in March and came into force in September 2013. MPs could already anticipate the tightening of disclosure law, and we therefore treat 2012 as the reform year for the second reform.

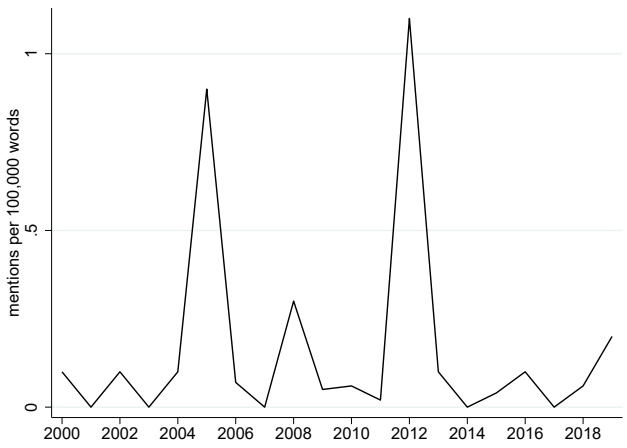
Tightening of disclosure rules The new law aimed to provide more detailed information on high-earning MPs. More specifically, seven new income categories were added to the reporting scheme, so that top-censoring occurred at 250,000€ instead of 7,000€. This makes it possible to distinguish between a MP earning moderate amounts and top-earners. Figure 3.1 visualizes the bracket structure of both reforms. The solid line refers to the first reform, where every activity that is remunerated with more than 7,000€ is is categorized as level 3. The dashed line shows the bracket structure under the new regime and thereby the increase in disclosure of outside earnings to voters. As a reference, federal MPs receive around 90,000€ as a yearly salary for their work as a politician across our period under study.

For the disclosure rules to be effective, there has to be sufficient attention paid to the

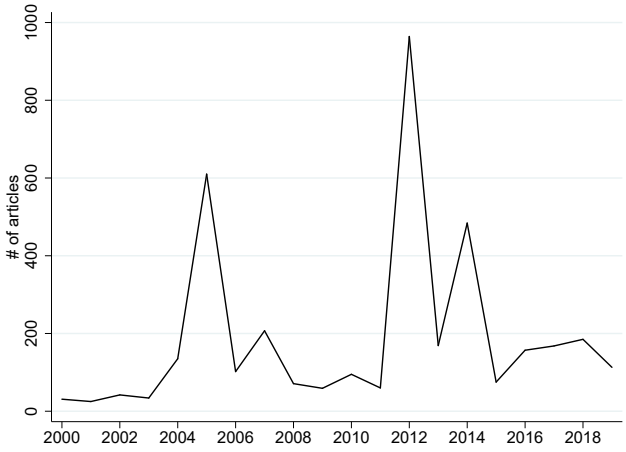
Figure 3.2: Interest in outside activities and earnings



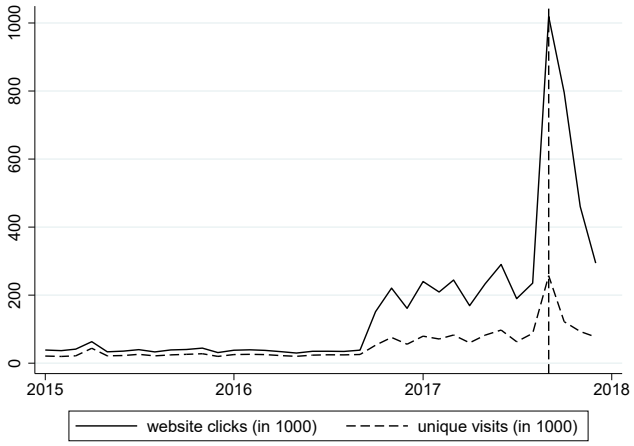
(a) Google searches



(b) Mentions in parliamentary speeches



(c) # of articles mentioning outside earnings



(d) clicks on web pages of the Bundestag

Notes: Panel (a) plots the search interest relative to the highest point in the chart for the selected region in the specified time period. The value 100 stands for the highest popularity of this search term. Source: Google Trends; search term: 'Nebeneinkünfte' (engl: outside earnings); Search Period: 01.01.2004-31.12.2019 in Germany. Panel (b) plots the number of times outside earnings were mentioned in speeches held in parliament per 100,000 words. Source: Die Zeit. Panel (c) plots the number of articles mentioning outside earnings of politicians from the newspaper archive *GENIOS*. Panel (d) plots website clicks and unique visitors (in 1000) on the webpages of the German Bundestag from January 2015 to January 2018 on a monthly basis. The solid line indicate the federal election in September 2017. Source: Deutscher Bundestag (own freedom of information request of 18.11.2019).

reported earnings. This can either be archived through the media, which made extensive use of the reported earnings, or by citizens themselves. To test the first channel, we plot the number of articles mentioning politicians outside earnings found in the newspaper archive GENIOS from 2000 to 2019 in Figure 3.3(c). One can clearly see the spikes in articles in 2005 and 2012 when the two big scandals happened. More generally, the number of articles clearly increased after MPs had to disclose their earnings. To test whether citizens themselves look up their

MPs earnings, we obtain data on unique visitors and clicks on the webpages of the Bundestag where the earnings are reported.[8] As one can see in Figure 3.3(d), the number of clicks and unique visitors increases one year before the federal election in September 2017. There were 61.7 million eligible voters and 47.0 million voters, implying a turnout of 76.2%. In the month of election clicks spike at roughly 1,000,000 clicks and 200,000 unique visitors. Together with the large amount of newspaper articles documenting the existence of outside earnings and activities, we argue that sufficient attention was and still is paid to these issues.

3.2.3 Voting System in Germany

The German Bundestag is the national Parliament of the Federal Republic of Germany, while state Parliaments (*Landtage*) are the legislative bodies for the individual German states. The competence of legislation is split between the 16 State Parliaments and the Federal Parliament. Elections for the German Bundestag as well as for the German State Parliaments are based on a "personalized" proportional representation system. Its goal is to combine the advantages of both proportional representation and majority voting system. Each citizen has two votes. The first vote is directly attributed to a candidate representing her electoral district. As there are 299 federal electoral districts, the same number of mandates in the Bundestag are distributed to the candidates winning the plurality of first votes in their districts (directly elected candidates). The second vote supports a political party at the national level. Based on their share of the second vote, political parties send their candidates from predefined electoral lists into the federal parliament. The electoral lists are determined by the parties at the state level. This way 299 additional mandates are distributed to the parties who have received at least 5 percent of the valid second votes.[9] The Bundestag is elected for four years, while State Parliament elections are held every five years.

In our analysis, we will distinguish between MPs that are directly elected and those who entered parliament through the party list. In particular, directly elected MPs should face a higher level of electoral accountability since voters have the possibility to punish (or reward) them directly given their published information on outside earnings and activities. Furthermore, we will compare MPs with a safe ranking on the electoral list to those with a more insecure ranking. Again, the less secure the rank is, the higher the degree of electoral accountability should be.

[8] Unfortunately, the data is only available from January 2015 to January 2018.

[9] If a party receives more mandates via the first vote than the second vote, all directly elected candidates gain additional seats in the Bundestag (*Überhangmandate*). To keep proportional representation intact, parties whose share of candidates lies below their share of second votes are also given additional seats (*Ausgleichsmandate*).

3.3 Data

We employ the German Taxpayer Panel for the years 2001 to 2014 (henceforth called *TPP*), which comprises the universe of German tax returns. In addition, we collect publicly disclosed outside activities and earnings for the years 2005 to 2017 as well as publicly available information on demographics, committee membership and voting statistics (henceforth called *reported data*). The two data sets have distinct advantages and drawbacks. The TPP allows us to precisely measure outside income before *and* after the reforms both for federal *and* state MPs. By this, we can causally evaluate the reforms in a difference-in-difference setting. The main drawback of the TPP is the low number of demographic and political variables. Given the strict data protection rules when working with tax return records, we cannot identify individuals' names or party affiliations. In contrast to the tax return data, our reported data offers a rich set of demographic and political variables, but the publicly disclosed information on earnings are imprecisely measured. Given the nature of the reported data, we can only observe federal MPs after the reform and state MPs are not covered at all. We use the reported data to provide some suggestive evidence on the characteristics of outside activities and demographics, but also to support potential mechanisms. Importantly, we are not allowed to combine these two data sets and both will be evaluated separately.

3.3.1 German Taxpayer Panel

The German Taxpayer Panel (TPP) covers all tax units for the period 2001 – 2014. It is an administrative data set collected by German tax authorities, provided and administered by the German Federal Statistical Office. The unit of observation is a tax unit, i.e., either a single individual or a couple filing jointly. It contains all information necessary to calculate a taxpayer's annual income tax, including basic socio-demographic characteristics such as age, gender, state of residence, marital status, as well as detailed information on income sources and tax base parameters such as work related expenses and (claimed and realized) deductions on a yearly level. Hence, the advantage of tax return data lies in its precise measurement of pre- as well as post-reform income related variables. However, it does not contain of information about the specific type of outside activity (e.g. speech or ongoing work as a lawyer) or personal information (e.g. party affiliation).[10]

Treatment and control group Our empirical strategy compares federal MPs (treatment

[10] Data access is subject to very strict data security rules and we only work with these data via remote-access. Every single request requires a confidentiality check. Moreover, it is impossible to combine these data with any other information.

group) to state MPs (control group). Now, we outline how we determine the two groups in the TPP. First, we identify all members of federal, state, and EU parliament by having positive income from parliamentary activities. Next, we gather data on the remuneration and election dates of all 16 state parliaments as well as well the federal and European parliament from 2001 to 2014.[11] Since state MPs earn less than federal MPs, we discriminate between the two groups within state-year cells. Until 2009, members of the European parliament received the same amount of remuneration as federal MPs. To identify those units we exploit an increase in their compensation in 2009 due to a EU-wide harmonization of their salaries. Hence we drop observations whose income from parliamentary activities discontinuously jumps in 2009 by the reform-induced amount.[12] Further, we drop households, in which both the head and the spouse are MPs since they could be part of both the treatment and the control group.[13] Next, we exploit the panel structure of our data to exclude individuals who just entered parliament for a given year, since we would wrongly classify their pre-politician earnings as outside earnings. MPs leaving parliament receive a transitional payment (*Übergangsgeld*). We make use of the fact that (i) most MPs leave parliament after elections, and (ii) the transitional payment is lower than the regular salary. This allows us to pinpoint MPs whose income from parliamentary activities drops right after a state or federal election. We classify these MPs as dropouts.[14] As a robustness check, we will report results both with and without dropouts. Finally, we drop all MPs from the three German city-states (Berlin, Hamburg and Bremen) since being an MP is only a part-time job in their state parliaments (so-called *Feierabendparlamente*).

In 2013, Bavaria was the first state that introduced a public disclosure law for its state MPs. One year later, five further states introduced similar laws (see Table C.8). Therefore, we exclude observations from these states when disclosure laws where in effect to avoid a contamination of our control group. In Figure 3.3, we verify the accuracy of our allocation mechanism and compare the amount of units identified in the tax data with the actual number of units that are present in parliament. We match the number of state and federal parliamentarians quite closely.

Outcome variables We capture disclosed outside earnings as closely as possible. We take

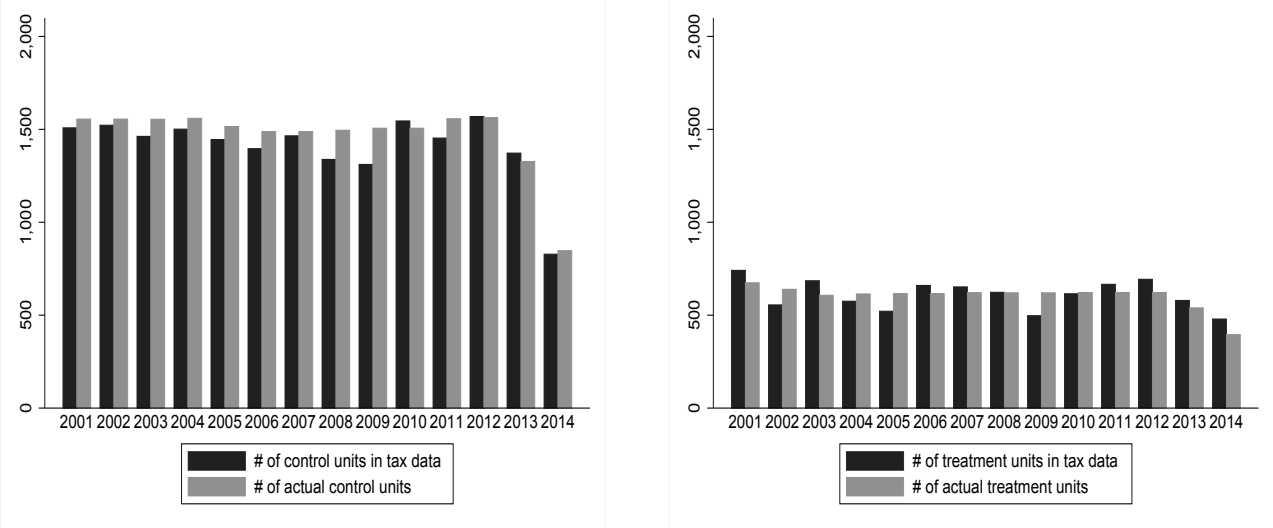
[11] Appendix Figure C.2 plots the average remuneration for the federal, EU and all state parliaments over our sample period.

[12] We can identify about two thirds of the 99 EU parliamentarians since one third newly enters the European parliament and is therefore indistinguishably from newly entering federal MPs. Note, that this induces a bias towards zero since a (small) part of the treatment group is not actually treated. Over our sample period there were no changes with respect to income disclosure for members of the European parliament.

[13] This involves only a very small number of couples in our sample period. Including them does not change our results.

[14] Federal MPs receive one additional month of transitional payments for each year they spend in parliament. The transitional payments are capped at 18 months. Starting with the second month after leaving parliament, transitional payments are reduced one to one by all other income a former MP receives.

Figure 3.3: Comparison between tax data and actual numbers



(a) control units (b) treatment units

Notes: Panel (a) shows the number of state MPs that are identified in the tax data (in black) and the actual or expected number of MPs (in grey). Panel (b) shows the number of federal MPs that are identified in the tax data (in black) and the actual number of MPs (in grey). We exclude parliamentarians from Berlin, Hamburg, and Bremen for both groups. We further exclude those units that newly enter parliament and those who leave parliament in a given year. Hence, in our baseline estimations, we only consider 'full year' units such that our results do not get contaminated by e.g. individuals directly entering employment right after leaving parliament. Source: German tax return data, 2001-2014 (Taxpayer Panel, TPP)

advantage of the fact that earnings are divided into seven different types of income (e.g. income from business operations or income from employed work) in the German income tax system. Our main outcome is the total income from sources that MPs have to disclose. This amounts to all income from (i) salaries and wages (ii) (non-corporate) businesses and self-employment (iii) agriculture and forestry, as well as other sources. We will also evaluate the effect on each of the categories (i) to (iii) separately. Furthermore, we use rental income as a placebo outcome since such income does not need to be disclosed.[15]

3.3.2 Reported Data

Our second data set consists of several publicly available sources (henceforth called *reported data*). The most important part of this data are the reported (and disclosed) outside earnings and activities from web pages of the German federal parliament. We enrich this data with further demographic and political variables. Our reported data covers every MP who was at least present in one of the following three legislative periods of the German Bundestag: 16th

[15] We do not consider capital income in our analysis, since MPs were not required to disclose such earnings and investment income is only observable until 2009 in the tax data.

legislative period (2005-2009), 17th legislative period (2009-2013) and 18th legislative period (2013-2017).[16] In the following, we describe the different data sources in greater detail.

Demographic variables Using the handbook of German MPs, we extract a number of demographic variables. We observe a politician's name, gender, age, marital status, and number of children. Additionally, we know whether a politician has a PhD degree and their resident state. We classify a politician's (former) occupation into ten groups. Importantly, as opposed to the tax data, we know the party membership of each MP. For our sample period about half of MPs are part of a center-right party (CDU/CSU and FDP), while the other half is a member of one of the left-wing parties (SPD, Greens and The Left). Moreover, we group MPs by their political experience into three categories: newcomers (first term), those serving for two to three terms, and MPs with four or more terms in parliament. Lastly, we construct dummies for MPs that leave (or join) parliament in the middle of an election period since they have less time to accumulate outside earnings. Summary statistics of all these variables can be found in Appendix Table C.3.

Political and electoral variables A MP can be voted into the *Bundestag* either via party list or direct ballot (see Section 3.2.3). To capture this distinction, we construct a dummy for being elected directly. We also create a dummy for MPs who entered through a safe rank on the party list (above-median ranking) as opposed to those that where placed on a less safe rank (below-median ranking). Furthermore, when a MP ran for direct ballot in one of the 299 electoral districts, we obtain her own as well as her party's vote share in that district. Then, we calculate the vote margin of each MP as the difference to the second-placed candidate for winning candidate and the difference to the first placed candidate for all other candidates. To account for political offices and to capture a politician's policy expertise and interest more accurately, we construct dummies for membership in one each of the 23 committees of the German federal parliament. In addition, to capture the rank and status of the MP, we create dummies for being part of party leadership and for being a committee chair, respectively. Summary statistics are again displayed in Appendix Table C.4.

Published data on outside earnings We collect every disclosed activity, its income level (0 to 3 for election period 16 & 17 and 0 to 10 for election period 18), its starting and end date as well as frequency (monthly, yearly, once), and the respective employer. Table 3.2 provides information about the number of MPs with at least one activity and positive outside

[16] Table C.2 provides an overview about these three election periods under study as well as the composition of MPs in federal parliament by party.

Table 3.2: Number of MPs with at least one activity and positive outside earnings

	EP16		EP17		EP18		Total	
	N	in %	N	in %	N	in %	N	in %
MPs who report at least one activity	573	89.81	581	89.11	582	88.45	1736	89.12
MPs with positive outside earnings	241	37.77	250	38.34	252	38.30	743	38.14

Notes: This table provides an overview about federal MPs who report outside activities and who report outside earnings for the election periods 16-18 and the average across all three election periods. All percentages refer to the total amount of MPs for a given election period. Source: Reported Data, own calculations.

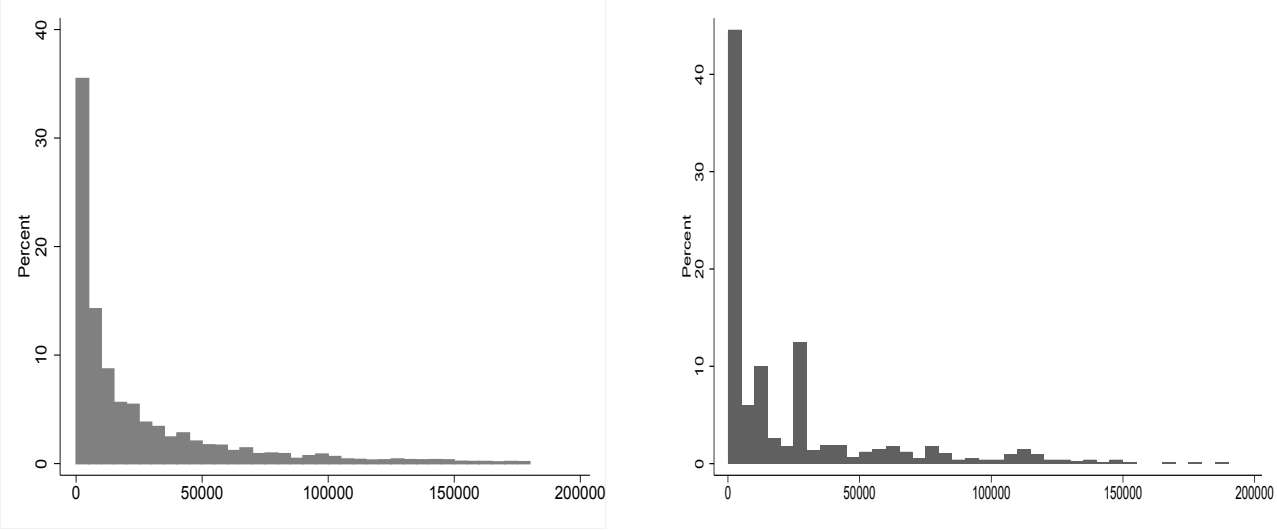
earnings. 89.12% of all MPs report an activity and 38.14% report positive outside earnings. This is due to the fact that many activities are voluntary work and thus not remunerated. In Appendix Table C.7 we display the distribution of each activity's bracket and frequency. 18% of all activities are assigned level three of higher across all election periods. 94% of all activities happen only once and only 2% and 4% of all activities happen on a yearly or monthly basis.

To determine a value of outside earnings, we assign the mean value of each bracket to every activity (e.g. an activity with level 0 is measured with 500€). The value assigned to the last bracket is determined by polynomial extrapolation, i.e. an activity with level 3 is assigned 9,500€ (see Appendix Figure C.3). Since the addition of 7 new levels in election period 18 mechanically increases this measure, we code every activity of level 4 or higher as a level 3 activity. More precisely, an activity with level 0 is assigned a value of 500€ , level 1 2,250€ , level 2 5,250€ and level 3 and above 9,500€ .[17] This is likely to underestimate the true level of outside earnings, but ensures comparability over time. In a last step, we calculate the total amount of reported outside earnings of every federal MP for a given election period and divide it by four to ensure comparability to the yearly tax data.

Published data on outside activities The composition of the main activities that MPs undertake are displayed in Appendix Table C.6. 32% pursue a remunerated activity, 40% hold functions in enterprises and 59% hold functions in public corporations. The most popular remunerated activities are classified as law (10% of all MPs report at least one law activity), 10% of all MPs have at least one management and consulting activity and 9% were giving at least one speech. Typical functions in enterprises are member of advisory board (*Mitglied des Beirates*) or member of supervisory board (*Mitglied des Aufsichtsrates*). 11% of all MPs report shareholdings in private corporations with a share larger than 25%, but we cannot observe their income from these shareholdings.

[17] As a robustness check, we also use a lower bound measure, where we we assign the lower threshold of 7,000€ to level 3 (and above) activities.

Figure 3.4: Distribution of outside earnings



(a) tax data

(b) reported data: baseline

Notes: Panel (a) displays the distribution of (positive) outside earnings from federal parliamentarians excluding the top 2% for privacy reasons based on the tax return data. Panel (b) shows the corresponding distribution for the baseline measure of outside earnings based on the reported data. Source: German tax return data, 2001-2014 (Taxpayer Panel, TPP) (Panel (a)); Reported Data EP 16 - 18 (Panel (b))

3.3.3 Descriptive Analysis: Outside Earnings

The reported data consists of 1,952 MP-election period observations and covers election period 16-18 of the German Bundestag. We observe 1,108 individual MPs, 264 of which are present throughout all election periods.[18]

Outside earnings Figure 3.4 plots the distribution of federal MPs outside earnings both from the reported data as well as from the tax data. Outside earnings is extremely unequally distributed in both data sets. The outside earnings from the tax data closely traces a pareto distribution, while the reported distribution exhibits bunching at different points. Between these bunching points, one can see the missing mass that is caused by the bracket reporting system. In our tax data, half of those MPs who do have positive earnings, have less than 10,000€ and around 30% have more than 30,000€ across the period under study. Next, we compare the outside earnings that were publicly disclosed with the actual outside earnings that we can observe in the tax data.

[18] We provide details of the composition of the German Bundestag for the election periods under study in the Appendix.

Table 3.3: Descriptive statistics: outside earnings (reported data & tax data)

	mean	sd	min	max	N
tax data					
all MPs					
outside earnings	29,358	146,151			27,974
wages & salaries	14,633	136,463			27,974
business & self-employment	11,762	113,943			27,974
renting	-986	17,880			27,974
other sources	2,963	15,770			27,974
federal MPs					
outside earnings	21,546	75,968			8,537
wages & salaries	8,230	42,613			8,537
business & self-employment	10,390	59,358			8,537
renting	-1,830	14,363			8,537
other sources	2,926	16,702			8,537
state MPs					
outside earnings	32,789	167,837			19,437
wages & salaries	17,445	161,184			19,437
business & self-employment	12,364	130,909			19,437
renting	-616	19,212			19,437
other sources	2,980	15,344			19,437
reported data					
federal MPs					
outside earnings: baseline	9,677	26,957	0	251,875	1,952
outside earnings: lower bound	8,478	23,205	0	227,562	1,952

Notes: Both panels refer to yearly values. The upper panel reports earnings based on the German tax return data, 2001-2014 (Taxpayer Panel, TPP). Outside earnings amounts to all income from (i) salaries and wages, (ii) business and self-employment income and (iii) other sources (except for income from parliamentary activities). Income from renting is our placebo outcome. Due to privacy reasons minimum and maximum values are omitted in the tax return data. In out reported data, outside earnings are calculated as follows: *baseline*: an activity with level 0 is assigned a value of 500€ , level 1 2,250€ , level 2 5,250€ and level 3 and above 9,500€ . In our *lower bound* definition, we assign a value of 7,000€ for each activity with level 3 and above. Source: Outside earnings are based on reported data for the election periods 16, 17 and 18 (lower panel);

Table 3.3 shows that the mean outside earnings in the tax data is around 29,000€ across all MPs. Federal MPs receive on average 21,000€ of outside earnings, while state MPs earn on average 32,000€. The large difference might be surprising since the focus of the political debate is usually on federal MPs. Possible explanations might be the lower public attention placed on state MP's or simply because they still have a closer relation to their hometown and thereby their initial occupation. The major income source is business and self-employment income for federal MPs, while state MPs earn (on average) the most from wages and salaries. The mean in the reported data is around 10,000€. The values reported in the tax data are almost twice as high as our baseline measure from the reported data. This confirms one frequent criticism of

Table 3.4: Outside earnings: correlations

	(1) outside earnings	(2) outside earnings	(3) outside earnings	(4) outside earnings	(5) outside earnings	(6) outside earnings
left-wing	-7,408***					-3,624***
	(1,488)					(1,402)
female		-7,267***				-3,815***
		(1,429)				(1,302)
East Germany			-5,987***			-6,755***
			(1,375)			(1,486)
age between 50 and 60				1,307		-122
				(1,380)		(1,438)
age 60 above				3,191*		1,188
				(1,919)		(1,988)
terms: 2 - 3					1,270	84
					(1,272)	(1,429)
terms: >3					2,069	-1,882
					(1,703)	(2,127)
controls						Yes
N	1,952	1,952	1,952	1,952	1,952	1,952
# politicians	1,108	1,108	1,108	1,108	1,108	1,108

Notes: The outcome variable is outside earnings as described in Section 3.3.2. SPD, Greens and The Left are coded as left-wing (parties). Controls include all variables in Tables C.3 and C.4 for which we have full observations. Robust standard errors clustered at the individual level. * p<0.1, ** p<0.05, *** p<0.01. Source: Reported data for EP 16 - 18 (2005–2017).

the public disclosure law. The bracket system, and in particular the highest bracket, mask the real extent of moonlighting that politicians engage in.

Correlations We classify SPD, Greens and the Left Party as left-wing parties and show that they earn less compared to members of other parties, a result often found in the existing literature (Becker et al., 2009; Eggers and Hainmueller, 2009). Table 3.4 shows that the unconditional difference amounts to about 7,400€ per year. This difference shrinks to 3,600€ when including all control variables, such as for example their former occupation, but is still statistically significant and of a economically meaningful size. Furthermore, in our sample both female and East German MPs earn significantly less outside earnings. Meanwhile, there is no significant difference by age and experience once we control for all other variables.[19]

3.4 Empirical Strategy

In this section, we outline our empirical strategy. First, we describe both our simple difference-in-differences setting and our identification strategy. Furthermore, we extend our model to a dynamic difference-in-difference strategy. Second, to analyze who particularly responded to disclosure of outside earnings and activities, we run a quantile regression approach and

[19] Appendix Figure C.4 shows that there is also substantial variation in outside earnings by committee membership. MPs in the economics, agriculture and exterior committee earn on average over 13,000€, while members of the environmental and digital committee earn 3,000€ and less.

we use different income categories as outcome variables. Last, we explore the mechanisms behind our results using the reported data by comparing MPs with different levels of electoral accountability.

3.4.1 Difference-in-Differences Strategy

Since 2005 Federal MPs are obliged to privately disclose their outside activities and earnings. Starting from 2007 and onwards these information is publicly disclosed (Also retroactively to 2005). We exploit the fact that members of the federal parliament (*Bundestag*) are affected by disclosure rules, while members of state parliaments (*Landtag*) do not face such legal requirements. Thus, members of the federal parliament are our treatment units and members of state parliaments form our control group. This setup gives rise to a difference-in-difference design by comparing federal to state MPs before and after the reform. This identification strategy will uncover the casual effect of the public disclosure law if the assumption of parallel trends between the treatment and control group holds. We implicitly validate this assumption using a dynamic difference-in-difference approach.

Our baseline estimation is structured as follows: Let Y_{ist} be an outcome of politician i resident in state s in year t. We then estimate

$$Y_{it} = \beta Treat_i Reform_t + \gamma_i + \lambda_{st} + \epsilon_{it} \tag{3.1}$$

where $Treat_i$ is a dummy taking the value one if i is a federal MP and $Reform_t$ is an indicator equal to 1 from 2007 onwards. We also include individual fixed effects γ_i to control for potentially unobserved and time-constant features of MPs. The state-year fixed effects λ_{st} absorb aggregate movements as well as state-specific shocks such as local economic conditions. Finally, we cluster our standard errors at the individual level to allow for serial correlation. The coefficient of interest is β, which identifies the casual effect of the public disclosure law. Our sample period runs from 2001 to 2009 for the first reform. Note that, since this is classical 2x2 difference-in-difference setup, we do not have to assume homogeneous treatment effects for our estimator to be consistent (Goodman-Bacon, 2021).

We evaluate the tightening of the public disclosure law in much the same manner as its introduction with one exception. We drop observations in which state MPs were also subject to disclosure rules (see Section 3.3.1). Next, we estimate equation 3.1 on the sample from 2010 to 2014 with the reform dummy being one for $t \geq 2012$. Standard errors are again clustered on the individual level.

Dynamic difference-in-difference As mentioned above, we also estimate a more dynamic

version of equation 3.1 both to test for pre-trends and to allow for dynamic post-treatment effects. To do so, we define a set of dummy variables $\mathbb{1}_{k=t}$, which takes the value one if k equals t and zero otherwise. To estimate the effects of the introduction of public disclosure rules, we run the following equation:

$$Y_{it} = \sum_{k=2001}^{2005} \beta_k Treat_i \mathbb{1}_{k=t} + \sum_{l=2007}^{2009} \beta_l Treat_i \mathbb{1}_{l=t} + \gamma_i + \lambda_{st} + \epsilon_{it} \tag{3.2}$$

where we omit the interaction of the 2006 dummy to normalize our estimates to the pre-reform year. Therefore, $\beta_k \ \forall k \in \{2001, ..., 2005\}$ refer to differences in trends between the treatment and control group before the reform, while $\beta_l \ \forall l \in \{2007, ..., 2009\}$ represent the dynamic treatment effects.

Analogous to equation 3.2, we adjust the dynamic difference-in-difference equation such that we check for pre- and post-treatment effects for the second reform:

$$Y_{it} = \sum_{k=2010}^{2010} \beta_k Treat_i \mathbb{1}_{k=t} + \sum_{l=2012}^{2014} \beta_l Treat_i \mathbb{1}_{l=t} + \gamma_i + \lambda_{st} + \epsilon_{it} \tag{3.3}$$

where we omit the interaction of the 2011 dummy to normalize our estimates to the pre-reform year. Again, β_{2010} refers to differences in trends between the treatment and control group before the reform, while $\beta_l \ \forall l \in \{2012, ..., 2014\}$ represent the dynamic treatment effects.

3.4.2 Who responds to the Disclosure of Outside Earnings and why?

Increased transparency makes politicians more accountable. In which way politicians adjust their earnings depend on the preferences of voters. If voters perceive outside income negatively, increased transparency could make politicians more accountable such that they reduce outside activities. We discuss direct ways to test for the effect of electoral accountability in the reported data in Section 3.4.3.

Income components Income disclosure by politicians might have counteracting effects on different categories of outside income. On the one hand, the effect depends on the preferences of voters on incomes from different sources. For example, Campbell and Cowley (2015) show via a survey experiment that voters do not penalize business owners or the self-employed for continuing their business. On the other hand, the literature on behavioral responses towards taxation shows that the self-employed can more easily adjust their labor supply and also the reporting of their income (Saez et al., 2012).

Another possible behavioral effect can occur if income disclosure affects tax compliance. By

increasing the possibility to detect evasion behaviour, income disclosure laws incentives tax payers to declare their true income (Slemrod et al., 2020; Bø et al., 2015). Given strict third-party reporting standards in Germany, we expect this possible effect only to be present for income from business operations and self-employment, since these income categories are self-declared by the tax payer. Both of these effects should (at least partially) materialize already in 2005 when private disclosure was applied and politicians had to assume that there is a decent chance for public disclosure to be applied retroactively. In contrast, if the effect is only observed from 2007, it is more likely that it is connected to the information that was publicly released.

Social norms towards having outside work might have changed after the introduction of the public disclosure law. Initially, the very conservative top-coding at 7,000€, has prevented voters to distinguish between a high- and moderate-earning MP and might have lead voters to underestimate the true extent of outside earnings. Therefore, from a voter's point of view it might have become more acceptable to have a second job as a politician. The second reform, which introduced more brackets and thereby increased the amount of information available to voters, however, could have had the opposite effect. In response, politicians might then reduce the amount of outside income.

Public disclosure could also have changed a previous social norm of not pursuing outside activities among MPs to a market transaction by putting a price on it (Gneezy and Rustichini, 2000).[20] Given that MPs are paying a price, which is the reporting requirement itself, they might engage in more outside work. Moreover, politicians might have misperceived social norms and learned from the behavior of their peers, which causes them to update their beliefs about the acceptability of outside earnings (Bursztyn et al., 2020). Last, the reported income could also be used as a signal of skill to (certain) voters. This could be potentially heterogeneous with some MPs wanting to highlight the importance of their mandate by having no outside jobs, while others explicitly start to have outside jobs to signal competence.

Quantile regression As already seen in Figure 3.4, outside earnings of politicians are highly unequally distributed. To shed light into the full distribution of outside earnings, we use (unconditional) quantile regressions. Whereas ordinary least squares regressions allow us to estimate the effect of a given variable at the mean, quantile regressions tell us about the effect of a policy change on the entire distribution of outside earnings.

We apply the estimator suggested by Firpo et al. (2009) to estimate the effect of the reform on all nine deciles of the outside earnings distribution. We apply this estimator to both data periods: 2001 – 2009 (first reform) and 2010 – 2014 (second reform). The results are particular

[20] This is also connected to the concept of moral licensing, where an individual, after doing something perceived as morally good, i.e. a politician being transparent about their outside earnings, it gives herself license to do something that is perceived to be morally bad, i.e. increasing her outside earnings (Merritt et al., 2010).

interesting for the second reform, since it has changed only the bracket structure. More precisely, until 2012 every activity that was remunerated with more than 7,000€ was top-coded and appeared as level 3 on the web pages of the German Bundestag. After the tightening of the rules, activities that are remunerated with more than 250,000€ are top-coded. Therefore, we expect most of the effect to be concentrated at the top of the distribution.

3.4.3 Mechanisms: Electoral Accountability

To further investigate the mechanism of the reform, we look at variation in electoral accountability. As explained in Section 3.2.3, we exploit the fact that there are two ways to become a federal MP in Germany: direct ballot election and party lists. Since it is impossible to differentiate between the two groups of MPs in the tax data, we will test this hypothesis using the reported data. As we do not have a control group in this data set, all evidence has to be considered suggestive.

Election via direct ballot or party list Politicians, who enter parliament by direct ballot election, are arguably more accountable to voters. In case for any perceived misbehaviour, voters have the opportunity to directly vote specific politicians out of office. In contrast, voters cannot (directly) vote out specific politicians that enter trough the party list. Therefore, directly elected MPs are more electorally accountable and should react more strongly to the reform if electoral accountability matters. We test the prediction by looking at the subset of electoral districts, from which the second-placed candidate also entered parliament (through the party list). This allows us to compare directly elected MPs to their runner-ups in the following way:

$$Y_{ide} = \beta_e D_{ie}^{direct} + \delta X_{ie} + \gamma_d + \epsilon_{ide} \quad \forall\, e \in \{16, 17, 18\} \tag{3.4}$$

where Y_{ide} are outside earnings for MP i in district d in election period e. D_{ie}^{direct} is a dummy for being directly elected, and γ_d are district fixed effects ensuring that we identify the effect within electoral districts. We estimate this equation both for the two election periods before the second reform and for the period after the second reform.

We expect β_e to be negative for all election periods, since they are subject to a higher level of electoral accountability. If the tightening of the disclosure rules, which went into effect, in election period 18, increased electoral accountability, directly elected MPs should reduce their outside income relative to MPs entering parliament through the party list. That is, we expect β_e to be even more negative in election period 18.

Safe and unsafe ranking on party list In contrast, MPs entering parliament via party list are

only at risk to be voted out of office if they are close to the marginal rank, meaning the last rank which gets into parliament. Therefore, we also compare MPs with a safe list rank to those with an unsafe rank. Given the higher risk of being voted out of office for MPs with an unsafe rank, we argue that they are subject to a higher level of electoral accountability. Since party lists are organized at the state-party level, we construct a dummy $D_{ie}^{unsaferank}$ that takes the value one if a politician has an above median rank. For example, 22 politicians entered through the list of the Bavarian Social Democrats in election period 18. According to our classification, those ranked 1 to 11 had safe list ranks, whereas ranks 12 to 22 were unsafe. We then estimate the following equation:

$$Y_{ispe} = \beta_e D_{ie}^{unsaferank} + \delta X_{ie} + \gamma_{sp} + \epsilon_{ie} \quad \forall\, e \in \{16, 17, 18\} \tag{3.5}$$

where Y_{ispe} are outside earnings for MP i in state s and party p and election period e. γ_{sp} are state-party fixed effects controlling for the (potentially) different assignment procedures of the state-level party associations. Similar to above, β_e should generally be negative and become even more negative in election period 18 if electoral accountability plays a mediating role.

3.5 Results

3.5.1 Introduction of the Public Disclosure Law

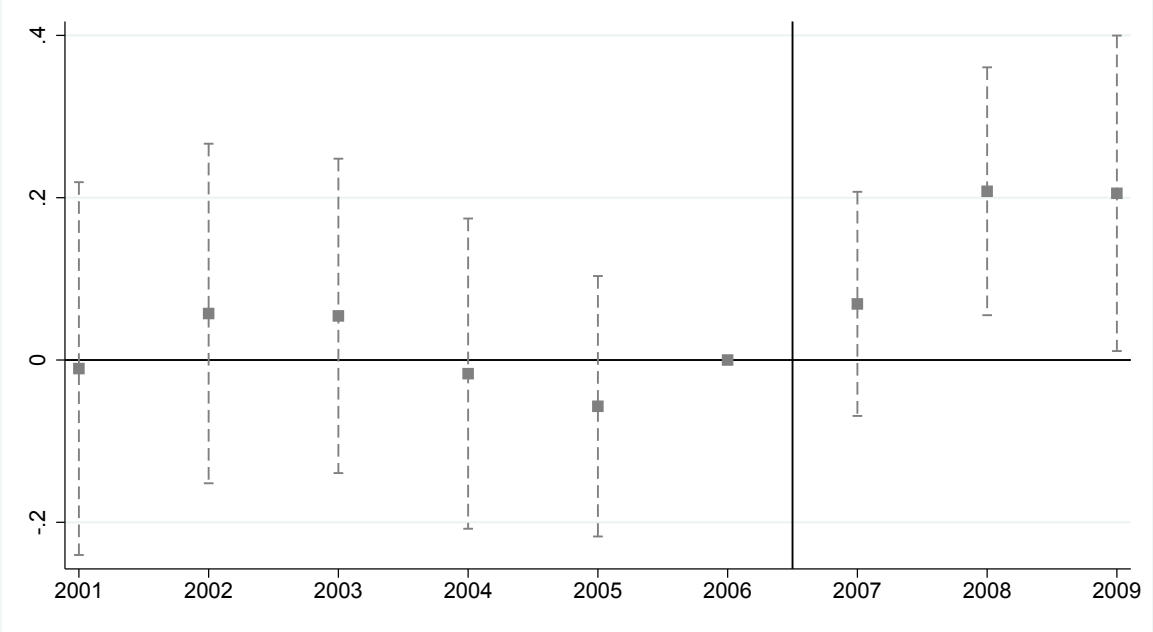
Table 3.5: Introduction of the disclosure law: extensive and intensive margin

	(1) log outside income	(2) log outside income	(3) outside income > 0	(4) outside income > 0
treatment x reform	0.155**	0.153**	0.049***	0.045***
	(0.064)	(0.066)	(0.017)	(0.017)
politician FE	Yes	Yes	Yes	Yes
state-year FE	Yes	Yes	Yes	Yes
w/o dropouts		Yes		Yes
N	14,135	12,955	19,993	18,412
# politicians	3,189	3,013	3,652	3,546

Notes: This tables displays estimates from equation 3.1 using log outside earnings (columns 1 & 2) and a dummy for positive outside earnings (columns 3 & 4) as outcome variables. Robust standard errors are clustered at the individual level. * p<0.1, ** p<0.05, *** p<0.01. Source: German tax return data, 2001-2009 (Taxpayer Panel, TPP)

Baseline results We first present the results from our baseline Difference-in-Differences approach (see equation 3.1). Table 3.5 shows the causal effects of the introduction of disclosure laws. Outside earnings did actually *increase* by about 15%. Also, the probability of having positive outside income increased by 4.5 percentage points. Both of these effect are statistically

Figure 3.5: Introduction of the disclosure law: dynamic difference-in-difference



Notes: This graphs displays the coefficients $\beta_t \; \forall \; t \in \{2001, ..., 2009\}$ and the corresponding 95% confidence intervals estimated by equation 3.2 using outside earnings as the outcome variable. Robust standard errors are clustered at the individual level. Source: German tax return data, 2001-2009 (Taxpayer Panel, TPP)

significant at conventional levels. One potential concern is that we include politicians who just dropped out of parliament in our sample conflating outside earnings with their regular income. To test this possibility, we exclude these MPs (labelled as *dropouts*) from our sample (see column (2) and (4) in Table 3.5). This leaves our estimates almost unchanged.

Figure 3.5 visualizes the estimates of our dynamic difference in differences approach (see equation 3.2). The effect only emerges after the introduction of public disclosure in 2007. Importantly, there is no evidence for any significant differential trend between the treatment and control group before the reform. This is reinforcing the parallel trends assumption underlying our research design. In addition, we do not observe any differential trend in the time period of private disclosure from 2005 to 2006. Politicians are only reacting to *public*, but not to *private* disclosure. The effect in 2007 is positive, but insignificant. In the following years, the effect becomes stronger and significant at conventional levels.

Income components To disentangle the total effect of an increase in outside earnings, we apply our baseline difference-in-difference setup to different income categories (see equation 3.1). Table 3.6 shows the results for wages & salaries (column 1 and 2), business & self-employment (column 3 and 4), other sources (column 5 and 6) and last, renting as our placebo

Table 3.6: Introduction of the disclosure law: income categories

income category	wages & salaries		business & self-employment		other sources		renting (placebo)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	log income	income > 0	log income	income > 0	log income	income > 0	log income	income > 0
treatment x reform	0.089	0.001	0.193**	0.037**	0.060	0.009	0.095	0.018
	(0.089)	(0.011)	(0.089)	(0.018)	(0.111)	(0.014)	(0.179)	(0.014)
politician FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
state-year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
w/o dropouts	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	5,608	18,412	9,046	18,412	4,463	18,412	3,799	18,412
# politicians	1,518	3,546	2,319	3,546	1,229	3,546	1,550	3,546

Notes: This tables displays estimates from equation 3.1 using log outside earnings and a dummy for positive earnings from wages and salaries (column 1 & 2), business operations and self-employment (column 3 & 4), forest and agriculture and other sources (column 5 & 6), and renting (column 7 & 8) as outcome variables. Robust standard errors are clustered at the individual level. * p<0.1, ** p<0.05, *** p<0.01. Source: German tax return data, 2001-2009 (Taxpayer Panel, TPP)

outcome (column 7 and 8). The results show that the increase is solely driven by income from business and self-employment, which increased by 19.3% at the intensive margin and 3.7 percentage points at the extensive margin. All other coefficients are positive and insignificant. Lastly, rental income, which was not affected by the disclosure law, does also not react to the reform. This increased credibility of that the measured effect is solely driven by the introduction of the disclosure law and not by some other shock occurring at the same time.

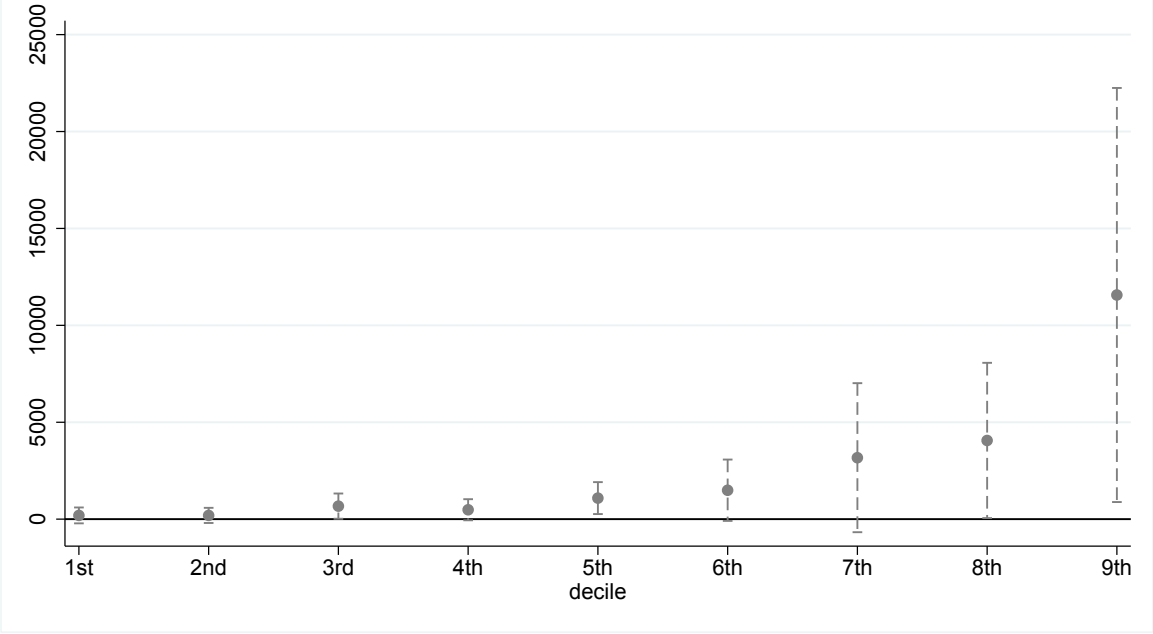
Now, we discuss if an increase in tax compliance might be an explanation of why the introduction of public disclosure leads to an increase in outside earnings, particularly in business and self-employment income. We do not think that tax compliance (or previous tax evasion) is a driving force behind our results. The timing of the effect is not consistent with an increase in tax compliance. If politicians were concerned about being caught evading taxes, they should have already reacted in 2005 when private disclosure was introduced. Since it was known that the privately disclosed income would become public retroactively, MPs should have anticipated the possibility of public disclosure and, at least partially, increased their tax compliance starting in 2005. Moreover, tax evasion is a criminal offence and caught MPs not only would loose their mandate, they would also face severe penalties.

Instead, the increase in 2007 is consistent with a change in social norms towards outside activities and earnings. These social norms could only have changed when outside earnings became *public*, not when they were privately disclosed. As the reported amounts were kept artificially low by top-coding at 7,000€, this could have induced voters (and subsequently politicians) to view outside earnings less negatively. This mechanism is also consistent with the increase being driven by income from self-employment as this income category has been shown to be acceptable by voters (Campbell and Cowley, 2015).

Social norms might also change when previously intrinsically motivated is replaced by extrinsically motivated behavior. Gneezy and Rustichini (2000) show in a field experiment that

the introduction of a fine for parents who pick up their children late from a day-care center actually increased late-coming. Before the fine, it was simply a social norm to be on time and afterwards it was perceived as a market transaction. Apply this finding to our setting, it might be that it was a social norm not to have little (or no) outside earnings. After the policy change, the price an MP pays for earning outside income, is the duty to report it. Therefore, since politicians pay the price, earning outside income becomes more acceptable simply because they report it. Another explanation might be that social norms were initially misperceived. Bursztyn et al. (2020) define the term 'pluralistic ignorance'. It refers to a situation where most people privately hold an opinion, but they incorrectly believe that most other people hold the contrary opinion, and end up acting against their own view. When politicians believe having outside jobs are stigmatized, they might be reluctant to reveal their private views to others for fear of social sanctions. If most politicians act this way, they might end up believing their private views are only shared by a small minority at most. In our setting, MPs might have misperceived the norms regarding outside activities since it was not public knowledge. Although the private view of MPs was that having outside earnings is not necessarily a bad thing, they might have been reluctant to have any because they thought that others disapprove such behavior. When outside income became public and were seen to be wide-spread, they engage more in such behavior.

Figure 3.6: Introduction of the disclosure law: quantile regression



Notes: This graphs displays the coefficient β on log of outside earnings and the corresponding 95% confidence interval when estimating equation 3.1 using unconditional quantile regression for the first to ninth decile. Robust standard errors are clustered at the individual level. Source: German tax return data, 2001-2009 (Taxpayer Panel, TPP)

Quantile regressions We test whether the effect is driven by different parts of the outside income distribution by conducting (unconditional) quantile regressions on the deciles of the outside earnings distribution. That is, we estimate not the average effect, but the effect on all nine deciles (Firpo et al., 2009). The results are plotted in Figure 3.6. The treatment effect is very small for the lower and middle part of the distribution, whereas the effect on the eighth and ninth decile is considerably larger. This implies that most of the treatment effect is driven by high-income MPs that are likely top-censored.

3.5.2 Tightening of the Public Disclosure Law

Table 3.7: Tightening of the disclosure law: extensive and intensive margin

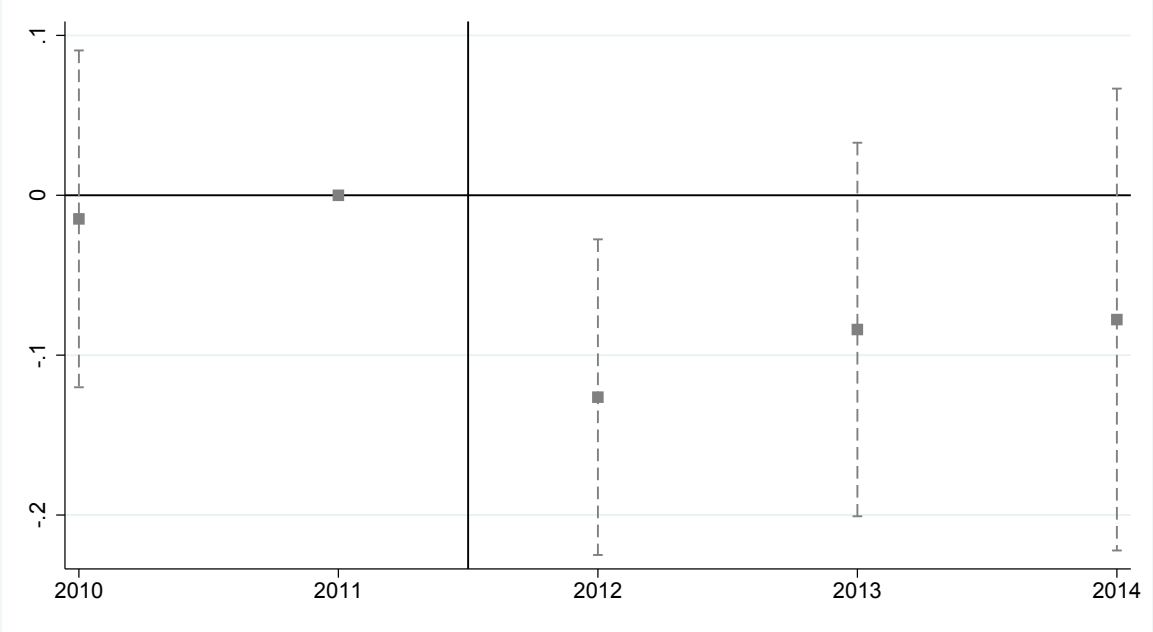
	(1) log outside income	(2) log outside income	(3) outside income > 0	(4) outside income > 0
treatment x reform	-0.092*	-0.096**	0.011	0.008
	(0.047)	(0.048)	(0.013)	(0.013)
politician FE	Yes	Yes	Yes	Yes
state-year FE	Yes	Yes	Yes	Yes
w/o dropouts		Yes		Yes
N	8,622	8,299	11,223	10,849
# politicians	2,716	2,600	3,212	3,096

Notes: This tables displays estimates from equation 3.1 using log outside earnings (columns 1 & 2) and a dummy for positive outside earnings (columns 3 & 4) as outcome variables. Robust standard errors are clustered at the individual level. * p<0.1, ** p<0.05, *** p<0.01. Source: German tax return data, 2010-2014 (Taxpayer Panel, TPP)

Baseline result Recall, that the reform only introduced seven new brackets such that it shifted top-coded incomes from 7,000€ to 250,000€. Therefore, voters are now able to differentiate between medium- and high-earning MPs. Our baseline difference-in-difference estimates using equation 3.1 are presented in Table 3.7. The tightening of disclosure law significantly decreased total outside income by 9.6%, while leaving the extensive margin unchanged. This result is line with the institutional details of the new rules, since the introduction of new brackets did not change the reporting requirements at the extensive margin. As one can see in Figure 3.7, the effect emerges in 2012 with parallel trends between the treatment and control group in the year before. Importantly, the effect occurs before the federal election in 2013 and can therefore not be driven by a changed composition of the federal parliament.

Income categories When we decompose the total effect into the different income categories, we find that the negative intensive margin effect is driven by a reduction of 15.8% of income from wages and salaries (see column 1 of Table 3.8). We do not find any significant negative effect on self-employment or business income. This is consistent with the tightening of the

Figure 3.7: Tightening of the disclosure law: dynamic difference-in-difference



Notes: This graphs displays the coefficients $\beta_t \; \forall \; t \in \{2010, ..., 2014\}$ and the corresponding 95% confidence intervals estimated by equation 3.3 using outside earnings as the outcome variable. Robust standard errors are clustered at the individual level. Source: German tax return data, 2010-2014 (Taxpayer Panel, TPP)

rules inducing a sizeable transparency effect as this income category is viewed more favourably among voters (Campbell and Cowley, 2015).

We do not find consistent evidence for a change in the other income categories. Similarly to the introduction of the law, we do not find any effect on rental income, which acts as our placebo outcome.

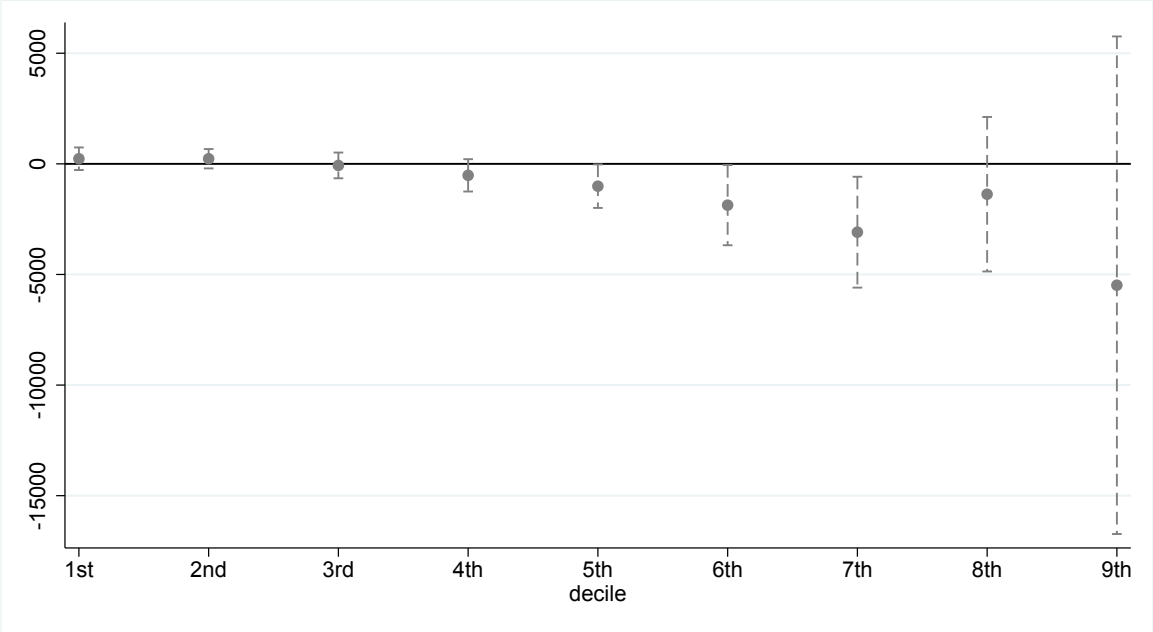
Table 3.8: Tightening of the disclosure law: income categories

income category	wages & salaries		business & self-employment		other sources		renting (placebo)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	log income	income > 0	log income	income > 0	log income	income > 0	log income	income > 0
treatment x reform	-0.158***	-0.000	-0.035	0.034**	-0.116	-0.027**	0.003	-0.017
	(0.052)	(0.009)	(0.064)	(0.015)	(0.073)	(0.011)	(0.095)	(0.012)
politician FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
state-year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
w/o dropouts	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	3,580	10,849	5,808	10,849	3,163	10,849	2,554	10,849
# politicians	1,256	3,096	1,978	3,096	1,064	3,096	964	3,096

Notes: This tables displays estimates from equation 3.1 using log outside earnings and a dummy for positive earnings from wages and salaries (column 1 & 2), business operations and self-employment (column 3 & 4), forest and agriculture and other sources (column 5 & 6), and renting (column 7 & 8) as outcome variables. Robust standard errors are clustered at the individual level. * p<0.1, ** p<0.05, *** p<0.01. Source: German tax return data, 2010-2014 (Taxpayer Panel, TPP)

Quantile regression Given that the introduction of the new income brackets mainly affected top-earning MPs, we expect the treatment effect to be concentrated at the top of the distribution. We test this hypothesis by estimating quantile regressions for every decile of the distribution. As one can see in Figure 3.8, the effect is very small and insignificant for the first deciles and then becomes larger the further one goes along the distribution.

Figure 3.8: Tightening of the disclosure law: quantile regression



Notes: This graphs displays the coefficient β on log of outside earnings and the corresponding 95% confidence interval when estimating equation 3.1 using unconditional quantile regression for the first to ninth decile. Robust standard errors are clustered at the individual level. Source: German tax return data, 2010-2014 (Taxpayer Panel, TPP)

Electoral accountability Next, we explore potential mechanisms of the decrease in outside earnings following the tightening of the disclosure rules.[21] As we argued before, we expect the effect to be stronger the more accountable politicians are to their voters. Since we cannot test this hypothesis in the tax data, we make use of the reported data. In a first step, we compare MPs elected by direct ballot and their runner-up peers, who entered via party list. We additionally add electoral district fixed effects to only compare the winner of a direct election and the second placed candidate. Panel A in 3.9 shows that there was no significant difference between MPs elected by direct ballot and MPs joining via the party list before election period 18.[22] In election period 18, when the new rules became effective, the difference increases to

[21] We cannot use the reported data for the first reform since we cannot observe report outside income before the reform.

[22] The negative, but insignificant coefficients are consistent with the introduction of the law causing minor

Table 3.9: Electoral accountability

	(1) EP 16 outside earnings	(2) EP 17 outside earnings	(3) EP 18 outside earnings
Panel A: directly elected			
D^{direct}	-8,501	-6,112	-13,997***
	(5,653)	(10,725)	(5,282)
electoral district FE	Yes	Yes	Yes
controls	Yes	Yes	Yes
N	318	238	404
# politicians	318	238	404
Panel B: unsafe rank			
$D^{unsaferank}$	-2,790	-605	-5,907**
	(2,471)	(3,968)	(2,360)
party-state FE	Yes	Yes	Yes
controls	Yes	Yes	Yes
N	562	578	593
# politicians	562	578	593

Notes: The outcome variable is outside earnings as described in Section 3.3.2. In Panel A, the sample contains only MPs from districts, where both the first- and second-placed candidate entered parliament to estimate equation 3.4. In Panel B, we use only MPs that were ranked on a party list to estimate equation 3.5. Controls refer to all variables in Tables C.3 and C.4. Robust standard errors. * p<0.1, ** p<0.05, *** p<0.01 Source: reported data EP 16 - 18

roughly 14,000€ and becomes significant at the 1% level (see column (3) of Table 3.9). This suggests that directly elected MPs reduced their outside earnings more dramatically because of electoral concerns. We observe a similar pattern for MPs inhabiting more and less safe party list ranks. Before election period 18, there is no significant difference between those, who just made it in, and MPs, who were relatively safe (see columns (4) and (5) of Table 3.9). After the reform, we observe a significant difference of about 6,000€. Both results are robust to the lower bound measure of outside earnings (see Appendix Table C.9). Taken together, these estimates provide support for the mechanism of electoral accountability.

3.6 Conclusion

In this paper, we evaluate the effects of public disclosure rules on politicians outside earnings. Since 2005, members of the German federal parliament are obliged to publish their outside activities and associated earnings in a (top-coded) bracket-based reporting scheme on web pages of the German Bundestag. By law, the execution of the mandate of an MP should be central to his or her activity. The intention of the reform was to indicate any conflicts of interests

electoral pressure.

that might have implications on the political work. First, we exploit the introduction of this policy as exogenous variation. We can observe both federal and state MPs in administrative tax records before and after the policy change. Thereby, we use unaffected state MPs as a control group in a difference-in-difference design. Second, we can differentiate between private and public disclosure. Since 2005, information on outside activities and earnings was initially privately disclosed to the administration of the Bundestag. In 2007, the Federal Court decided that the information must be publicly disclosed involving a public disclosure of the information for the years 2005 and 2006. Third, we evaluate a second reform that tightened existing rules by introducing seven new income brackets in the reporting scheme causing reported outside income to be top-coded at 250,000€ instead of 7,000€. Last, given the sparse number of demographic variables in the tax return data and the inability to merge this data with any other data set, we collect various other data sets to uncover potential mechanisms behind our findings.

We show that the introduction of public disclosure of outside activities and earnings lead to an increase of 15.3% in outside earnings. This effect is mainly present at the top end of the distribution and is largely driven by income from self-employment and businesses. Importantly, the effect only emerges when disclosure is public, not when it is private. Therefore, it is unlikely that it is driven by increased tax compliance since MPs should have anticipated that there is a significant chance that their privately disclosed income would become public retroactively. A more likely explanation is a change in social norms regarding outside income that made the practice more acceptable. Next, we find that the tightening of the disclosure decrease outside income, in particular, income from salaries and wages drop by 15.8%, while other income categories are largely unaffected. Using the reported data on outside income, we provide evidence that electoral accountability might explain the decrease in outside income. More specifically, we show that outside income of directly elected MPs drops relative to MPs joining via party list after the reform. Similarly, MPs with an unsafe rank on the party list decrease their outside income relative to MPs with a safe rank. Taken together, our results suggest that the effect of income disclosure laws crucially depend on their exact implementation. If the disclosed information is very limited and lacks precision such that voters cannot identify top-earners, public income disclosure can increase outside activities and earnings and thereby, might increase the risk of exertion of influence.

Our project faces various limitations. Earnings in the tax data does not necessarily reflect the time an MP has invested into his or her outside work. Activities differ in the type of activity (for example, giving a speech or being a member of a supervisory board), the time invested, and the degree of interdependence with third parties, all of which we cannot observe in the tax data. Therefore, we cannot make statements about the impact on the quality of parliamentary

work or potential conflicts of interest.

Appendix to Chapter 2

A.1 Meta (Estimation) Sample

A.1.1 List of Included Studies

Aarbu, K. O. and Thoresen, T. O. (2001), 'Income Responses to Tax Changes – Evidence from the Norwegian Tax Reform', *National Tax Journal* 54(2), 319–335.

Almunia, M., and Lopez-Rodriguez, D. (2019), 'The Elasticity of Taxable Income in Spain: 1999-2014', *Working Paper (July 2019) forthcoming: SERIEs - Journal of the Spanish Economic Association*.

Arrazola, M., de Hevia, J., Romero, D. and Sanz-Sanz, J. F. (2014), 'Personal Income Tax Reforms and the Elasticity of Reported Income to Marginal Tax Rates: An Empirical Analysis Applied to Spain', *Working Paper Series in Public Finance provided by Victoria Business School (September 2014)*.

Arrazola-Vacas, M., Sanz-Sanz, J. F., , Rueda-Lopez, N. and Romero-Jordan, D. (2015), 'Reported Gross Income and Marginal Tax Rates: Estimation of the Behavioural Reactions of Spanish Taxpayers', *Applied Economics* 47(5), 466–484.

Auten, G. and Carroll, R. (1999), 'The Effect of Income Taxes on Household Income', *Review of Economics and Statistics* 81(4), 681–693.

Auten, G., Carroll, R. and Gee, G. (2008), 'The 2001 and 2003 Tax Rate Reductions: An Overview and Estimate of the Taxable Income Response', *National Tax Journal* 61(3), 345–364.

Auten, G. and Joulfaian, D. (2009), 'The Taxable Income Elasticity of High-Income Taxpayers: Evidence from a Long Panel', *Working Paper (May 2009)*.

Auten, G. and Kawano, L. (2014), 'How the Rich Respond to Anticipated Tax Increases: Evidence from the 1993 Tax Act', *Working Paper (May 2014)*.

Bakos, P., Benczúr, P. and Benedek, D. (2010), 'The Elasticity of Taxable Income: Estimates and

Flat Tax Predictions using the Hungarian Tax Changes in 2005', *Working Paper (September 2010)*.

Berg, K. and Thoresen, T. (2018), 'Problematic Response Margins in the Estimation of the Elasticity of Taxable Income' , *Working Paper (June 2018)*.

Blomquist, S. and H. Selin (2010), 'Hourly wage rate and taxable labor income responsiveness to changes in marginal tax rates', *Journal of Public Economics* 94 (11-12), 878-889.

Bosch N. and de Boer, H.-W. (2019) 'Income and Occupational Choice Responses of the Self-employed to Tax Rate Changes: Heterogeneity across Reforms and Income' , *Labour Economics* 58, 1-20.

Burns, S. K. and Ziliak, J. P. (2017), 'Identifying the elasticity of taxable income', *The Economic Journal* 127, 297–329.

Carey, S., Creedy, J., Gemmell, N. and Teng, J. (2015), 'Estimating the Elasticity of Taxable Income in New Zealand', *Economic Record* 91(292), 54–78.

Carroll, R. (1998), 'Do Taxpayers Really Respond to Changes in Tax Rates? Evidence from the 1993 Tax Act', *Office of Tax Analysis Working Paper* (79), 145–77 (November 1998).

Chetty, R., Friedman, J. N., Olsen, T. and Pistaferri, L. (2011), 'Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records', *Quarterly Journal of Economics* 126, 749–804.

Creedy J., Gemell, N. and Teng J. (2018) 'Income Effects and the Elasticity of Taxable Income' , *New Zealand Economic Papers* 52(2), 158-203.

Díaz-Caro, C. and Onrubia, J. (2018) 'How do Taxable Income Responses to Marginal Tax Rates Differ by Sex, Marital Status and Age? Evidence from Spanish Dual Income Tax' , *Economics: The Open-Access, Open-Assessment E-Journal* 2018-67(12), 1-25.

Doerrenberg, P., Peichl, A. and Siegloch, S. (2017), 'The Elasticity of Taxable Income in the Presence of Deduction Possibilities', *Journal of Public Economics* 151, 41–55.

Ericson, P., Flood, L. and Islam, N. (2015), 'Taxes, Wages and Working Hours', *Empirical Economics* 49(2), 503–535.

Gelber, A. M. (2014), 'Taxation and the Earnings of Husbands and Wives: Evidence from Sweden', *Review of Economics and Statistics* 96(2), 287–305.

Giertz, S. H. (2007), 'The Elasticity of Taxable Income over the 1980s and 1990s', *National Tax Journal* pp. 743–768.

Giertz, S.H. (2010), 'Panel Data Techniques and the Elasticity of Taxable Income', *National Tax*

Association 2009 Annual Conference Proceedings

Giertz, S. H. (2010), 'The Elasticity of Taxable Income during the 1990s: New Estimates and Sensitivity Analyses', *Southern Economic Journal* 77(2), 406–433.

Gottfried, P. and Schellhorn, H. (2004), 'Empirical Evidence on the Effects of Marginal Tax Rates on Income – The German Case', *IAW-Diskussionspapiere* (15) (February 2004).

Gottfried, P. and Witczak, D. (2009), 'The Responses of Taxable Income Induced by Tax Cuts - Empirical Evidence from the German Taxpayer Panel', *IAW-Diskussionspapiere* (57) (November 2009).

Gruber, J. and Saez, E. (2002), 'The Elasticity of Taxable Income: Evidence and Implications', *Journal of Public Economics*, vol. 84(1), pp. 1–32.

Hansson, Å. (2007), 'Taxpayers' Responsiveness to Tax Rate Changes and Implications for the Cost of Taxation in Sweden', *International Tax and Public Finance* 14(5), 563–582.

Harju, J. and Matikka, T. (2016), 'The Elasticity of Taxable Income and Income-Shifting: What is "real"and what is not?', *International Tax and Public Finance* 23(4), 640–669.

Heim, B. T. (2010), 'The Responsiveness of Self-Employment Income to Tax Rate Changes', *Labour Economics* 17(6), 940–950.

Heim, B. T. and Mortenson, J. A. (2016), 'The Effect of Recent Tax Changes on Taxable Income: Correction and Update', *Journal of Policy Analysis and Management* 37(3), 686-694.

He, D., Peng, L. and Wang, X. (2018) 'Understanding the Elasticity of Taxable Income: A Tale of Two Approaches' , *Working Paper (September 2018)*

Hermle, J. and Peichl, A. (2018) 'Jointly Optimal Taxes for Different Types of Income' , *Working Paper (September 2018)*

Holmlund, B. and Söderström, M. (2011), 'Estimating Dynamic Income Responses to Tax Reform', *The BE Journal of Economic Analysis & Policy* 11(1), 1–38.

Igdalov, S., Frish, R. and Zussman, N. (2017) 'The Wage Response to a Reduction in Income Tax Rates: The 2003–2009 Tax Reform in Israel' , *Working Paper (December 2017)*

Jongen, E. L. and Stoel, M. (2019), 'The Elasticity of Taxable Labour Income in the Netherlands', *Working Paper (January 2019)*.

Kemp J. (2017), 'The Elasticity of Taxable Income: The case of South Africa', *Working Paper (May 2017)*.

Kiss, A. and Mosberger, P. (2014), 'The Elasticity of Taxable Income of High Earners: Evidence from Hungary', *Empirical Economics* 48(2), 883–908.

Kleven, H. J. and Schultz, E. A. (2014), 'Estimating Taxable Income Responses Using Danish Tax Reforms', *American Economic Journal: Economic Policy* 6(4), 271–301.

Kopczuk, W. (2005), 'Tax Bases, Tax Rates and the Elasticity of Reported Income', *Journal of Public Economics* 89(11), 2093–2119.

Kopczuk, W. (2015), 'Polish Business Flat Tax and its Effect on Reported Incomes: Preliminary Analysis', *Working Paper (June 2015)*.

Kumar, A. and Liang, C.-Y. (2017), 'Estimating Taxable Income Responses with Elasticity Heterogeneity', *Working Paper (March 2017)*.

Lehmann, E., Marical, F. and Rioux, L. (2013), 'Labor Income Responds Differently to Income-Tax and Payroll-Tax Reforms', *Journal of Public Economics* 99, 66–84.

Lindsey, L. B. (1987), 'Individual Taxpayer Response to Tax Cuts: 1982–1984: With Implications for the Revenue Maximizing Tax Rate', *Journal of Public Economics* 33(2), 173–206.

Looney, A. and Singhal, M. (2006), 'The Effect of Anticipated Tax Changes on Intertemporal Labor Supply and the Realization of Taxable Income', *Working Paper (July 2006)*.

Massarrat Mashhadi, N. and Werdt, C. (2012), 'Estimating Dynamic Income Responses to Tax Changes', *Working Paper (December 2012)*.

Matikka, T. (2016), 'The Elasticity of Taxable Income: Evidence from Changes in Municipal Income Tax Rates in Finland', *Scandinavian Journal of Economics* 120(3), 943-973 .

Miyazaki, T. and Ishida, R. (2016), 'Estimating the Elasticity of Taxable Income: Evidence from Top Japanese Taxpayers', *Working Paper (October 2016)*.

Moffitt, R. and Wilhelm, M. O. (2000), 'Taxation and the Labor Supply Decisions of the Affluent', *Does Atlas Shrug?: The Economic Consequences of Taxing the Rich* p. 193.

Mortenson, J. A. (2016), 'All Income is Not Created Equal: Cross-Tax Elasticity Estimates in the United States', *Working Paper (March 2016)*.

Pirttilä, J. and Selin, H. (2011a), 'Income Shifting within a Dual Income Tax System: Evidence from the Finnish Tax Reform of 1993', *The Scandinavian Journal of Economics* 113 (1), 120–144.

Saez, E. (2003), 'The Effect of Marginal Tax Rates on Income: a Panel Study of 'Bracket Creep', *Journal of Public Economics* 87(5), 1231–1258.

Saez, E., Slemrod, J. and Giertz, S. H. (2012), 'The Elasticity of Taxable Income with Respect to Marginal Tax Rates: A Critical Review', *Journal of Economic Literature* 50(1), 3–50.

Schmidt, T.-P. and Müller, H. (2012), 'Die Elastizität des zu versteuernden Einkommens in Deutschland: Eine empirische Untersuchung auf Basis des deutschen Taxpayer-Panels', *Working Paper (June 2012)*.

Sillamaa, M.-A. and Veall, M. R. (2001), 'The Effect of Marginal Tax Rates on Taxable Income: A Panel Study of the 1988 Tax Flattening in Canada', *Journal of Public Economics* 80(3), 341–356.

Singleton, P. (2011), 'The Effect Of Taxes On Taxable Earnings: Evidence From The 2001 And Related U.S. Federal Tax Acts', *National Tax Journal* 64(2), 323.

Thomas, A. (2012), 'The Elasticity of Taxable Income in New Zealand: Evidence from the 1986 Tax Reform', *New Zealand Economic Papers* 46(2), 159–167.

Thoresen, T. O. and Vattø, T. E. (2015), 'Validation of the Discrete Choice Labor Supply Model by Methods of the New Tax Responsiveness Literature', *Labour Economics* 37, 38–53

Weber, C. (2014a), 'Measuring Treatment for Tax Policy Analysis', *Working Paper (November 2014a)*.

Weber, C. (2014b), 'Toward Obtaining a Consistent Estimate of the Elasticity of Taxable Income Using Difference-in-Differences', *Journal of Public Economics* 117, 90–103.

Werdt, C. (2015), 'The Elasticity of Taxable Income for Germany and its Sensitivity to the Appropriate Model', *Working Paper (Jan 2015)*.

A.1.2 Distribution of Estimates by Study: Published vs Working Paper

Table A.1: Distribution of Estimates by Study: Published vs Working Paper

Published articles			**Working Papers**		
Study	**# estimates**	**in %**	**Study**	**# estimates**	**in %**
Aarbu and Thoresen (2001)	8	0.47	Almunia and Lopez-Rodriguez (2019)	83	4.83
Arrazola-Vacas et al. (2015)	27	1.57	Arrazola et al. (2014)	8	0.47
Auten and Carroll (1999)	20	1.16	Auten and Joulfaian (2009)	24	1.40
Auten et al. (2008)	10	0.58	Auten and Kawano (2014)	12	0.70
Blomquist and Selin (2010)	10	0.58	Bakos et al. (2010)	21	1.22
Bosch (2019)	44	2.56	Berg and Thoresen (2018)	4	0.23
Burns and Ziliak (2017)	68	3.95	Carroll (1998)	12	0.70
Carey et al. (2015)	6	0.35	Giertz (2010)	72	4.19
Chetty et al. (2011)	6	0.35	Gottfried and Schellhorn (2004)	11	0.64
Creedy et al. (2018)	3	0.17	Gottfried and Witczak (2009)	15	0.87
Diaz-Caro and Onrubia (2018)	29	1.69	He et al. (2018)	4	0.23
Dörrenberg et al. (2017)	16	0.93	Hermle and Peichl (2018)	4	0.23
Ericson et al. (2015)	5	0.29	Igdalov et al. (2017)	19	1.10
Gelber (2014)	16	0.93	Jongen and Stoel (2019)	99	5.76
Giertz (2007)	69	4.01	Kemp (2017)	18	1.05
Giertz (2010)	127	7.38	Kopczuk (2015)	30	1.74
Gruber and Saez (2002)	35	2.03	Kumar and Liang (2017)	21	1.22
Hansson (2007)	30	1.74	Looney and Singhal (2017)	15	0.87
Harju and Matikka (2016)	14	0.81	Massarrat Mashhadi and Werdt (2012)	9	0.52
Heim (2010)	14	0.81	Miyazaki and Ishida (2016)	8	0.47
Heim and Mortenseon (2018)	14	0.81	Mortenson (2016)	42	2.44
Holmlund and Söderström (2011)	36	2.09	Schmidt and Müller (2012)	18	1.05
Kiss and Mosberger (2014)	15	0.87	Weber (2014)	5	0.29
Kleven and Schultz (2014)	114	6.63	Werdt (2015)	11	0.64
Kopczuk (2005)	91	5.29			
Lehmann et al. (2013)	18	1.05			
Lindsey (1987)	14	0.81			
Matikka (2018)	18	1.05			
Moffitt and Wilhelm (2000)	39	2.27			
Pirttilä and Selin (2011)	10	0.58			
Saez (2003)	91	5.29			
Saez et al. (2012)	24	1.40			
Sillamaa and Veall (2001)	25	1.45			
Singleton (2011)	25	1.45			
Thomas (2012)	8	0.47			
Thoresen and Vatto (2015)	21	1.22			
Weber (2014)	35	2.03			
Total (published)	1155	67.15%		565	32.85%

Note: The data covers only observations with a given or calculable standard error. # estimates denote the number of estimates collected in a particular study and the corresponding percentage share shows the share a study has in the final sample.

A.2 Additional Descriptives

A.2.1 Summary Statistics by Income Concept

Table A.2: Distributions of Estimates by Income Concept

Tax Base	Mean	Median	Std. Dev.	Obs.	Studies
Before Deductions	0.287	0.185	1.212	940	46
Adjusted Gross Income	0.319	0.236	2.607	278	
Gross Income	0.312	0.230	0.542	414	
Earned Income	0.125	0.062	0.257	129	
Self employed Income	0.675	0.858	0.510	20	
Wage Income	0.230	0.114	0.744	99	
After Deductions	0.403	0.353	0.564	780	41
Taxable Income	0.4	0.343	0.578	737	
Taxable Earnings	0.445	0.444	0.186	43	
Total	0.34	0.270	0.975	1720	61

Note: The data covers only observations with a given or calculable standard error.

A.2.2 Distribution of Estimates by Country and Income Concepts

Table A.3: Income Concepts by Country

Variable	Adj. G. Income	Gross Income	Taxable Income	Earned Income	Self employed	Wage Income	Taxable Earnings	Total
Canada	15	2	2	2	2	2	0	25
China	0	0	4	0	0	0	0	4
Denmark	0	18	18	78	0	6	0	120
Finland	0	6	17	0	0	19	0	42
France	0	0	0	0	0	18	0	18
Germany	3	20	61	0	0	0	0	84
Hungary	0	0	36	0	0	0	0	36
Israel	0	19	0	0	0	0	0	19
Japan	0	0	8	0	0	0	0	8
Netherlands	99	0	44	0	0	0	0	143
New Zealand	0	0	13	0	0	4	0	17
Norway	0	0	12	21	0	0	0	33
Poland	0	30	0	0	0	0	0	30
South Africa	0	9	9	0	0	0	0	18
Spain	0	53	94	0	0	0	0	147
Sweden	12	26	17	12	0	0	30	97
USA	149	231	402	16	18	50	13	879
Total	278	414	737	129	20	99	43	1,720

Note: The sample covers only observations with a given or calculable standard error.

A.2.3 Distribution of Estimates by Year of Publication

Table A.4: Year of Publication and Published Type

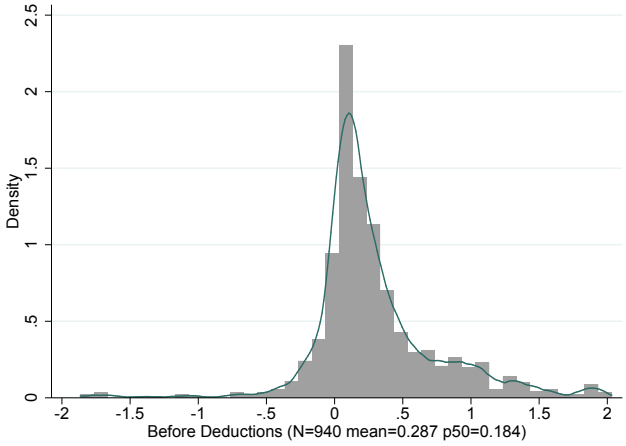
Year of Publication	Working Paper	Published	Total
1987	0	14	14
1998	12	0	12
1999	0	20	20
2000	0	39	39
2001	0	33	33
2002	0	35	35
2003	0	91	91
2004	11	0	11
2005	0	91	91
2006	15	0	15
2007	0	99	99
2008	72	10	82
2009	39	0	39
2010	21	151	172
2011	0	77	77
2012	27	32	59
2013	0	18	18
2014	25	191	216
2015	4	96	147
2016	50	82	124
2017	58	0	58
2018	12	32	44
2019	182	44	226
Total	565	1155	1,720

Note: The sample covers only observations with a given or calculable standard error.

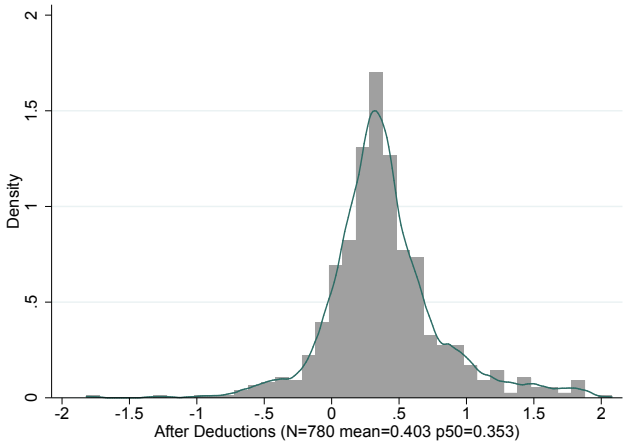
A.3 Distribution of Elasticities and Details on Explanatory Variables

A.3.1 Distribution of Elasticities

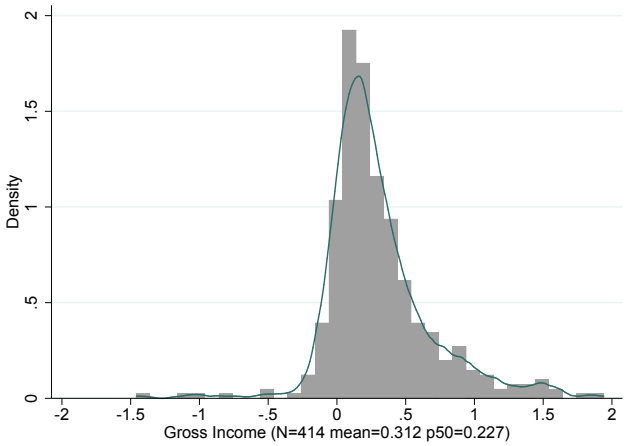
Figure A.1: Distribution of Elasticities



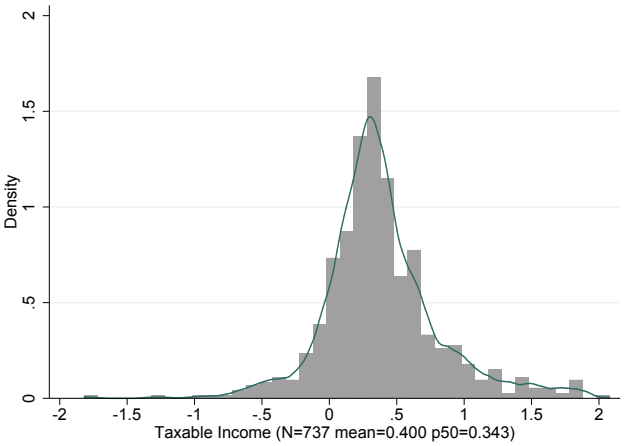
(a) Before Deductions



(b) After Deductions



(c) Gross Income Elasticities



(d) Taxable Income Elasticities

Note: The data cover only observations with a given standard error or z-statistic. I restrict the sample to elasticity estimates that belong to the (a) before deductions subsample or (b) the after deduction subsample. Subfigures (c) and (d) are based on a narrower definition (gross or taxable income respectively).

A.3.2 Explanatory Variables: Details

Regression technique: Most approaches use an Instrument for Δ NTR = $ln\left[\frac{(1-\tau_{it}(z_{it}))}{(1-\tau_{it-k}(z_{it-k}))}\right]$ to achieve a causal relationship:

IV: mechanical tax rate changes: $\Delta ln(1-\tau_{it}^{p}) = ln\left[\frac{(1-\tau_{it}^{p}(z_{it-k}))}{(1-\tau_{it-k}(z_{it-k}))}\right]$, where τ_{it}^{p} is the marginal tax rate that an individual would face given her synthetic income. Example: In year 3, τ_{it}^{p} would be calculated based on income of year two (assume time length of one year). Introduced by Auten and Carroll (1999) / Gruber and Saez (2002) and often referred to as the most standard specification.

IV: (lagged) mechanical tax rate changes: $\Delta ln(1-\tau_{it}^{p,lag})$, where $\tau_{it}^{p,lag}$ is based on income further in the past. $\Delta ln(1-\tau_{it}^{p}, lag) = ln\left[\frac{(1-\tau_{it}^{p}(z_{it-k-lag}))}{(1-\tau_{it-k}(z_{it-k}))}\right]$

IV: other: This category summarizes all other instruments. (1) Blomquist and Selin (2010): They use a single difference and an imputed taxable income $\hat{z_{it}}$ to calculate their instrument: $\left(\frac{1-\tau_{it}(\hat{z_{it}})}{1-\tau_{it-k}(\hat{z_{it-k}})}\right)$. (2) Burns and Ziliak (2017): use a grouping estimator/instrument. (3) Carey et al. (2015): Two instruments based on a time period with no tax changes to estimate dynamics of taxable income. (4) Carroll (1998): proxy for permanent income and calculate synthetic tax rate. (5) Ericson et al. (2015): instrument based on individual/household-specific variables/no measure of previous or future taxable income. (6) Harju and Matikka (2016): use Gruber and Saez (2002) and Weber (2014) but include separate NTR for wage and dividend (plus, separate instruments). (7) Homlund and Söderström (2011): use a dynamic model to explicitly measure short and long run responses. (8) Looney and Singhal (2006): NTR change based on family income stays the same; predict the change in marginal tax rates faced by families assuming that family income remains constant in real terms between year 1 and year 2. (9) Matikka (2018): use changes in flat municipal income tax rates as an instrument for overall changes in marginal tax rates. This instrument is not a function of individual income, which is the basis for an exogenous instrument. (10) Gelber (2014) explicitly control for NTR for wife and husband and extend the most standard specification to allow each spouse's earnings to depend not only on his or her own tax rate and unearned income, but also on the tax rate and unearned income of the other spouse.

DID and IV: Combination of a classical DID and an IV- estimation procedure. The instrument is a binary dummy variable. It determines treatment and control. (e.g. Saez, 2003 or Kopczuk, 2015)

DID classic.

Income Controls: For the majority of coded specifications, there is no information available about what type of income (e.g. gross or taxable) is used.

Auten and Carroll (1999): '**Auten Carroll**' describes the use of log base year income $ln(z_{i,t-k})$ as an income control.

Mostly old studies and robustness checks deliver estimates that use no income control (**none)** at all.

Gruber and Saez (2002): '**Gruber Saez**' defines the inclusion of a spline of base year income as an income control.

Kopczuk (2005): '**Kopczuk**' defines the inclusion of two income control variables. The deviation of log base year income and lagged base year income and lagged base year income separately. To be more precise: $ln(z_{i,t-k-1})$, $ln(z_{i,t-k}) - ln(z_{i,t-k-1})$, spline of $ln(z_{i,t-k-1})$, spline of $ln(z_{i,t-k}) - ln(z_{i,t-k-1})$, combination of $ln(z_{i,t-k-1})$ and $ln(z_{i,t-k}) - ln(z_{i,t-k-1})$, combination of $ln(z_{i,t-k})$ and spline of $ln(z_{i,t-k}) - ln(z_{i,t-k-1})$, combination of spline of $ln(z_{i,t-k})$ and $ln(z_{i,t-k}) - ln(z_{i,t-k-1})$ and combination of spline of $ln(z_{i,t-k})$ and spline of $ln(z_{i,t-k}) - ln(z_{i,t-k-1})$.

The category '**other**' involves all other kinds of income controls. Example: Burns and Ziliak (2017) use a cohort-state-year income control in some specifications.

Difference Length The term difference length defines the time window *k*. If researchers relate 2005 to 2002, the time window will be 3 years.

Weighting by Income: This is a dummy variable that indicates whether (primary) estimation results are weighted by income.

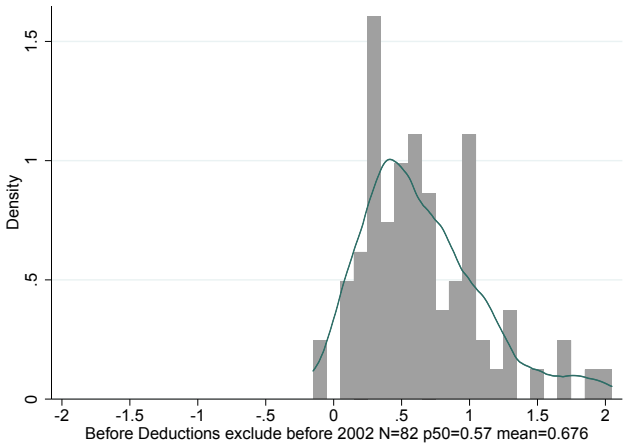
Sample Restrictions:
Age Cutoff: It is a dummy variable that indicates whether an age cutoff is used.
Income Cutoff: I create subcategories: 0-10k, 10-12k, 12-31k and none. Some researchers do not apply any kind of income restrictions. However, sometimes it is not clear if they simply do not mention them, applied no income restriction on purpose or if their dataset considers a subgroup of tax-units in the first place. It often remains unclear what type of income is used (e.g. taxable or gross) to restrict the sample. I coded the values in national currency and recalculated them in US-Dollar. Purchasing power parities do not lead to different results.
Employment type: I distinguish between no restriction with respect to employment type (none), only wage earner, and only self employed individuals.
Marital Status: I distinguish between no restriction with respect to marital status (none), only married tax-units and only singles.

Variations across time and country:
Country Group: USA, Scandinavia (Denmark, Norway, Sweden) and Rest (Canada, Finland, France, Germany, Hungary, Netherlands, New Zealand, Poland, Spain)
Mean year in study data: I calculate the (rounded) mean year of observation based on time start and time end of dataset.
Estimation/Data Decade: I used the mean year of the study data and assigned the respective decade: $<$ 1990, 1990-2000 and $>=$ 2000.

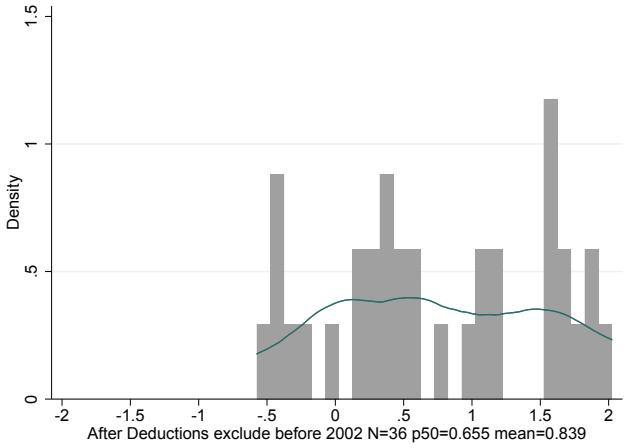
Publication Characteristics:
Publication Decade: 2001-1010, $<$ = 2000 and $>$ 2011.
Published Type: I distinguish between (1) published in a peer reviewed journal and (2) Working Paper.
Extension: Contextual Variables: For a particular estimate, I compare start and end year of (restricted) data period and add the tax related characteristics. Economy related characteristics are merged via the mean year of observation.
Tax Reform Characteristics: It is difficult and almost impossible to code precisely if taxes are increased, and if so, by how much. As an example, think of an estimate that uses data from 2001 to 2010 and exploits three tax changes at different points in the income distribution which differ additionally in magnitude. Therefore, I decided to focus on: (1) introduction of a top tax bracket.
Intro of top tax bracket: information if reform involves an introduction of top. Source: Paper itself plus OECD Tax Database
Economy related characteristics merged via link to mean year of observation (= use start and end year of (restricted) data period for collected primary estimate:
Gini (disposable income, post taxes and transfers)/Income Definition till 2011. To improve (regression) interpretation, I standardized the Gini Coefficient by multiplying it with 100. Remark: These tables are updated on a regular basis. No data is available for China and South Africa. Source: http://stats.oecd.org (07.11.2016/18.06.2019)
Top Income Shares: Pre-tax national income share held by a given percentile group (here top 1% and

top 10%). Pre-tax national income is the sum of all pre-tax personal income flows accruing to the owners of the production factors, labour and capital, before taking into account the operation of the tax/transfer system, but after taking into account the operation of pension system. No data available for Israel and South Africa. Source: World Inequality Database (extracted 16.07.2018/18.06.2019)
Unemployment Rate: The unemployment rate is the number of unemployed people as a percentage of the labour force, with the latter consisting of the unemployed plus those in paid or self-employment. Unemployed people are those who report that they are out of work, that they are available for work and that they have taken active steps to find work in the last four weeks. When unemployment is high, some people become discouraged and stop looking for work; they are then excluded from the labour force. This implies that the unemployment rate may fall, or stop rising, even though there has been no underlying improvement in the labour market. For South Africa and China no data available. (Source: OECD, Short-Term Labour Market Statistics; extracted 17.07.2018/18.06.2019.)
Fraction self-employed: fraction self-employed is defined crudely as all non employees (self-employed, employers, and non classifiable workers) as a fraction of the workforce. For Israel no data available. Source: Kleven - How Can Scandinavians Tax So Much? (2014, Journal of Economic Perspectives)
Modern taxes/GDP: Kleven et al. (2016) decompose the tax take (=tax/GDP) into modern and traditional taxes. Modern taxes include individual and corporate income taxes, payroll taxes and social security contributions, and value added taxes. Traditional taxes include all the other taxes. For Israel no data available. Source: Kleven et al. - Why Can Modern Governments Tax So Much? (2016, Economica)

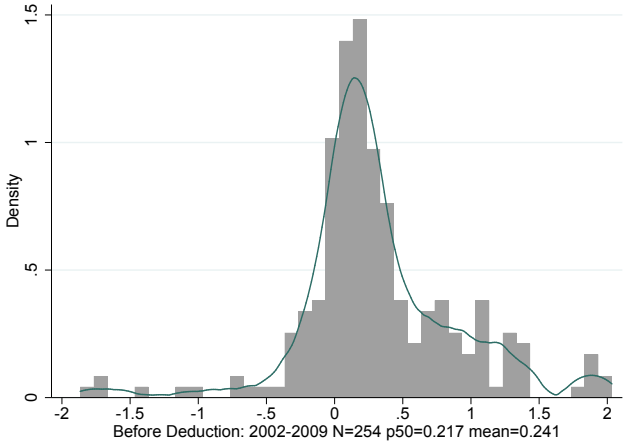
Figure A.2: Distribution of Estimates by Publication Decade.



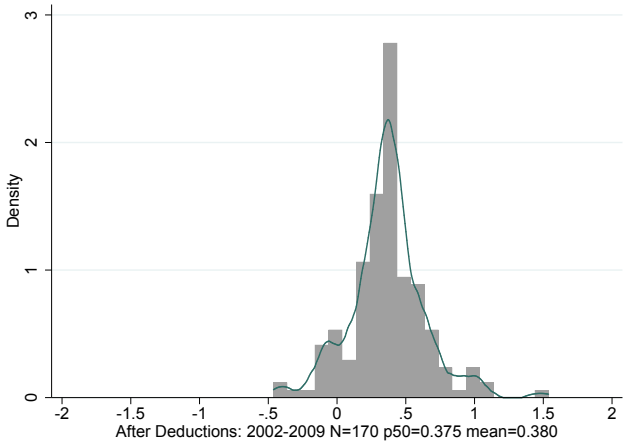
(a) Before Deductions <2002



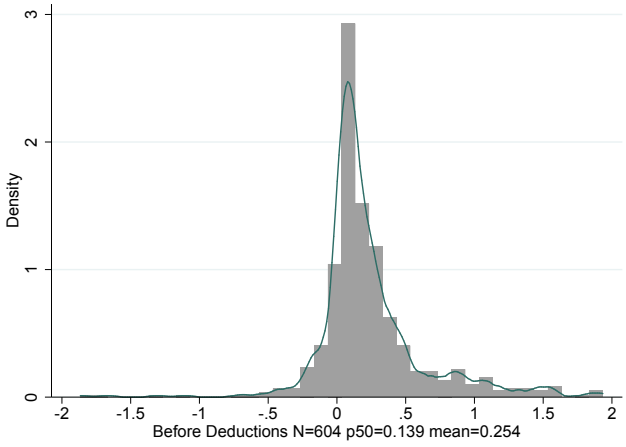
(b) After Deductions <2002



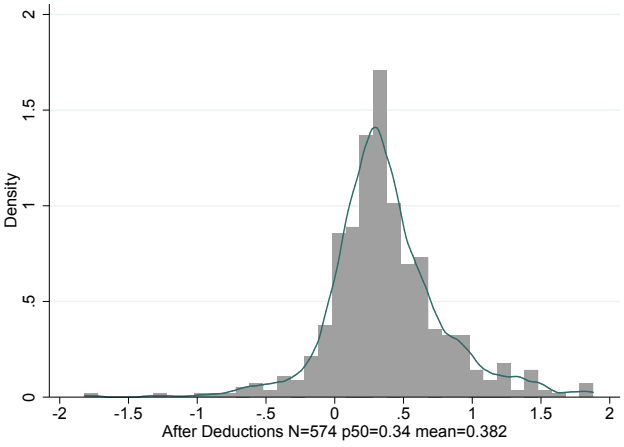
(c) Before Deductions >=2002 and <2009



(d) After Deductions >=2002 and <2009



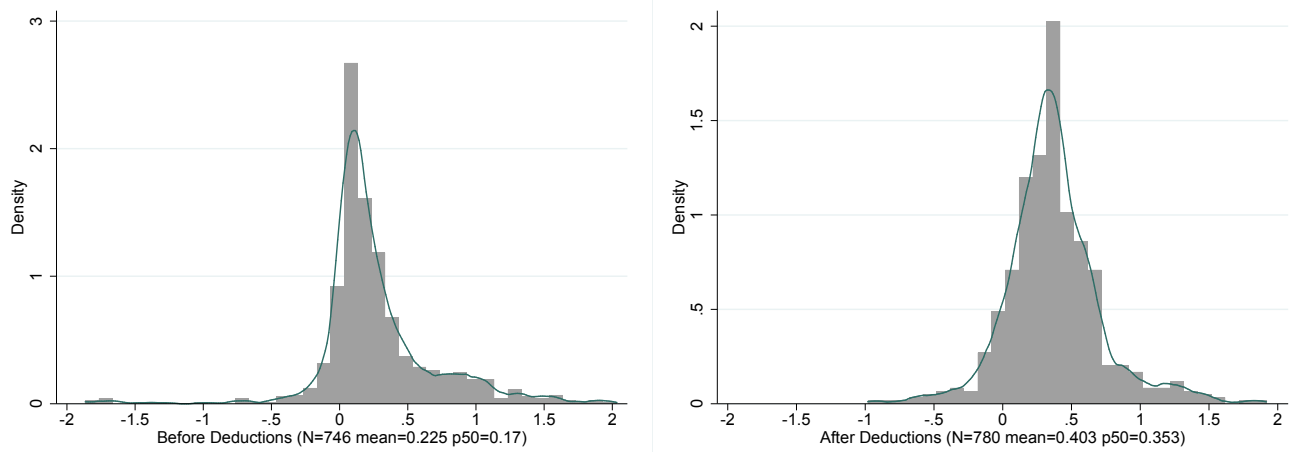
(e) Before Deductions >=2009



(f) After Deductions >=2009

Note: All graphs plot the distribution of elasticities by subsample and publication decade.

Figure A.3: Distribution of Estimates (only income control(s)).



(a) Before Deductions

(b) After Deductions

Note: Both graphs plot the distribution of elasticities that are derived with a specification using income control(s).

Table A.5: Descriptive Statistics: Categories of Heterogeneity

	Before Deductions (BD) (N=940) # studies	**After Deductions (AD) (N=780)** # studies
Estimation Techniques		
Regression technique		
IV: mechanical tax rate changes	32	32
IV: (lagged) mechanical tax rate changes	9	12
IV: other	6	11
DID and IV	7	3
classic DID	1	4
Income Control		
Auten Carroll (1999)	23	23
none	28	28
Gruber Saez (2002) spline	18	14
Kopczuk (2005) type	19	21
other	7	4
Difference Length		
3 years	20	24
1 year	25	24
2 years	13	14
4+ years	9	8
Weighted by Income	16	15
Sample Restrictions		
Age Cutoff	23	27
Income Cutoff		
0-10k	15	17
none	11	11
10k-12k	17	11
12-31k	19	15
$>$ 31k	23	21
Variations across Countries and Time		
Country Group		
USA	20	19
Scandinavia	5	67
other countries	16	20
Estimation decade		
$<$ 1999	15	17
1990 - 2000	15	10
$>$ 2000	16	23
Publication Characteristics		
Publication decade		
2001-2010	11	15
$<=$ 2000	3	3
$>$ 2011	27	28
Published Type		
published in peer reviewed journal	26	27
working paper	15	19

Note: see text for description of sample. I present descriptive results separately for two subsamples: before (BD) and after deductions (AD). The sample covers only observations with a given standard error or t-statistic. Reference categories are given in italics.

A.4 Additional Sample Restrictions - Before (BD) and After Deductions

Researchers often conduct subgroup analysis by *marital status* or *employment type*. Single taxpayers might respond differently than married couples and it is obvious that a self-employed person has more control over his or her income compared to someone receiving only wage income.

Table A.6: Descriptive Statistics: Sample Restrictions

	Before Deductions (BD) (N=940)		**After Deductions (AD) (N=780)**	
	Mean	**Std. Dev.**	**Mean**	**Std. Dev.**
Sample Restrictions				
Employment type				
none	0.710	0.454	0.877	0.329
wage earner	0.218	0.413	0.040	0.195
self-employed	0.072	0.259	0.083	0.277
Marital Status				
none	0.845	0.362	0.858	0.350
married	0.110	0.313	0.092	0.290
single	0.046	0.209	0.050	0.218

Note: see text for description of sample. I present descriptive results separately for two subsamples: before (BD) and after deductions (AD). The sample covers only observations with a given standard error or t-statistic.

In line with expectations, a BD elasticity estimated on a subsample of only wage earners leads to a lower elasticity compared to a specification with no restriction on employment type. Greater coverage of third party information reporting and the associated lower evasion opportunities might be a reason (Kleven et al., 2011). If primary studies restrict their sample according to marital status, it appears that single taxpayers reveal a lower BD elasticity compared to no restriction.

Table A.7: WLS before deductions results with add. sample restrictions

Dependent Variable: Income Elasticity BEFORE deductions	(1)	(2)	(3)	(4)	(5)	(6)
Estimation Technique:						
Reg. Technique (omitted: IV: mechanical tax rate changes)						
IV: (lagged) mechanical tax rate changes	0.060*	0.054*	0.061*	0.055*	0.028**	0.025*
	(0.031)	(0.029)	(0.032)	(0.030)	(0.013)	(0.015)
IV-other	0.075	0.081*	0.070	0.078*	0.074	0.107*
	(0.056)	(0.044)	(0.055)	(0.042)	(0.053)	(0.056)
DID-IV	0.298***	0.224**	0.291***	0.218*	0.319***	0.313***
	(0.053)	(0.105)	(0.058)	(0.109)	(0.046)	(0.075)
DID-classic	0.332***	0.068	0.309***	0.049	0.184***	0.149**
	(0.059)	(0.132)	(0.078)	(0.137)	(0.060)	(0.065)
Income Control (omitted: Auten Carroll)						
none	-0.213***	-0.212***	-0.213***	-0.211***	-0.207***	-0.207***
	(0.024)	(0.025)	(0.024)	(0.025)	(0.029)	(0.029)
Gruber Saez Spline	-0.020***	-0.021***	-0.021***	-0.022***	-0.014**	-0.016***
	(0.005)	(0.007)	(0.006)	(0.007)	(0.006)	(0.005)
Kopczuk	-0.017**	-0.014**	-0.018**	-0.015**	-0.012*	-0.010**
	(0.007)	(0.005)	(0.008)	(0.006)	(0.006)	(0.005)
other	-0.034**	-0.029**	-0.033*	-0.029*	-0.012	-0.009
	(0.017)	(0.013)	(0.018)	(0.015)	(0.010)	(0.009)
Difference Length (omitted: 3-years)						
1 year	0.060	0.033	0.058	0.032	0.031	0.012
	(0.063)	(0.045)	(0.062)	(0.044)	(0.051)	(0.040)
2 years	-0.013	-0.030*	-0.015	-0.033*	-0.042***	-0.035***
	(0.021)	(0.016)	(0.021)	(0.019)	(0.010)	(0.008)
4 years and more	0.082*	0.068**	0.084*	0.068**	0.012	0.026
	(0.042)	(0.030)	(0.043)	(0.029)	(0.020)	(0.021)
Age Cutoff applied (omitted: no restriction)						
Age Cutoff applied		-0.282**		-0.278**		-0.267
		(0.122)		(0.123)		(0.174)
Income Cutoff applied (omitted: 0-10k)						
none		0.018		0.019		-0.020*
		(0.021)		(0.022)		(0.010)
10k-12k		0.024		0.026		-0.015**
		(0.016)		(0.016)		(0.007)
12k-31k		0.009		0.009		0.007
		(0.007)		(0.010)		(0.008)
>31k		0.021		0.023		-0.005
		(0.017)		(0.019)		(0.012)
Employment Type (omitted: no restriction)						
noindent wage earner			-0.008*	-0.005		
			(0.005)	(0.005)		
self-employed			0.006	0.007		
			(0.009)	(0.011)		
Marital Status (omitted: no restriction)						
married			0.021	0.022		
			(0.031)	(0.036)		
sinlge			0.012	0.009		
			(0.030)	(0.028)		
Country Group (omitted: USA)						
Scandinavia				0.074	0.239*	
					(0.081)	(0.123)
other countries					0.191**	0.343***
					(0.081)	(0.126)
(Publication) Decade (omitted: 2001-2010)						
prior to 2001					0.226	0.426**
					(0.141)	(0.207)
after 2010					-0.254**	-0.205***
					(0.098)	(0.073)
Constant	0.073***	0.351***	0.079***	0.350***	0.244***	0.296***
	(0.007)	(0.123)	(0.009)	(0.123)	(0.043)	(0.054)
Observations	940	940	940	940	940	940
Adjusted R^2	0.566	0.615	0.566	0.615	0.628	0.655

Note: Columns (1) to (6) estimated using WLS with the inverse of an estimate's variance as analytical weights. Reported coefficients need to be interpret as a deviation from the reference category (in bold). Standard errors (in parentheses) are clustered at the study level. Significance levels are * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.8: WLS after deductions results with add. sample restrictions

Dependent Variable: Income Elasticity AFTER deductions	(1)	(2)	(3)	(4)	(5)	(6)
Estimation Technique:						
Reg. Technique (omitted: IV: mechanical tax rate changes)						
IV: (lagged) mechanical tax rate changes	0.409***	0.420***	0.461***	0.455***	0.232***	0.207***
	(0.088)	(0.061)	(0.047)	(0.054)	(0.048)	(0.074)
IV-other	-0.265*	-0.246**	-0.224***	-0.227*	0.403*	0.197
	(0.145)	(0.118)	(0.078)	(0.130)	(0.230)	(0.218)
DID-IV	-0.590**	-0.702**	-0.423***	-0.583**	-0.152	-0.289
	(0.224)	(0.281)	(0.155)	(0.240)	(0.403)	(0.475)
DID-classic	-0.188	-0.189	-0.152	-0.162	-0.167	-0.178
	(0.372)	(0.363)	(0.320)	(0.324)	(0.323)	(0.305)
Income Control (omitted: Auten Carroll)						
none	0.108	0.045	0.100	0.043	-0.225	-0.249
	(0.078)	(0.089)	(0.069)	(0.096)	(0.176)	(0.159)
Gruber Saez Spline	-0.100	-0.137**	-0.086	-0.120*	-0.087	-0.119
	(0.068)	(0.068)	(0.054)	(0.067)	(0.066)	(0.088)
Kopczuk-type	-0.371***	-0.387***	-0.375***	-0.383***	0.027	0.025
	(0.043)	(0.075)	(0.047)	(0.087)	(0.068)	(0.104)
other	-0.195**	-0.331**	-0.174**	-0.308**	0.108	0.048
	(0.075)	(0.132)	(0.085)	(0.136)	(0.074)	(0.124)
Difference Length (omitted: 3-years)						
1 year	-0.048	0.073	-0.049	0.066	-0.001	0.119
	(0.106)	(0.074)	(0.121)	(0.085)	(0.127)	(0.090)
2 years	0.033	0.019	0.021	0.008	0.043	0.057
	(0.086)	(0.119)	(0.091)	(0.117)	(0.102)	(0.105)
4 years and more	0.285	0.182	0.290	0.188	-0.329	-0.362
	(0.191)	(0.212)	(0.189)	(0.210)	(0.247)	(0.242)
Age Cutoff applied (omitted: no restriction)						
Age Cutoff applied		0.252**		0.245**		0.140
		(0.113)		(0.113)		(0.124)
Income Cutoff applied (omitted: 0-10k)						
none		0.154***		0.147**		0.254***
		(0.054)		(0.057)		(0.087)
10k-12k		0.109		0.099		0.353
		(0.090)		(0.088)		(0.236)
12k-31k		0.111*		0.105		0.068
		(0.063)		(0.063)		(0.059)
>31k		0.468		0.462		0.518
		(0.424)		(0.423)		(0.353)
Employment Type (omitted: no restriction)						
wage earner			-0.007	-0.019		
			(0.050)	(0.031)		
self-employed			-0.274***	-0.208*		
			(0.052)	(0.105)		
Marital Status (omitted: no restriction)						
married			-0.074	-0.035		
			(0.096)	(0.071)		
single			0.010	0.012		
			(0.098)	(0.090)		
Country Group (omitted: USA)						
Scandinavia					0.121	0.410
					(0.112)	(0.305)
other countries					0.416***	0.632**
					(0.136)	(0.304)
(Publication) Decade (omitted: 2001-2010)						
prior to 2001					1.060*	1.164*
					(0.599)	(0.662)
after 2010					-0.468***	-0.500***
					(0.161)	(0.173)
Constant	0.445***	0.208***	0.455***	0.222***	0.376***	-0.019
	(0.040)	(0.066)	(0.041)	(0.068)	(0.098)	(0.272)
Observations	780	780	780	780	780	780
Adjusted R^2	0.405	0.479	0.414	0.483	0.553	0.621

Note: Columns (1) to (6) estimated using WLS with the inverse of an estimate's variance as analytical weights. Reported coefficients need to be interpret as a deviation from the reference category (in bold). Standard errors (in parentheses) are clustered at the study level. Significance levels are * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.5 Contextual Factors - Full Results

A.5.1 Contextual Factors - Before Deductions (BD) - Full Results

Table A.9: WLS before deductions - Contextual Variables

Dependent Variable: Income Elasticity BEFORE deductions	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Reg. Technique (omitted: IV: Δ mechanical tax rate)							
IV: lagged Δ mechanical tax rate	0.063	0.040***	0.048**	0.058**	0.066*	0.044**	0.038**
	(0.039)	(0.014)	(0.019)	(0.028)	(0.036)	(0.019)	(0.017)
IV-other	0.095**	0.090	0.063	0.073	0.181*	0.050	0.077
	(0.037)	(0.057)	(0.059)	(0.058)	(0.106)	(0.057)	(0.061)
DID-IV	0.297***	0.272***	0.271***	0.291***	0.318***	0.296***	0.253***
	(0.055)	(0.068)	(0.077)	(0.060)	(0.090)	(0.050)	(0.083)
DID-classic	0.325***	0.270***	0.267***	0.319***	0.337***	0.359***	0.134*
	(0.073)	(0.066)	(0.077)	(0.066)	(0.076)	(0.048)	(0.075)
Income Control (omitted: Auten Carroll)							
none	-0.213***	-0.213***	-0.214***	-0.214***	-0.206***	-0.211***	-0.211***
	(0.024)	(0.023)	(0.023)	(0.023)	(0.023)	(0.026)	(0.025)
Gruber Saez Spline	-0.020***	-0.011	-0.017***	-0.019***	-0.018**	-0.022***	-0.013*
	(0.005)	(0.009)	(0.005)	(0.005)	(0.008)	(0.006)	(0.008)
Kopczuk-type	-0.016**	-0.013*	-0.016**	-0.016**	-0.018**	-0.021**	-0.011*
	(0.006)	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)	(0.006)
other	-0.035*	-0.065***	-0.053**	-0.043*	-0.044***	0.002	-0.047***
	(0.017)	(0.020)	(0.022)	(0.024)	(0.014)	(0.010)	(0.013)
Difference Length (omitted: 3-years)							
1 year	0.066	0.057	0.063	0.064	0.058	0.035	0.049
	(0.079)	(0.060)	(0.066)	(0.066)	(0.079)	(0.049)	(0.056)
2 years	-0.013	-0.005	-0.004	-0.008	-0.012	-0.040*	-0.014***
	(0.021)	(0.007)	(0.020)	(0.020)	(0.016)	(0.024)	(0.004)
4 years and more	0.083*	0.047**	0.083*	0.086*	0.068*	0.063**	0.038*
	(0.043)	(0.020)	(0.043)	(0.044)	(0.035)	(0.030)	(0.021)
Additional Variables							
Intro top bracket	-0.027						
	(0.078)						
Gini Coefficient		0.008***					
		(0.002)					
Top 10%			0.814*				
			(0.442)				
Top 1%				0.330			
				(0.448)			
Unemployment Rate					-0.007		
					(0.004)		
Fraction of self-employed						0.016***	
						(0.006)	
modern taxes (in 2005)							-0.010***
							(0.002)
Constant	0.073***	-0.118***	-0.142	0.047	0.113***	-0.081	0.460***
	(0.006)	(0.043)	(0.114)	(0.032)	(0.022)	(0.054)	(0.084)
Observations	940	931	912	912	854	915	921
Adjusted R^2	0.566	0.614	0.585	0.576	0.569	0.611	0.614

Columns (1) to (7) estimated using WLS. Standard errors (in parentheses) are clustered at the study level. Significance levels are * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

For observations that are based on classic DID approach, I do not have information of the share of self employed that correspond to the respective mean year of observation.

A.5.2 Contextual Factors - After Deductions (AD) - Full Results

Table A.10: WLS after deductions - Contextual Factors

Dependent Variable: Income Elasticity AFTER deductions	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Reg. Technique (omitted: IV: Δ mechanical tax rate)							
IV: (lagged) Δ mechanical tax rate	0.410***	0.422***	0.287***	0.304***	0.393***	0.472***	0.510***
	(0.088)	(0.091)	(0.094)	(0.108)	(0.058)	(0.116)	(0.111)
IV-other	-0.265*	-0.279	-0.087	-0.016	-0.038	-0.300*	-0.391**
	(0.145)	(0.179)	(0.147)	(0.161)	(0.135)	(0.158)	(0.175)
DID-IV	-0.591**	-0.596**	-0.618***	-0.699***	-0.042	-0.653***	-0.498*
	(0.223)	(0.225)	(0.084)	(0.109)	(0.178)	(0.222)	(0.296)
DID-classic	-0.189	-0.201	-0.011	-0.009	-1.130**	-0.264	-0.305
	(0.372)	(0.398)	(0.363)	(0.340)	(0.482)	(0.377)	(0.376)
Income Control (omitted: Auten Carroll)							
none	0.107	0.107	0.029	-0.013	0.029	0.021	0.045
	(0.078)	(0.077)	(0.083)	(0.097)	(0.108)	(0.116)	(0.096)
Gruber Saez Spline	-0.100	-0.080	-0.112*	-0.107*	-0.043	-0.102	-0.084
	(0.068)	(0.064)	(0.064)	(0.060)	(0.049)	(0.085)	(0.067)
Kopczuk-type	-0.371***	-0.385***	-0.290***	-0.352***	-0.333***	-0.454***	-0.493***
	(0.043)	(0.111)	(0.061)	(0.053)	(0.067)	(0.122)	(0.123)
other	-0.190*	-0.240	-0.087	-0.147	-0.304*	-0.374*	-0.368*
	(0.096)	(0.151)	(0.099)	(0.122)	(0.158)	(0.189)	(0.184)
Difference Length (omitted: 3-years)							
1 year	-0.048	-0.042	-0.074	-0.094	-0.066	0.013	-0.009
	(0.106)	(0.114)	(0.105)	(0.100)	(0.122)	(0.100)	(0.099)
2 years	0.035	0.061	-0.060	-0.063	-0.017	0.088	0.100
	(0.084)	(0.133)	(0.081)	(0.090)	(0.095)	(0.121)	(0.126)
4 years and more	0.288	0.267	0.041	-0.042	0.430	0.209	0.482*
	(0.188)	(0.211)	(0.245)	(0.266)	(0.436)	(0.200)	(0.253)
Additional Variables							
Intro top bracket	-0.016						
	(0.132)						
Gini Coefficient		-0.002					
		(0.014)					
Top 10%			3.563**				
			(1.536)				
Top 1%				7.709**			
				(3.202)			
Unemployment Rate					0.067*		
					(0.039)		
Fraction of self-employed						-0.022	
						(0.023)	
modern taxes (in 2005)							0.016
							(0.012)
Constant	0.450***	0.513	-0.572	-0.159	-0.088	0.746**	-0.060
	(0.116)	(0.424)	(0.435)	(0.243)	(0.315)	(0.349)	(0.363)
Observations	780	767	771	771	703	771	780
Adjusted R^2	0.404	0.410	0.455	0.469	0.468	0.425	0.426

Columns (1) to (7) estimated using WLS. Standard errors (in parentheses) are clustered at the study level. Significance levels are * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

For observations that are based on classic DID approach, I do not have information of the share of self employed that correspond to the respective mean year of observation.

A.6 Sensitivity Analysis and Robustness Checks

A.6.1 Sensitivity Analysis

In this section, I limit the number of estimates along various dimensions: (i) I drop studies that are released prior to 2002, (ii) I consider only published articles or (iii) only US studies and (iv) I only consider taxable income elasticities. Results are presented in Table 17 and they vary slightly compared to the baseline results when I consider only published articles and only US studies. For US studies, the constant for BD elasticities is larger and smaller for AD elasticities compared to the baseline results shown in Table 2 and 3 (column 2).[1] Moreover, the degree of influence of other factors changes. The use of (lagged) mechanical tax rate changes lead to an increase of 0.541 compared to an approach that relies on mechanical tax rate changes as an instrument. On the other hand DID and DID IV does not make a big difference compared to an approach using the standard mechanical tax rate changes instrument. The coefficient of DID-classic is very large but mainly driven by older studies (reported < 2002).

Table A.11: Sensitivity Analysis: Different Sample Restrictions

Dependent Variable:	drop studies prior to 2002		(only) Published		(only) US studies		(only) Taxable
Income Elasticity ...	BD	AD	BD	AD	BD	AD	Income
Reg. Technique (omitted: IV:Δ mech. tax rate))							
IV: (lagged) Δ mech. tax rate	0.060*	0.410***	0.060*	0.420***	0.395**	0.271**	0.409***
	(0.031)	(0.088)	(0.031)	(0.086)	(0.155)	(0.123)	(0.088)
IV-other	0.055	-0.261*	0.055	0.690***	-0.003	0.309***	-0.274*
	(0.054)	(0.147)	(0.053)	(0.117)	(0.094)	(0.093)	(0.142)
DID-IV	0.295***	-0.248**	0.290***	0.239	-0.026	0.115*	-0.751***
	(0.055)	(0.113)	(0.055)	(0.239)	(0.120)	(0.064)	(0.220)
DID-classic	0.332***	-0.225	0.337***	0.076	-0.054	1.302***	-1.432***
	(0.059)	(0.395)	(0.058)	(0.173)	(0.116)	(0.128)	(0.474)
Income Control (omitted: Auten Carroll)							
none	-0.212***	0.103	-0.215***	-0.873***	-0.083	0.012	0.124
	(0.025)	(0.079)	(0.024)	(0.127)	(0.140)	(0.171)	(0.080)
Gruber Saez Spline	-0.019***	-0.105	-0.020***	-0.070	-0.150*	0.020	-0.089
	(0.005)	(0.069)	(0.005)	(0.054)	(0.074)	(0.073)	(0.065)
Kopczuk-type	-0.015**	-0.377***	-0.017**	-0.300***	-0.240**	-0.062	-0.360***
	(0.006)	(0.045)	(0.007)	(0.043)	(0.087)	(0.067)	(0.050)
other	-0.033*	-0.195**	-0.033*	-0.254***	-0.134	0.017	-0.186**
	(0.017)	(0.076)	(0.017)	(0.042)	(0.082)	(0.123)	(0.075)
Difference Length (omitted: 3-years)							
1 year	0.060	-0.052	0.056	0.024	0.078	-0.155**	-0.046
	(0.064)	(0.109)	(0.063)	(0.062)	(0.102)	(0.073)	(0.107)
2 years	-0.013	0.025	-0.015	0.081	-0.137	-0.079	0.032
	(0.021)	(0.086)	(0.021)	(0.104)	(0.161)	(0.061)	(0.087)
4 years and more	0.081*	0.125	0.081*	0.089	0.147*	0.023	0.431
	(0.043)	(0.175)	(0.043)	(0.335)	(0.083)	(0.137)	(0.285)
Constant	0.072***	0.451***	0.073***	0.369***	0.310***	0.258**	0.434***
	(0.006)	(0.042)	(0.007)	(0.045)	(0.066)	(0.093)	(0.048)
Observations	858	744	822	701	464	415	737
Adjusted R^2	0.571	0.407	0.592	0.623	0.063	0.363	0.434

Note: BD refers to the before deductions subsample and AD to the after deductions subsample. All results are based on Weighted Least Squares (WLS) with the inverse of an estimate's variance as analytical weights. The baseline specification involves only controls for estimation technique (regression technique, income control and difference length). Standard errors (in parentheses) are clustered at the study level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

[1] The results within the US subsample but also the baseline results remain remarkably robust even when I exclude all estimates extracted from Weber(2014).

A.6.2 Robustness Checks: Different Estimation Techniques

The upper (lower) part of the Table displays results based on the BD (AD) subsample. Column (1) display the baseline results obtained in column (2) of Tables 2 and 3. In Column (2), I present results based on a random effects meta-regression technique. The weights in the baseline WLS represent only the within study variance and neglect any possible between study variance. In contrast the estimation used here, it is equivalent to the baseline WLS with an additive between study component in the denominator of the weights. Stanley (2017) show that WLS is superior to conventional random-effects meta-regression estimation. In case of publication bias, in particular, WLS always reveals a smaller bias than the random effects model. Moreover, random effects estimates are highly sensitive to the accuracy of the estimate of the between study variance.

For illustration, results based on a simple OLS are presented in column (4). Since we observe large heteroscedasticity among estimates, an OLS procedure is never appropriate in a meta analysis. To increase efficiency, a WLS procedure is always preferable.

Column (5) shows results that are based on WLS with weights that are based on the inverse of the share of observations per study in relation to the full sample. Given that my collected sample does not consist only of one estimate per study but of all available estimates a particular study provides, there's a risk that the baseline results are driven only by a small number of studies that offer a lot of estimates.

It seems reasonable to assume that extracted estimates themselves are influenced by their sample size. For instance, a dataset that almost covers the entire population might produce a different estimate and standard error compared to a dataset of a few hundred observations. In column (6) I weight each primary estimate with the sample size of the respective study. The difference between those results compared to a standard WLS with precision as a weight should be small, since the sampling error is to large extent determined by the respective sample size.

The BD subsample is based on 38 studies and the AD subsample on 37 studies. To check whether clustering in the meta-analysis produces misleading inferences, I apply a wild-cluster bootstrap procedure proposed by Cameron et al. (2008) for improved inference with only few cluster (see Column (3)).

The sample size in column (6) is lower because the sample size is not observed for every primary estimate.

Table A.12: Robustness Checks: Different Estimation Techniques

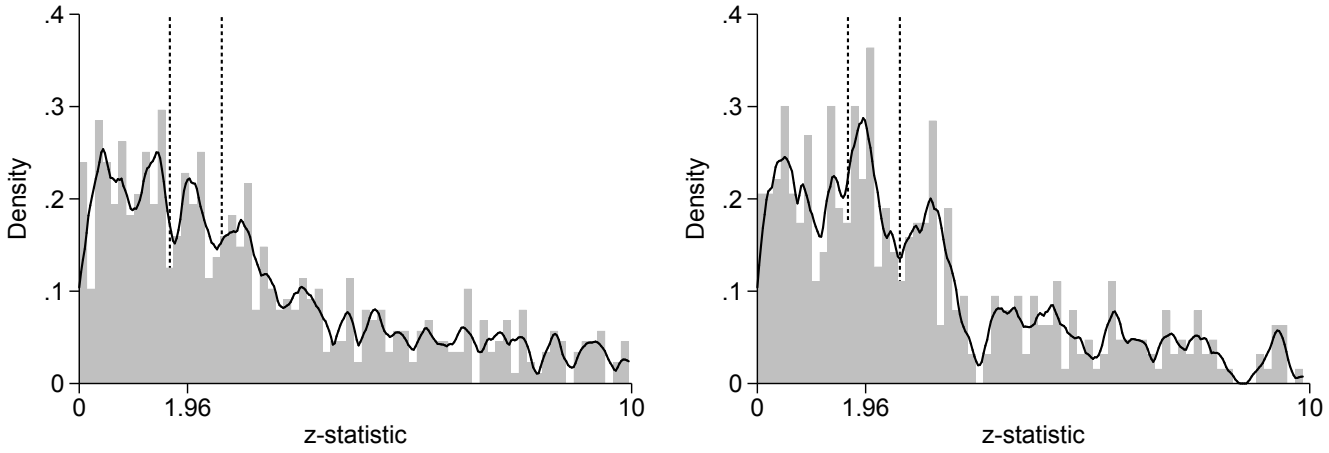
Dependent Variable:	(1)	(2)	(3)	(4)	(5)	(6)
Income Elasticity BEFORE deductions	WLS	META	WILD	OLS	EQUAL	NOBS
Reg. Technique (omitted: IV: mechanical tax rate changes)						
IV: (lagged) mechanical tax rate changes	0.060*	0.104*	0.060	0.254	0.400	0.124***
	(0.031)	(0.059)	(0.061)	(0.264)	(0.335)	(0.038)
IV-other	0.075	-0.096*	0.075	-0.228*	-0.428***	-0.016
	(0.056)	(0.057)	(0.065)	(0.135)	(0.154)	(0.093)
DID-IV	0.298***	0.080*	0.298***	-0.289	-0.230	0.475***
	(0.053)	(0.048)	(0.000)	(0.247)	(0.166)	(0.107)
DID-classic	0.332***	-0.065	0.332***	-0.583	-0.501***	0.173*
	(0.059)	(0.300)	(0.000)	(0.385)	(0.182)	(0.101)
Income Control (omitted: Auten Carroll)						
none	-0.213***	-0.156***	-0.213***	0.276	-0.044	-0.183***
	(0.024)	(0.036)	(0.069)	(0.322)	(0.170)	(0.062)
Gruber Saez Spline	-0.020***	-0.152***	-0.020***	-0.325**	-0.190	-0.040*
	(0.005)	(0.034)	(0.007)	(0.127)	(0.213)	(0.024)
Kopczuk-type	-0.017**	-0.195***	-0.017***	-0.243*	-0.371**	-0.015
	(0.007)	(0.031)	(0.005)	(0.125)	(0.164)	(0.013)
other	-0.034*	-0.248***	-0.034	-0.266**	-0.413***	-0.114***
	(0.017)	(0.040)	(0.031)	(0.109)	(0.118)	(0.037)
Difference Length (omitted: 3-years)						
1 year	0.060	0.179***	0.060	0.158	0.281**	0.174
	(0.063)	(0.029)	(0.089)	(0.140)	(0.138)	(0.104)
2 years	-0.013	-0.059	-0.013	-0.121	-0.141	0.047
	(0.021)	(0.038)	(0.022)	0.113)	(0.156)	(0.032)
4 years and more	0.082*	0.014	0.082	-0.016	0.047	0.117***
	(0.042)	(0.033)	(0.125)	(0.138)	(0.136)	(0.035)
Constant	0.073***	0.292***	0.073***	0.404***	0.519***	0.078***
	(0.007)	(0.026)	(0.000)	0.128)	(0.136)	(0.017)
Observations	940	940	940	940	940	869
Adjusted R^2	0.566		0.566	0.020	0.065	0.114
Income Elasticity AFTER deductions	(1)	(2)	(3)	(4)	(5)	(6)
Reg. Technique (omitted: IV: mechanical tax rate changes)						
IV: (lagged) mechanical tax rate changes	0.409***	0.294***	0.409***	0.326***	0.320**	0.445***
	(0.088)	(0.050)	(0.000)	(0.095)	(0.123)	(0.052)
IV-other	-0.265*	0.083	-0.265	0.181	0.411*	-0.108
	(0.145)	(0.052)	(0.293)	(0.123)	(0.226)	(0.127)
DID-IV	-0.590**	-0.129	-0.590	-0.104	-0.153	-0.146
	(0.224)	(0.081)	(0.530)	(0.125)	(0.160)	(0.093)
DID-classic	-0.188	0.578***	-0.188	0.551	0.814**	-0.144
	(0.372)	(0.071)	(0.278)	(0.331)	(0.401)	(0.296)
Income Control (omitted: Auten Carroll)						
none	0.108	-0.014	0.108*	-0.021	-0.276	0.030
	(0.078)	(0.044)	(0.059)	(0.130)	(0.206)	(0.065)
Gruber Saez Spline	-0.100	-0.000	-0.100	-0.056	-0.227	-0.126
	(0.068)	(0.045)	(0.080)	(0.068)	(0.169)	(0.100)
Kopczuk-type	-0.371***	-0.083**	-0.371***	-0.072	-0.193	-0.349***
	(0.043)	(0.041)	(0.120)	(0.088)	(0.177)	(0.094)
other	-0.195**	0.134	-0.195***	0.297	0.370	0.207
	(0.075)	(0.117)	(0.067)	(0.544)	(0.553)	(0.358)
Difference Length (omitted: 3-years)						
1 year	-0.048	0.018	-0.048	-0.044	0.088	-0.031
	(0.106)	(0.035)	(0.102)	(0.117)	(0.151)	(0.044)
2 years	0.033	0.091*	0.033	0.088	0.167*	-0.109***
	(0.086)	(0.053)	(0.074)	(0.105)	(0.087)	(0.039)
4 years and more	0.285	0.149**	0.285	0.264	0.651**	1.373**
	(0.191)	(0.066)	(0.229)	(0.221)	(0.251)	(0.644)
Constant	0.445***	0.295***	0.445***	0.317***	0.370***	0.484***
	(0.040)	(0.032)	(0.000)	(0.066)	(0.111)	(0.071)
Observations	780	780	780	780	780	728
Adjusted R^2	0.405		0.405	0.111	0.268	0.335

Except for column 3 standard errors (in parentheses) are clustered at the study level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.7 Selective Reporting Bias: more information

A.7.1 Distribution of z-statistics - only with income controls

Figure A.4: Distribution of z-statistics - only with income controls.



(a) Before Deductions

(b) After Deductions

Note: The left (right) figure is based on the before (after) deductions subsample. The 5% significance value (=1.96) is highlighted.

A.7.2 Selective Reporting Bias: BD - Full Results

Table A.13: WLS before deductions: Publication Bias Full Results

Dependent Variable: Income Elasticity BEFORE deductions	(1)	(2)	(3)	(4)
Reg. Technique (omitted: IV: Δ mechanical tax rate)				
IV: lagged Δ mechanical tax rate	0.031*	0.029*	0.022	0.025
	(0.018)	(0.016)	(0.014)	(0.016)
IV-other	-0.165*	-0.164*	-0.235**	-0.196*
	(0.096)	(0.087)	(0.113)	(0.106)
DID-IV	0.198*	0.216**	0.197*	0.205**
	(0.101)	(0.095)	(0.103)	(0.098)
DID-classic	-1.052***	-0.797**	-0.199	-0.135
	(0.293)	(0.300)	(0.269)	(0.344)
Income Control (omitted: Auten Carroll)				
none	-0.211***	-0.210***	-0.211***	-0.210***
	(0.026)	(0.026)	(0.026)	(0.026)
Gruber Saez Spline	-0.020***	-0.018***	-0.020***	-0.019***
	(0.005)	(0.005)	(0.006)	(0.005)
Kopczuk-type	-0.018***	-0.016***	-0.019***	-0.017***
	(0.007)	(0.006)	(0.007)	(0.006)
other	-0.026*	-0.022**	-0.022*	-0.023*
	(0.013)	(0.010)	(0.012)	(0.013)
Difference Length (omitted: 3-years)				
1 year	0.034	0.030	0.024	0.029
	(0.052)	(0.051)	(0.046)	(0.049)
2 years	-0.026	0.005	-0.033**	-0.028*
	(0.016)	(0.011)	(0.014)	(0.016)
4 years and more	0.053*	0.046*	0.041	0.050
	(0.031)	(0.027)	(0.026)	(0.030)
Standard Error	3.654***	4.084***	0.972	0.652
	(0.719)	(0.845)	(0.812)	(0.988)
Journal impact factor		-0.012		
		(0.008)		
Std.Error* Impact Factor		-0.051		
		(0.035)		
Dummy if obs > median(obs)			0.771***	
			(0.279)	
Std.Error*D if obs > median(obs)			4.375***	
			(1.416)	
Dummy reported prior to 2009				0.575**
				(0.267)
Std.Error*D reported prior to 2009				3.726***
				(1.322)
Constant	0.876***	0.982***	0.477***	0.460**
	(0.159)	(0.186)	(0.138)	(0.181)
Observations	940	940	940	940
Adjusted R^2	0.614	0.624	0.628	0.627

Columns (1) to (4) estimated using WLS. Standard errors (in parentheses) are clustered at the study level. Significance levels are * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Included standard errors as explanatory variables are normalized. It allows an interpretation as standard deviation.

A.7.3 Selective Reporting Bias: AD - Full Results

Table A.14: WLS after deductions Publication Bias Full Results

Dependent Variable: Income Elasticity AFTER deductions	(1)	(2)	(3)	(4)
Reg. Technique (omitted: IV: Δ mechanical tax rate)				
IV: lagged Δ mechanical tax rate	0.413***	0.205*	0.423***	0.426***
	(0.088)	(0.110)	(0.088)	(0.088)
IV-other	-0.264*	-0.066	-0.269*	-0.271*
	(0.143)	(0.167)	(0.140)	(0.138)
DID-IV	-0.577**	-0.390	-0.626**	-0.633**
	(0.230)	(0.246)	(0.258)	(0.299)
DID-classic	-0.186	-0.044	-0.266	-0.351
	(0.375)	(0.373)	(0.421)	(0.444)
Income Control (omitted: Auten Carroll)				
none	0.107	-0.020	0.125	0.134
	(0.075)	(0.097)	(0.086)	(0.084)
Gruber Saez Spline	-0.099	-0.139*	-0.069	-0.060
	(0.068)	(0.078)	(0.062)	(0.062)
Kopczuk-type	-0.372***	-0.052	-0.343***	-0.328***
	(0.042)	(0.092)	(0.055)	(0.054)
other	-0.193**	0.289	-0.168**	-0.160*
	(0.076)	(0.190)	(0.082)	(0.082)
Difference Length (omitted: 3-years)				
1 year	-0.048	-0.080	-0.030	-0.018
	(0.106)	(0.129)	(0.095)	(0.089)
2 years	0.034	0.031	0.046	0.061
	(0.089)	(0.114)	(0.090)	(0.093)
4 years and more	0.300	0.271	0.290*	0.354*
	(0.201)	(0.180)	(0.173)	(0.201)
Standard Error	-0.030	-0.834***	-0.223	-0.360
	(0.203)	(0.294)	(0.354)	(0.530)
Journal impact factor		0.030**		
		(0.014)		
Std.Error* Impact Factor		0.084***		
		(0.022)		
Dummy if obs > median(obs)			-0.066	
			(0.285)	
Std.Error*D if obs > median(obs)			0.113	
			(0.540)	
Dummy reported prior to 2009				-0.122
				(0.304)
Std.Error*D reported prior to 2009				0.217
				(0.614)
Constant	0.424**	-0.027	0.400**	0.416*
	(0.158)	(0.221)	(0.158)	(0.248)
Observations	780	780	780	780
Adjusted R^2	0.404	0.456	0.408	0.420

Columns (1) to (4) estimated using WLS. Standard errors (in parentheses) are clustered at the study level. Significance levels are * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

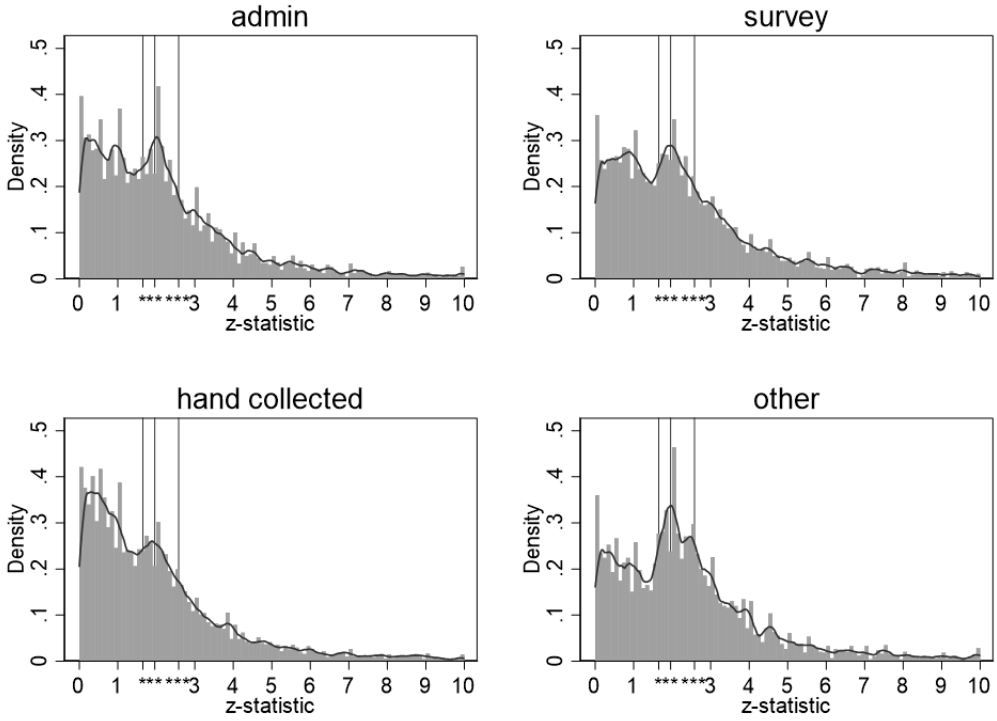
Included standard errors as explanatory variables are normalized. It allows an interpretation as standard deviation.

APPENDIX A. APPENDIX: THE ETI

Appendix to Chapter 3

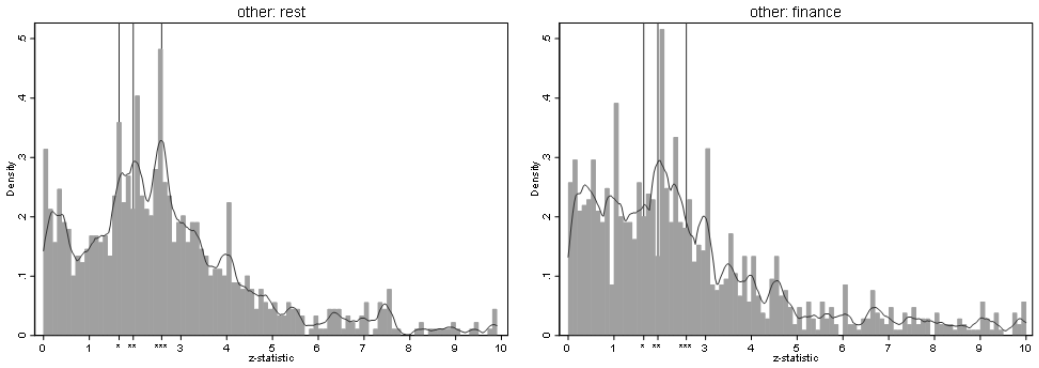
B.1 Appendix Figures and Tables

Figure B.1: z-statistics by Method of Data Collection (full sample used by Brodeur et al. (2020)



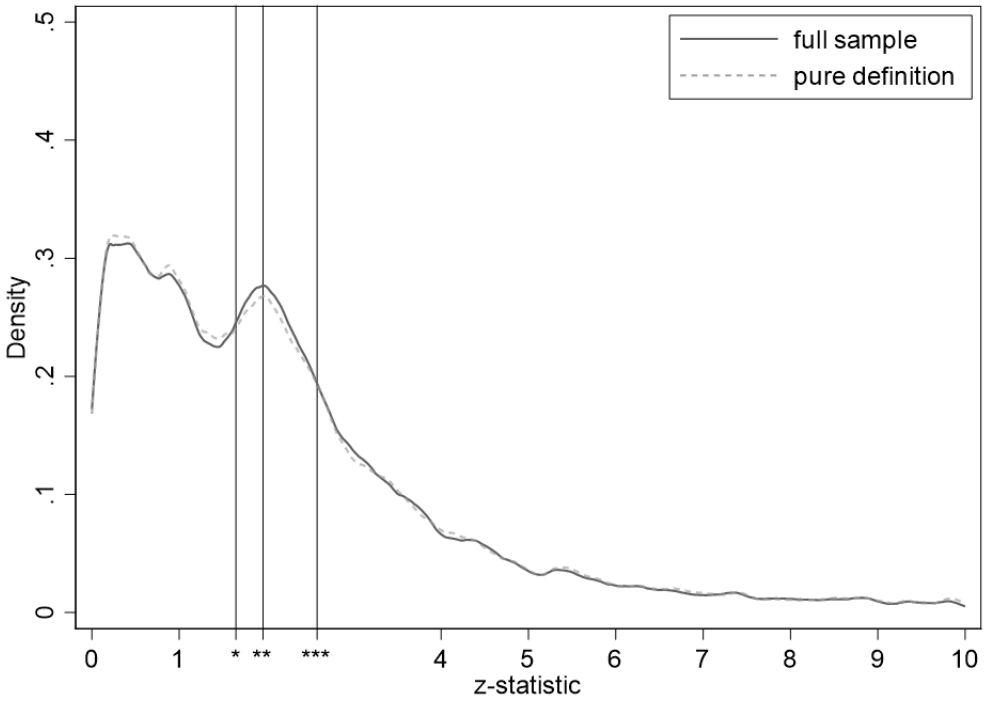
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by method of data collection: *admin*, *survey*, *hand collected* and *other*. In comparison to Figure 2.2, we consider the full sample used by Brodeur et al. (2020). The full sample consists of 21,440 test statistics. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.2: z-statistics for Method of Data Collection: *other*, Non-financial vs financial Data



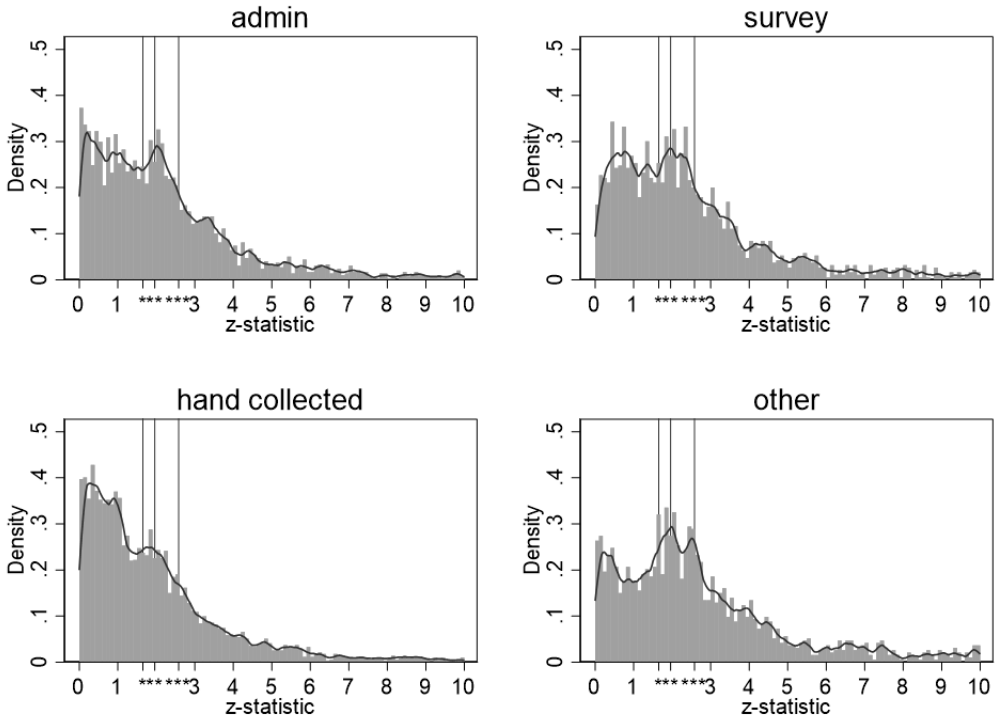
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ only for those that rely on *other data*. We split the data type category *other* into those test statistics that use non-financial data (left figure) and those that only rely on financial data (right figure). In total 2,019 test statistics belong to the data type category *other*. While 46% (N=934) belong to non-financial data, 53.74% (N=1,085) use financial data. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.3: z-statistics for all Estimates vs z-statistics for those Estimates that Rely Solely on One Data Type (De-rounded z-statistics)



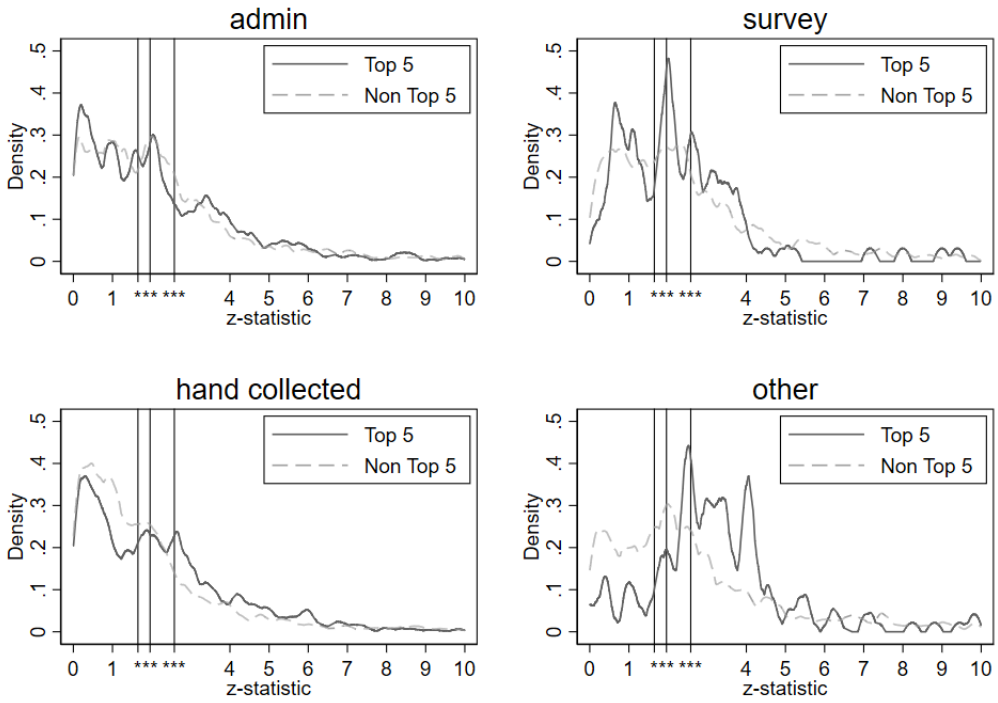
Notes: This figure displays two distributions. First, the solid line plots z-statistics for the sample used in Brodeur et al. (2020) and second, the dashed line plots z-statistics for the sub-sample of estimates that rely solely on one data type. Compared to figure 2.1 we use de-rounded z-statistics. Both figures are based on an Epanechnikov kernel with a bandwidth of 0.1. Estimates are not weighted.

Figure B.4: z-statistics by Method of Data Collection (De-rounded z-statistics)



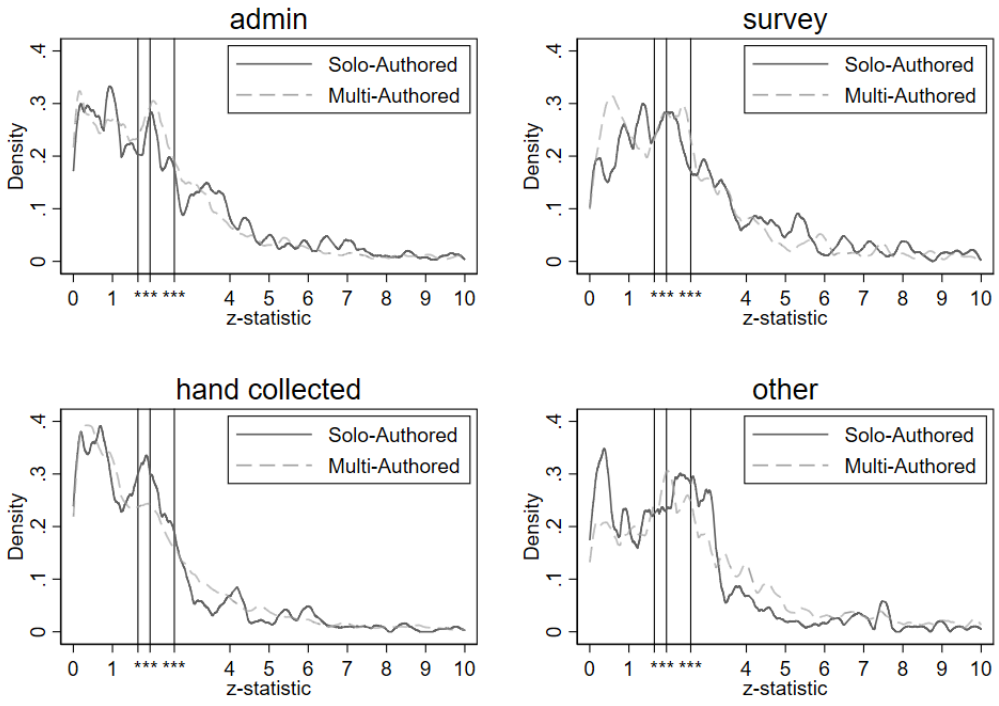
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by Method of Data Collection: *admin*, *survey*, *hand collected* and *other*. We only consider those observations that rely solely on one data type within each primary study. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. Compared to Figure 2.2 we use de-rounded z-statistics. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.5: z-statistics by Method of Data Collection and Journal Ranking



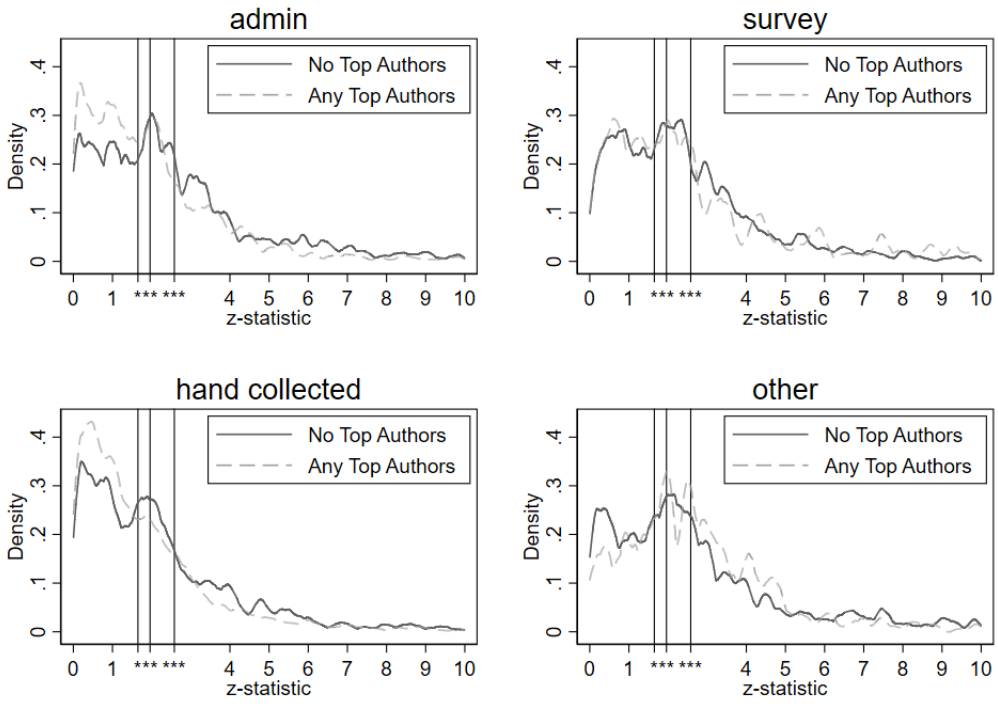
Notes: This figure displays histograms of test statistics for $z \in [0, 01]$. Test statistics are partitioned by method of data collection: admin, survey, hand collected and other. Lines in dark grey are for articles published in the top 5. Lines in light grey (dashes) are for articles published in non-top 5. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. All figures are based on Epanechnikov kernel. Estimates are not weighted.

Figure B.6: z-statistics by Method of Data Collection and Number of Authors



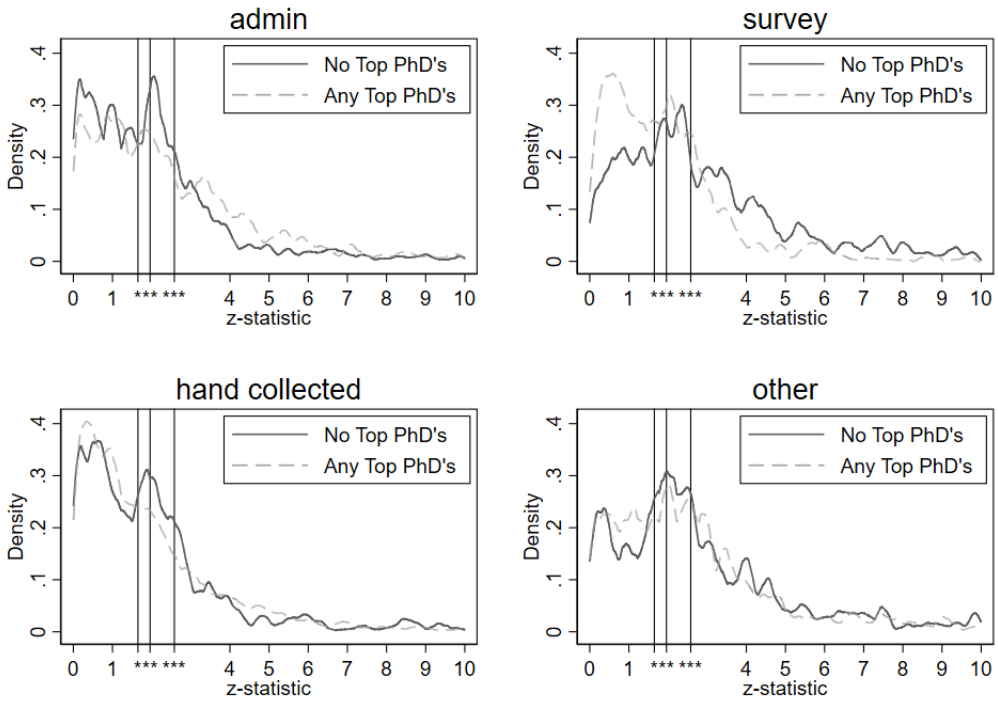
Notes: This figure displays histograms of test statistics for $z \in [0, 01]$. Test statistics are partitioned by method of data collection: admin, survey, hand collected and other. Lines in dark grey are for sole authored. Lines in light grey (dashes) are for multi authored articles. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. All figures are based on Epanechnikov kernel. Estimates are not weighted.

Figure B.7: z-statistics by Method of Data Collection and Affiliation



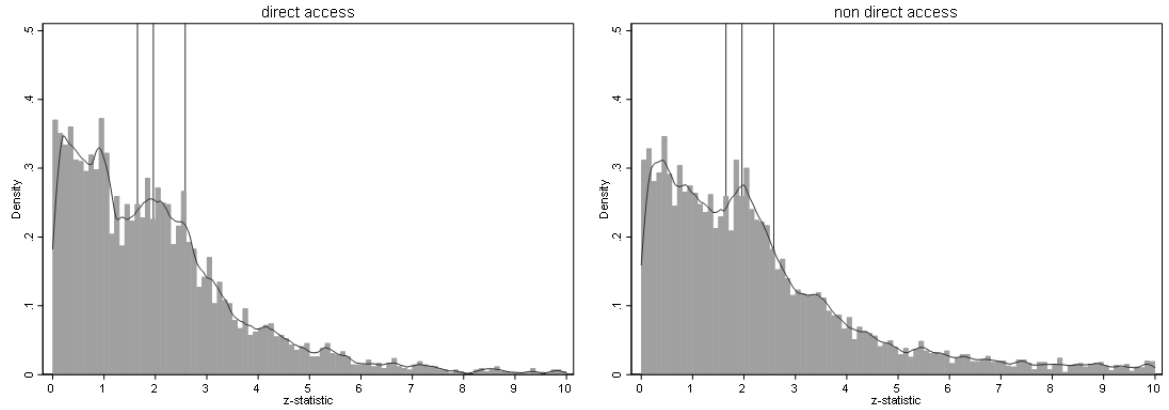
Notes: his figure displays histograms of test statistics for $z \in [0, 01]$. Test statistics are partitioned by method of data collection: admin, survey, hand collected and other. Lines in dark grey are for articles with at least one author affiliated to a top institution. Lines in light grey (dashes) are for articles with no author affiliated in a top institution. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. All figures are based on Epanechnikov kernel. Estimates are not weighted.

Figure B.8: z-statistics by Method of Data Collection and PhD Institution



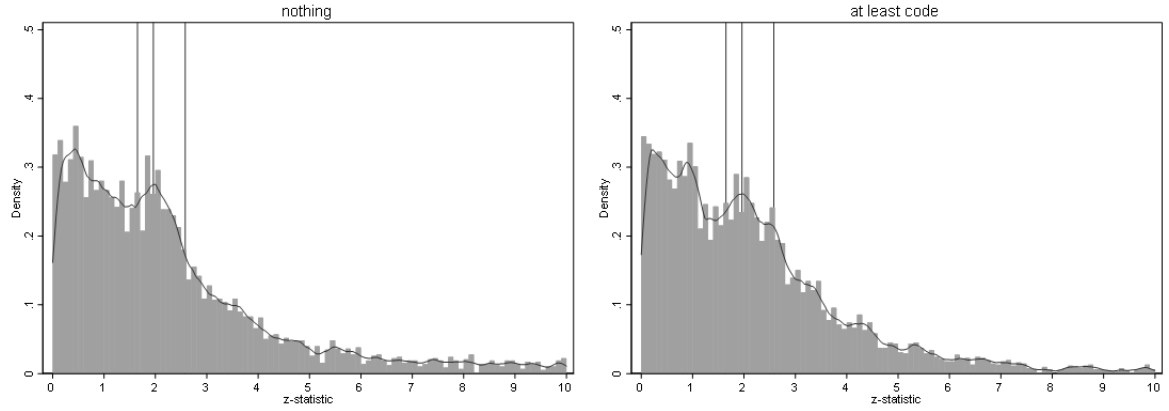
Notes: This figure displays histograms of test statistics for $z \in [0, 01]$. Test statistics are partitioned by method of data collection: admin, survey, hand collected and other. Lines in dark grey are for articles with at least one author who graduated from a top institution. Lines in light grey (dashes) are for articles with no author who graduated from a top institution. Bins are 0.1 wide. Reference lines are displayed at the conventional two-tailed significance levels. All figures are based on Epanechnikov kernel. Estimates are not weighted.

Figure B.9: z-statistics by Accessibility of Replication Material: Data *and* Code (De-rounded z-statistics)



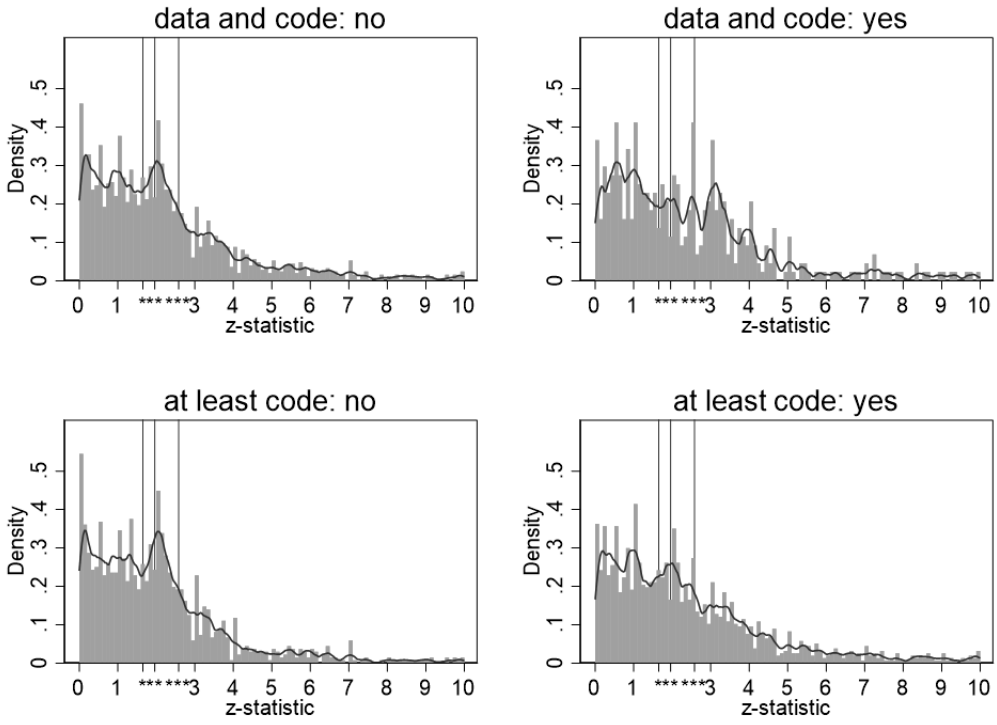
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. The left figure restricts the sample to estimates that provide direct access to data and code. The right figure restricts the sample to estimates that do not provide both data and code. We only consider those observations that rely solely on one data type within each primary study. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We use de-rounded z-statistics. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.10: z-statistics by Availability of Replication Material: At Least Code (De-rounded z-statistics)



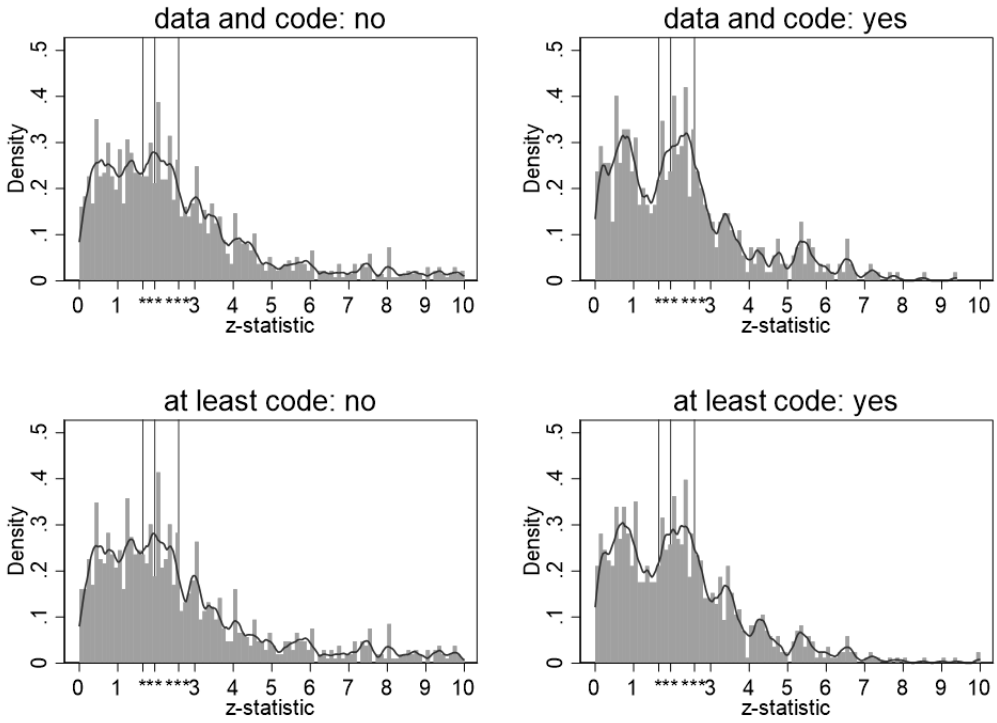
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. The left figure restricts the sample to estimates that do not provide any replication material (i.e., data and/or code). The right figure restricts the sample to estimates that at least provide code for replication. We only consider those observations that rely solely on one data type within each primary study. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. Compared to Figure 2.4 we use de-rounded z-statistics. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.11: z-statistics for Method of Data Collection: *admin*, Provision of Replication Material



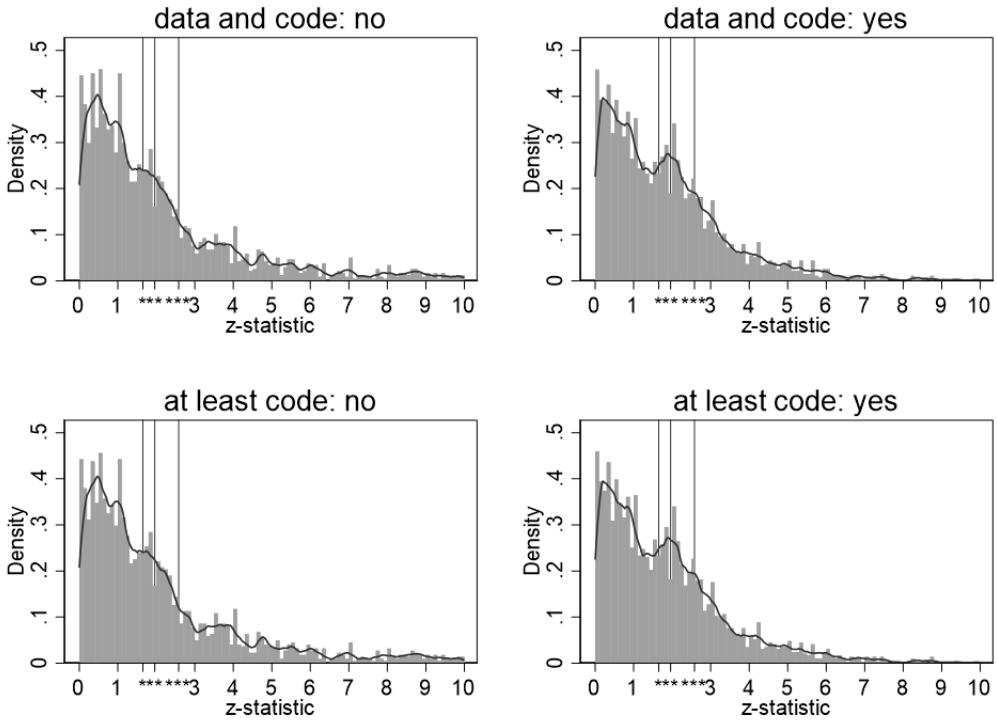
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ only for admin data. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.12: z-statistics for Method of Data Collection: *survey*, Provision of Replication Material



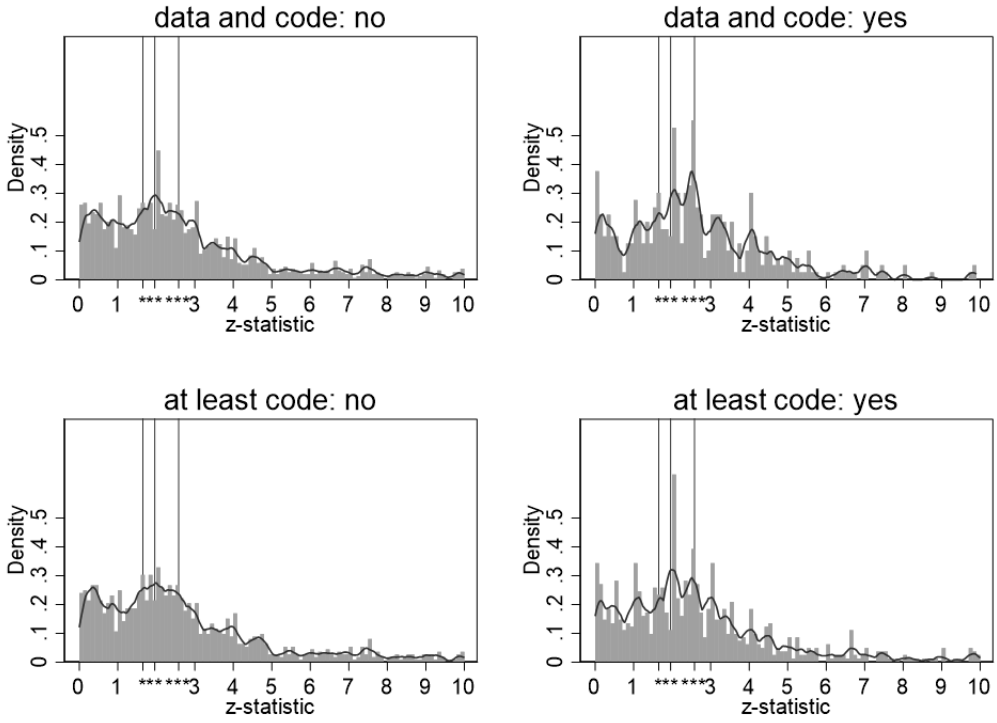
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ only for survey data. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.13: z-statistics for Method of Data Collection: *hand collected*, Provision of Replication Material



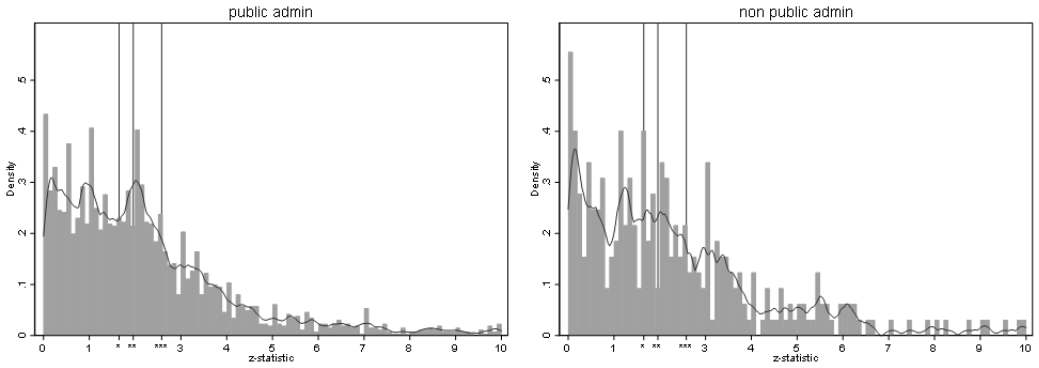
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ only for hand collected data. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.14: z-statistics for Method of Data Collection: *other*, Provision of Replication Material



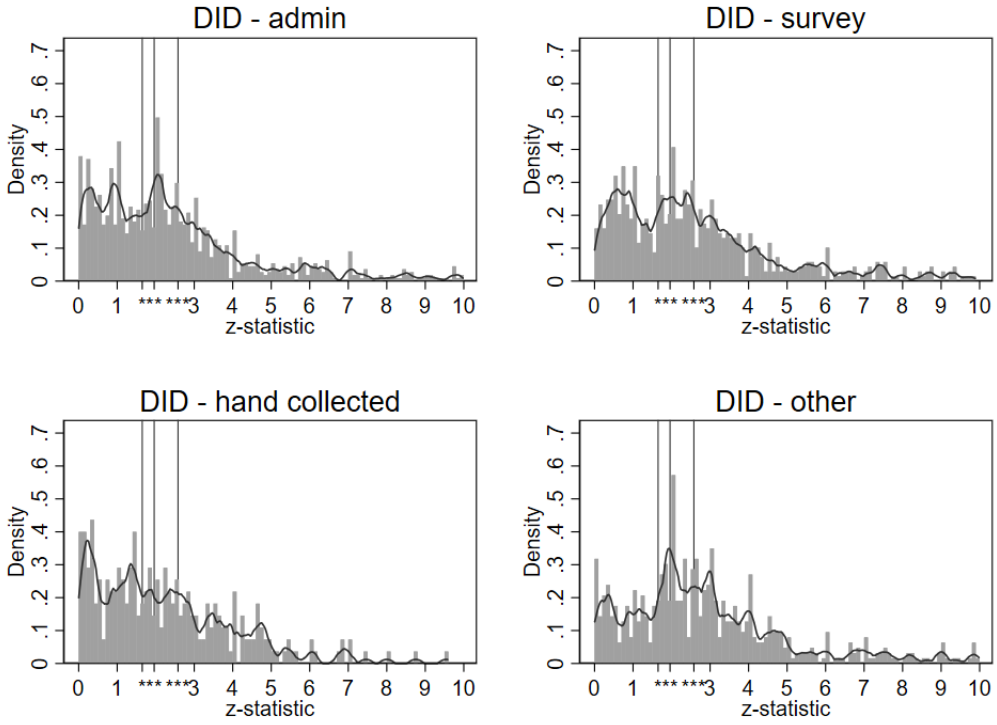
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ only for other data. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.15: z-statistics for those Estimates that rely on Admin Data: Public vs non-Public Admin Data



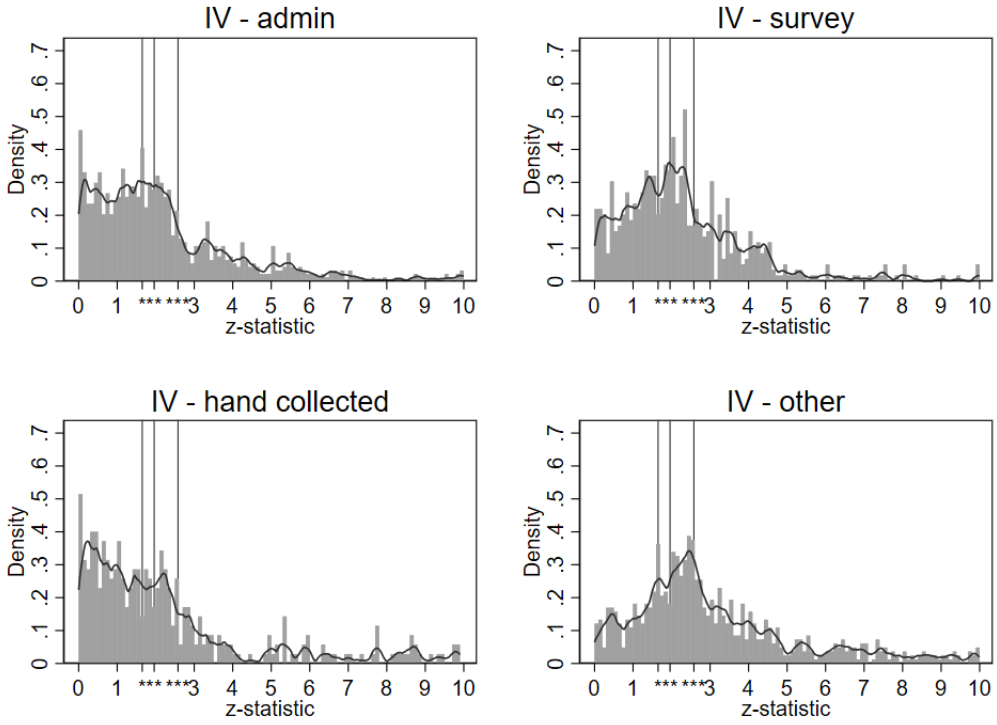
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$. The left figure restricts the sample to estimates that use publicly available admin data (e.g., tax return data) (N=2,883). The right figure restricts the sample to estimates that use non-publicly available data (e.g., electricity usage data) (N=329). Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.16: z-statistics using DID by Method of Data Collection



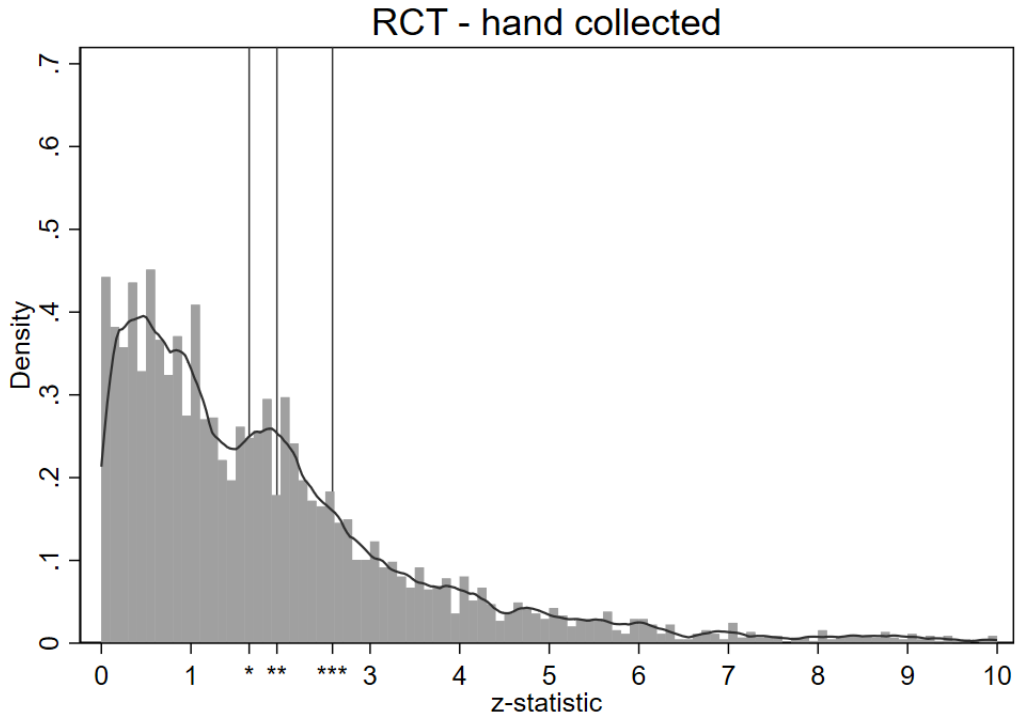
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ using a difference-in-differences (DID) approach by method of data collection: *admin*, *survey*, *hand collected* and *other*. We only consider those observations that rely solely on one type of data within each primary study. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.17: z-statistics using IV by Method of Data Collection



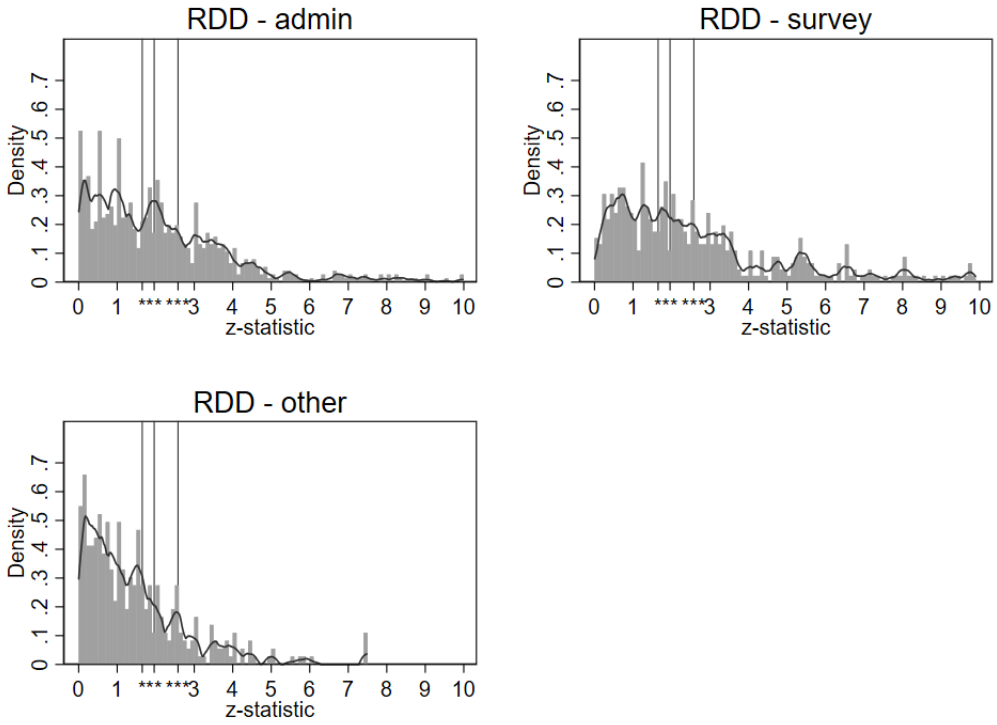
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ using an instrumental variables (IV) approach by method of data collection: *admin*, *survey*, *hand collected* and *other*. We only consider those observations that rely solely on one type of data within each primary study. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.18: z-statistics using RCT



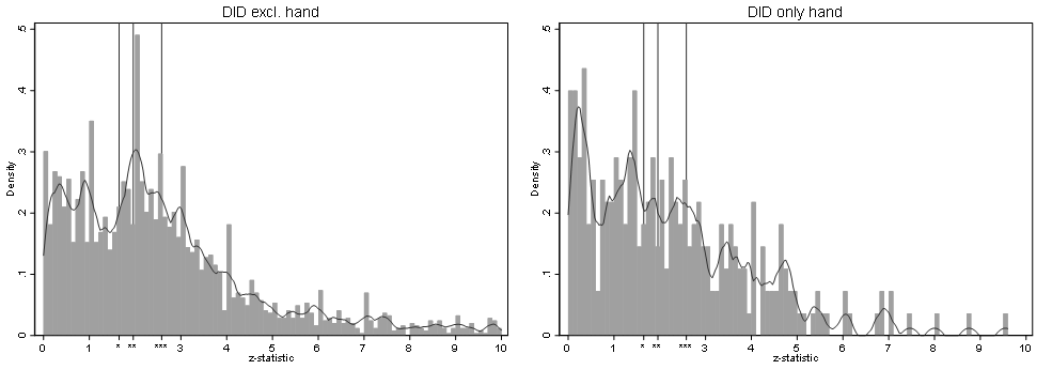
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ using randomized control trials (RCT) and *hand collected* data. We only consider those observations that rely on the same type of data within each primary study. Due to a small amount of observations, we do not display a graph for those estimates that rely on *admin*, *survey* or *other* data. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.19: z-statistics using RDD by Method of Data Collection



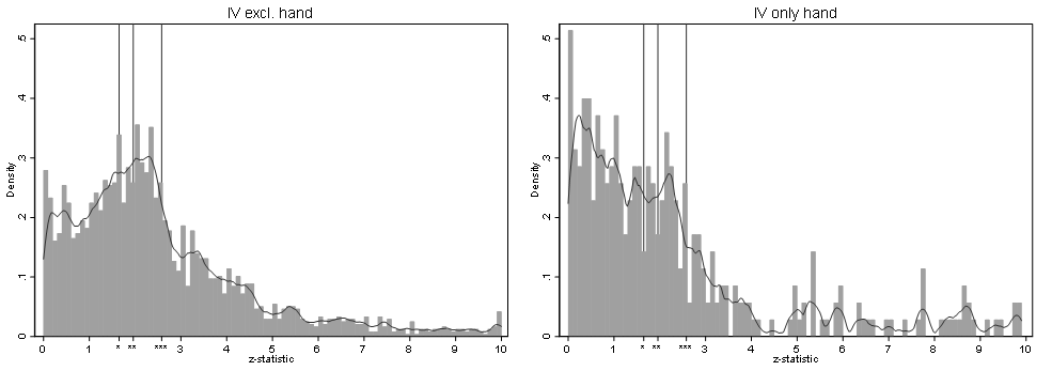
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ using regression discontinuity design (RDD) by method of data collection: *admin*, *survey* and *other*. We only consider those observations that rely on the same type of data within each primary study. Due to a small amount of observations, we do not display a graph for those estimates that rely on *hand collected* data. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.20: z-statistics using DID: hand vs non hand collected data



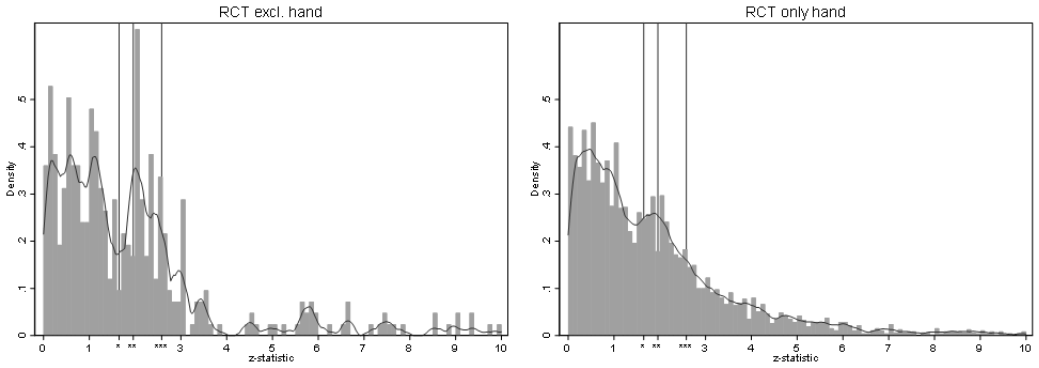
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ only for those that rely on DID. The left figure shows those tests that use non hand collected data (N=2,637) and the right figure only those that use hand collected data (N=279). Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.21: z-statistics using IV: hand vs non hand collected data



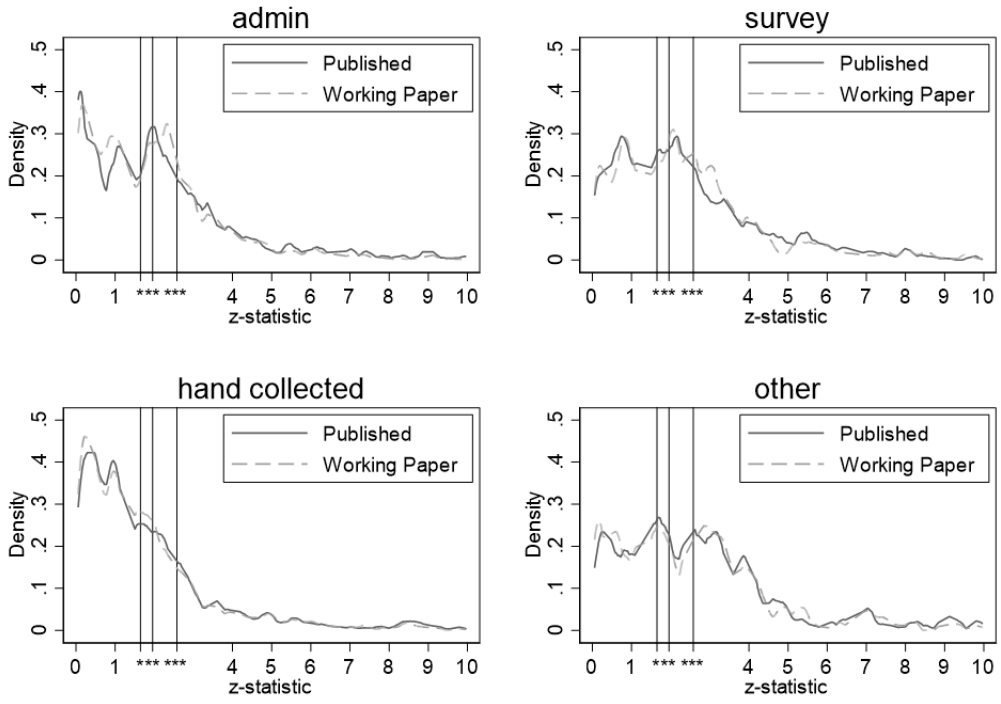
Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ only for those that rely on IV. The left figure shows those tests that use non hand collected data (N=2,467) and the right figure only those that use hand collected data (N=377). Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.22: z-statistics using RCT only: hand vs non hand collected data



Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ only for those that rely on RCT. The left figure shows those tests that use non hand collected data (N=451) and the right figure only those that use hand collected data (N=4583). Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates.

Figure B.23: Histogram by Publication Status and Method - Balanced Sample



Notes: This figure displays histograms of test statistics for $z \in [0, 10]$ by method of data collection: *admin*, *survey*, *hand collected* and *other*. The solid line represent published z statistics, while the dashed line represent those from working papers. The samples is accordingly restricted to estimates from published articles that had an associated working paper. Histogram bins are 0.1 wide. Reference lines are displayed at conventional two-tailed significance levels. We impose an Epanechnikov kernel and do not weight our estimates. No weights have been applied.

Table B.1: Examples by Method of Data Collection

Admin Data - examples:
School/ student data and test scores
Continuous Wage and Benefit History (CWBH) UI administration records
Medicare Beneficiaries and Claim Data
Crime/ Police Data/ Court Records
Tax Return Data
Survey Data - examples:
UChicago Consortium on School Research / Chicago Public Schools (CPS) survey
Medical Expenditure Panel Survey (MEPS)
American Community Survey (ACS)
Consumer Expenditure Survey (CE)
German Socio-Economic Panel (GSOEP)
Hand Collected Data - examples:
Own experiments (lab experiments, mTurk/ online experiments, field experiments)
Own Surveys
Self-collected school data ('School visits')
Self-collected Performance Data
Other Data - examples:
Fortune 500 list of companies
Compustat
Maddison Historical Statistics
Thomson Reuters Datastream
CRSP – stock market data

Notes: For each method of data collection, this table illustrates typical examples.

Table B.2: Summary Statistics: Method of Data Collection by Journal

	Share of Articles by Type of Data				Total	
	admin	survey	hand c.	other	Tests	Articles
American Economic Journal: Applied Economics	36.36	21.21	39.39	9.09	1545	33
American Economic Journal: Economic Policy	47.83	30.43	26.09	0	559	23
American Economic Journal: Macroeconomics	0	100	0	0	22	3
American Economic Review	37.84	16.22	32.43	16.22	1338	37
Econometrica	100	0	0	0	24	3
Economic Policy	0	33.33	0	66.67	53	3
Experimental Economics	0	0	100	0	73	5
Journal of Applied Econometrics	0	66.67	0	33.33	51	3
Journal of Development Economics	11.11	25	58.33	5.6	1618	36
Journal of Economic Growth	0	14.29	0	85.71	98	7
Journal of Financial Economics	16.67	5.556	0	77.78	294	18
Journal of Financial Intermediation	0	14.29	14.29	71.43	102	7
Journal of Human Resources	6.250	50	37.50	6.25	682	16
Journal of International Economics	20	30	0	50	295	10
Journal of Labor Economics	26.67	26.67	33.33	13.33	512	15
Journal of Political Economy	27.27	9.091	45.45	18.18	484	11
Journal of Public Economics	43.48	13.04	34.78	8.70	1297	46
Journal of Urban Economics	58.33	41.67	0	0	324	12
Journal of the European Economic Association	18.18	0	54.55	27.27	333	11
Review of Financial Studies	12.50	16.67	0	70.83	361	24
The Economic Journal	22.22	22.22	16.67	38.89	450	18
The Journal of Finance	13.33	6.667	20	60	696	15
The Quarterly Journal of Economics	50	0	50	0	471	12
The Review of Economic Studies	33.33	0	66.67	0	189	3
The Review of Economics and Statistics	42.31	23.08	26.92	11.54	624	26
Share of Articles by Datatype	28.11	19.90	29.10	22.89	.	.
Share of Tests by Datatype	25.71	16	42.13	16.16	.	.
Total Tests	3212	1999	5265	2019	12495	.
Total Articles	113	80	117	92	.	402

Notes: This table alphabetically presents our sample of Top 25 journals identified using RePEc's Simple Impact Factor: https://ideas.repec.org/top/top.journals.simple10.html. Some top journals did not have any eligible articles in the first data collection period: Journal of Economic Literature, Journal of Economic Perspectives, Journal of Monetary Economics, Review of Economic Dynamics, Annals of Economics and Finance, and the Annual Review of Economics. We also excluded Brookings Papers on Economic Activity from the sample. We only consider those estimates that rely solely on one type of data.

Table B.3: Summary Statistics: Replication Characteristics by Journal

Provision of:	Data *and* Code		at least Code	
Journal	No	Yes	No	Yes
American Economic Journal: Applied Econo	753	792	63	1,482
American Economic Journal: Economic Poli	277	282		559
American Economic Journal: Macroeconomic	8	14		22
American Economic Review	668	670		1338
Econometrica	9	15		24
Economic Policy	32	21	32	21
Experimental Economics	73		73	
Journal of Applied Econometrics	30	21	30	21
Journal of Development Economics	1,211	407	1,187	431
Journal of Economic Growth	62	36	62	36
Journal of Financial Economics	294		294	
Journal of Financial Intermediation	102		102	
Journal of Human Resources	682		682	
Journal of International Economics	295		295	
Journal of Labor Economics	100	412		512
Journal of Political Economy	65	419	36	448
Journal of Public Economics	1,297		1,297	
Journal of Urban Economics	324		324	
Journal of the European Economic Associa	39	294	29	304
Review of Financial Studies	361		361	
The Economic Journal	160	290	127	323
The Journal of Finance	696		672	24
The Quarterly Journal of Economics	215	256	215	256
The Review of Economic Studies	36	153		189
The Review of Economics and Statistics	447	177	259	365
Total	8,236	4,259	6,140	6,355

Notes: This table provides an overview for our two replication variables by journal: Direct Access to *Data (and Code)* and *(at least) Provision of Code*. Direct accessibility of data also involves the provision of code. The variable 'Provision of Code' consider test statistics that at least provide the code.

Table B.4: Summary Statistics: Method of Data Collection by Estimation Method

	admin	hand c.	other	survey	Total
Method: DID					
Total Tests	1,260	279	656	721	2,916
Tests in %	43.21	9.57	22.50	24.73	100.00
Method: IV					
Total Tests	974	377	876	617	2,844
Tests in %	34.25	13.26	30.80	21.69	100.00
Method: RCT					
Total Tests	143	4,583	119	189	5,034
Tests in %	2.84	91.04	2.36	3.75	100.00
Method: RDD					
Total Tests	835	26	368	472	1,701
Tests in %	49.09	1.53	21.63	27.75	100.00
Total					
Total Tests	3,212	5,265	2,019	1,999	12,495
Tests in %	25.71	42.14	16.16	16.00	100.00

Notes: This table provides an overview of test statistics by type of data and method under study. For example, 1,260 Tests employ a DID setting and use admin data. Among those tests that employ a DID setting 43.21% use admin data, while only 9.57% use hand collected data.

Table B.5: Caliper Test, Significant at the 5 percent level, bootstrap errors: Unweighted Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	-0.001	-0.001	0.029	0.016	0.018	0.016
	(0.037)	(0.038)	(0.047)	(0.052)	(0.048)	(0.047)
hand collected	-0.076	-0.062	-0.052	-0.087	-0.083	-0.087
	(0.026)	(0.028)	(0.038)	(0.066)	(0.056)	(0.056)
other	0.025	0.019	0.016	0.007	0.009	0.009
	(0.044)	(0.053)	(0.066)	(0.049)	(0.055)	(0.058)
Estimation Method: (omitted RCT)						
DID				-0.013	-0.015	-0.016
				(0.048)	(0.041)	(0.051)
IV				-0.032	-0.034	-0.035
				(0.065)	(0.055)	(0.049)
RDD				-0.106	-0.108	-0.108
				(0.074)	(0.067)	(0.068)
Controls						
Top 5		0.029	0.117			
		(0.033)	(0.048)			
Year=2018		0.009	0.011	0.012	0.012	0.013
		(0.029)	(0.034)	(0.026)	(0.033)	(0.030)
Experience		-0.004	-0.005	-0.006	-0.006	-0.006
		(0.008)	(0.009)	(0.009)	(0.008)	(0.009)
Experience^2		0.011	0.016	0.019	0.019	0.019
		(0.025)	(0.034)	(0.028)	(0.025)	(0.030)
Top Institution		-0.037	-0.021	-0.020	-0.020	-0.022
		(0.035)	(0.050)	(0.038)	(0.046)	(0.046)
PhD Top Institution		-0.001	-0.023	-0.029	-0.030	-0.028
		(0.038)	(0.039)	(0.040)	(0.032)	(0.053)
Replication Characteristics						
Direct Access to Data & Code					-0.011	
					(0.038)	
Provision of (at least) Code						-0.025
						(0.054)
Other Controls						
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Journal FE			Y	Y	Y	Y
Observations	2,904	2,904	2,904	2,904	2,904	2,904
Window	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.5]	[1.96±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Bootstrapped standard errors are in parentheses. Observations are unweighted.

Table B.6: Caliper Test, Significant at the 1 percent level: Unweighted Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.018	0.032	0.049	0.058	0.054	0.058
	(0.040)	(0.041)	(0.046)	(0.046)	(0.046)	(0.046)
hand collected	0.015	0.019	0.011	0.038	0.034	0.040
	(0.039)	(0.037)	(0.040)	(0.052)	(0.051)	(0.051)
other	0.081	0.089	0.076	0.099	0.097	0.097
	(0.042)	(0.043)	(0.044)	(0.042)	(0.042)	(0.042)
Estimation Method: (omitted RCT)						
DID				0.059	0.062	0.065
				(0.044)	(0.045)	(0.044)
IV				-0.044	-0.042	-0.039
				(0.044)	(0.045)	(0.044)
RDD				0.051	0.055	0.056
				(0.056)	(0.057)	(0.056)
Controls						
Top 5		0.105	0.072	0.071	0.060	0.026
		(0.037)	(0.102)	(0.101)	(0.103)	(0.108)
Year = 2018		0.019	0.022	0.021	0.022	0.020
		(0.028)	(0.028)	(0.027)	(0.027)	(0.027)
Experience		0.001	0.002	0.001	0.001	0.001
		(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Experience^2		-0.004	-0.011	-0.009	-0.009	-0.010
		(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Top Institution		-0.058	-0.055	-0.035	-0.036	-0.033
		(0.041)	(0.040)	(0.040)	(0.040)	(0.040)
PhD Top Institution		0.017	0.016	0.012	0.014	0.010
		(0.037)	(0.037)	(0.036)	(0.036)	(0.036)
Replication Characteristics						
Direct Access to Data & Code					0.017	
					(0.034)	
Provision of (at least) Code						0.054
						(0.056)
Other Controls						
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Journal FE			Y	Y	Y	Y
Observations	2,250	2,250	2,247	2,247	2,247	2,247
Window	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table B.7: Caliper Test, Significant at the 5 percent level, Unweighted Estimates, Replication Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	-0.004	-0.008	0.018	-0.001	-0.002	0.016
	(0.039)	(0.037)	(0.041)	(0.038)	(0.036)	(0.041)
hand collected	-0.082	-0.075	-0.083	-0.077	-0.064	-0.087
	(0.029)	(0.029)	(0.050)	(0.027)	(0.027)	(0.051)
other	0.023	0.014	0.009	0.025	0.018	0.009
	(0.046)	(0.048)	(0.054)	(0.046)	(0.046)	(0.052)
Estimation Method: (omitted RCT)						
DID			-0.015			-0.016
			(0.043)			(0.042)
IV			-0.034			-0.035
			(0.052)			(0.050)
RDD			-0.108			-0.108
			(0.058)			(0.056)
Controls						
Top 5		0.020	0.110		0.022	0.121
		(0.030)	(0.065)		(0.033)	(0.072)
Year=2018		0.011	0.012		0.011	0.013
		(0.027)	(0.026)		(0.026)	(0.026)
Experience		-0.005	-0.006		-0.004	-0.006
		(0.007)	(0.007)		(0.007)	(0.007)
Experience^2		0.012	0.019		0.011	0.019
		(0.021)	(0.023)		(0.021)	(0.023)
Top Institution		-0.036	-0.020		-0.036	-0.022
		(0.038)	(0.041)		(0.038)	(0.041)
PhD Top Institution		-0.001	-0.030		-0.004	-0.028
		(0.038)	(0.037)		(0.036)	(0.038)
Replication Characteristics						
Direct Access to Data & Code	0.014	0.029	-0.011			
	(0.025)	(0.026)	(0.034)			
Provision of (at least) Code				0.006	0.016	-0.025
				(0.025)	(0.026)	(0.043)
Other Controls						
Reporting Method		Y	Y		Y	Y
Solo Authored		Y	Y		Y	Y
Share Female Authors		Y	Y		Y	Y
Editor		Y	Y		Y	Y
Journal FE			Y			Y
Observations	2,904	2,904	2,904	2,904	2,904	2,904
Window	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Column (1)-(3) add direct access to data and code and Column (4)-(6) add provision of (at least) code as a control variable. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table B.8: Caliper Test, Significant at the 10 percent level, Unweighted Estimates, Replication Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.020	0.020	0.014	0.027	0.025	0.010
	(0.038)	(0.035)	(0.039)	(0.036)	(0.034)	(0.039)
hand collected	-0.056	-0.025	-0.028	-0.047	-0.017	-0.035
	(0.034)	(0.032)	(0.054)	(0.030)	(0.029)	(0.054)
other	0.037	0.044	0.034	0.042	0.044	0.029
	(0.037)	(0.038)	(0.042)	(0.036)	(0.037)	(0.042)
Estimation Method: (omitted RCT)						
DID			-0.029			-0.023
			(0.045)			(0.045)
IV			-0.062			-0.055
			(0.045)			(0.045)
RDD			-0.112			-0.107
			(0.053)			(0.053)
Controls						
Top 5		0.062	0.098		0.050	0.079
		(0.032)	(0.087)		(0.033)	(0.085)
Year=2018		-0.014	-0.011		-0.011	-0.010
		(0.024)	(0.024)		(0.023)	(0.024)
Experience		0.006	0.006		0.006	0.006
		(0.007)	(0.007)		(0.006)	(0.007)
Experience^2		-0.035	-0.027		-0.034	-0.027
		(0.019)	(0.020)		(0.019)	(0.020)
Top Institution		-0.057	-0.048		-0.056	-0.046
		(0.036)	(0.036)		(0.035)	(0.036)
PhD Top Institution		-0.059	-0.070		-0.067	-0.070
		(0.029)	(0.032)		(0.029)	(0.032)
Replication Characteristics						
Direct Access to Data & Code	0.028	0.029	-0.023			
	(0.028)	(0.027)	(0.032)			
Provision of (at least) Code				0.041	0.044	0.013
				(0.024)	(0.024)	(0.040)
Other Controls						
Reporting Method		Y	Y		Y	Y
Solo Authored		Y	Y		Y	Y
Share Female Authors		Y	Y		Y	Y
Editor		Y	Y		Y	Y
Journal FE			Y			Y
Observations	2,933	2,933	2,926	2,933	2,933	2,926
Window	[1.65±0.50	[1.65±0.50	[1.65±0.50	[1.65±0.50	[1.65±0.50	[1.65±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. Column (1)-(3) add direct access to data and code and Column (4)-(6) add provision of (at least) code as a control variable. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table B.9: Caliper Test, Significant at the 1 percent level, Unweighted Estimates, Replication Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.016	0.034	0.054	0.018	0.032	0.058
	(0.040)	(0.041)	(0.046)	(0.039)	(0.041)	(0.046)
hand collected	0.010	0.023	0.034	0.010	0.018	0.040
	(0.039)	(0.038)	(0.051)	(0.038)	(0.036)	(0.051)
other	0.081	0.090	0.097	0.084	0.088	0.097
	(0.042)	(0.042)	(0.042)	(0.041)	(0.042)	(0.042)
Estimation Method: (omitted RCT)						
DID			0.062			0.065
			(0.045)			(0.044)
IV			-0.042			-0.039
			(0.045)			(0.044)
RDD			0.055			0.056
			(0.057)			(0.056)
Controls						
Top 5		0.108	0.060		0.099	0.026
		(0.038)	(0.103)		(0.040)	(0.108)
Year=2018		0.018	0.022		0.021	0.020
		(0.029)	(0.027)		(0.028)	(0.027)
Experience		0.001	0.001		0.001	0.001
		(0.007)	(0.007)		(0.007)	(0.007)
Experience^2		-0.004	-0.009		-0.004	-0.010
		(0.018)	(0.019)		(0.019)	(0.019)
Top Institution		-0.058	-0.036		-0.058	-0.033
		(0.041)	(0.040)		(0.041)	(0.040)
PhD Top Institution		0.017	0.014		0.015	0.010
		(0.037)	(0.036)		(0.037)	(0.036)
Replication Characteristics						
Direct Access to Data & Code	0.011	-0.009	0.017			
	(0.031)	(0.031)	(0.034)			
Provision of (at least) Code				0.037	0.011	0.054
				(0.027)	(0.030)	(0.056)
Other Controls						
Reporting Method		Y	Y		Y	Y
Solo Authored		Y	Y		Y	Y
Share Female Authors		Y	Y		Y	Y
Editor		Y	Y		Y	Y
Journal FE			Y			Y
Observations	2,250	2,250	2,247	2,250	2,250	2,247
Window	[2.58±0.50	[2.58±0.50	[2.58±0.50	[2.58±0.50	[2.58±0.50	[2.58±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. Column (1)-(3) add direct access to data and code and Column (4)-(6) add provision of (at least) code as a control variable. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table B.10: Caliper Test, Significant at the 5 percent level. (aw weights)

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.021	0.017	0.030	0.024	0.023	0.026
	(0.057)	(0.056)	(0.057)	(0.056)	(0.054)	(0.056)
hand collected	-0.100	-0.096	-0.046	-0.045	-0.046	-0.043
	(0.049)	(0.045)	(0.043)	(0.066)	(0.065)	(0.066)
other	0.057	0.059	0.037	0.031	0.030	0.034
	(0.057)	(0.054)	(0.054)	(0.057)	(0.057)	(0.057)
Estimation Method: (omitted RCT)						
DID				-0.026	-0.026	-0.028
				(0.059)	(0.060)	(0.059)
IV				0.024	0.025	0.022
				(0.060)	(0.061)	(0.059)
RDD				-0.010	-0.010	-0.008
				(0.087)	(0.088)	(0.087)
Controls						
Top 5		-0.001	0.129	0.125	0.123	0.174
		(0.064)	(0.071)	(0.070)	(0.074)	(0.081)
Year=2018		-0.013	0.005	0.007	0.007	0.013
		(0.038)	(0.036)	(0.036)	(0.036)	(0.037)
Experience		-0.006	-0.008	-0.006	-0.006	-0.007
		(0.010)	(0.009)	(0.009)	(0.009)	(0.009)
Experience^2		0.005	0.019	0.014	0.014	0.017
		(0.025)	(0.024)	(0.025)	(0.026)	(0.026)
Top Institution		-0.005	0.033	0.028	0.028	0.027
		(0.069)	(0.056)	(0.057)	(0.057)	(0.057)
PhD Top Institution		-0.032	-0.048	-0.048	-0.047	-0.047
		(0.054)	(0.051)	(0.051)	(0.051)	(0.051)
Replication Characteristics						
Direct Access to Data & Code					0.004	
					(0.047)	
Provision of (at least) Code						-0.065
						(0.051)
Other Controls						
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Journal FE			Y	Y	Y	Y
Observations	2,904	2,904	2,904	2,904	2,904	2,904
Window	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table B.11: Caliper Test, Significant at the 10 percent level. (aw weights)

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.071	0.056	0.037	0.033	0.036	0.036
	(0.058)	(0.059)	(0.058)	(0.057)	(0.057)	(0.057)
hand collected	-0.022	-0.017	-0.016	-0.066	-0.063	-0.066
	(0.048)	(0.051)	(0.052)	(0.083)	(0.084)	(0.083)
other	0.066	0.095	0.044	0.044	0.046	0.047
	(0.053)	(0.051)	(0.055)	(0.055)	(0.055)	(0.055)
Estimation Method: (omitted RCT)						
DID				-0.040	-0.041	-0.042
				(0.080)	(0.080)	(0.080)
IV				-0.083	-0.084	-0.087
				(0.069)	(0.069)	(0.068)
RDD				-0.094	-0.095	-0.095
				(0.083)	(0.084)	(0.083)
Controls						
Top 5		0.108	0.187	0.172	0.178	0.208
		(0.051)	(0.107)	(0.104)	(0.105)	(0.112)
Year=2018		0.017	0.018	0.018	0.018	0.023
		(0.039)	(0.038)	(0.037)	(0.037)	(0.037)
Experience		0.009	0.002	-0.000	-0.000	-0.001
		(0.010)	(0.008)	(0.009)	(0.009)	(0.009)
Experience^2		-0.035	-0.010	-0.003	-0.003	-0.001
		(0.025)	(0.023)	(0.025)	(0.025)	(0.026)
Top Institution		0.011	-0.014	-0.018	-0.018	-0.019
		(0.065)	(0.053)	(0.053)	(0.053)	(0.053)
PhD Top Institution		-0.154	-0.124	-0.129	-0.130	-0.129
		(0.051)	(0.047)	(0.047)	(0.047)	(0.046)
Replication Characteristics						
Direct Access to Data & Code					-0.011	
					(0.047)	
Provision of (at least) Code						-0.056
						(0.057)
Other Controls						
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Journal FE			Y	Y	Y	Y
Observations	2,933	2,933	2,926	2,926	2,926	2,926
Window	[1.65±0.50	[1.65±0.50	[1.65±0.50	[1.65±0.50	[1.65±0.50	[1.65±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations..

Table B.12: Caliper Test, Significant at the 1 percent level. (aw weights)

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.024	0.011	0.016	0.036	0.053	0.036
	(0.078)	(0.075)	(0.073)	(0.071)	(0.070)	(0.071)
hand collected	-0.011	0.008	0.021	0.035	0.060	0.035
	(0.058)	(0.055)	(0.059)	(0.085)	(0.083)	(0.084)
other	0.020	0.041	0.004	0.021	0.029	0.021
	(0.064)	(0.057)	(0.056)	(0.055)	(0.053)	(0.055)
Estimation Method: (omitted RCT)						
DID				0.074	0.054	0.074
				(0.078)	(0.079)	(0.079)
IV				-0.049	-0.058	-0.049
				(0.076)	(0.077)	(0.077)
RDD				-0.010	-0.025	-0.010
				(0.092)	(0.091)	(0.092)
Controls						
Top 5		0.020	-0.106	-0.094	-0.011	-0.089
		(0.068)	(0.134)	(0.129)	(0.136)	(0.143)
Year=2018		0.012	0.042	0.039	0.035	0.040
		(0.048)	(0.047)	(0.047)	(0.046)	(0.047)
Experience		0.007	0.009	0.007	0.006	0.007
		(0.010)	(0.010)	(0.011)	(0.011)	(0.011)
Experience^2		-0.011	-0.022	-0.018	-0.012	-0.017
		(0.031)	(0.032)	(0.034)	(0.034)	(0.035)
Top Institution		-0.018	0.018	0.046	0.051	0.046
		(0.070)	(0.068)	(0.067)	(0.066)	(0.067)
PhD Top Institution		-0.021	-0.039	-0.061	-0.065	-0.061
		(0.059)	(0.058)	(0.055)	(0.055)	(0.055)
Replication Characteristics						
Direct Access to Data & Code					-0.118	
					(0.056)	
Provision of (at least) Code						-0.007
						(0.081)
Other Controls						
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Journal FE			Y	Y	Y	Y
Observations	2,250	2,250	2,247	2,247	2,247	2,247
Window	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. Robust standard errors are in parentheses, clustered by article. We use the inverse of the number of tests presented in the same article to weight observations.

Table B.13: Caliper Test, Significant at the 5 percent level, different thresholds: Unweighted Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.018	-0.024	-0.024	0.016	-0.030	-0.035
	(0.041)	(0.045)	(0.055)	(0.041)	(0.044)	(0.054)
hand collected	-0.083	-0.082	-0.081	-0.087	-0.094	-0.100
	(0.050)	(0.058)	(0.064)	(0.051)	(0.057)	(0.061)
other	0.009	0.011	0.060	0.009	0.010	0.057
	(0.054)	(0.052)	(0.060)	(0.052)	(0.051)	(0.059)
Estimation Method: (omitted RCT)						
DID	-0.015	-0.031	-0.061	-0.016	-0.032	-0.058
	(0.043)	(0.051)	(0.056)	(0.042)	(0.050)	(0.055)
IV	-0.034	-0.055	-0.041	-0.035	-0.056	-0.039
	(0.052)	(0.057)	(0.062)	(0.050)	(0.056)	(0.061)
RDD	-0.108	-0.111	-0.121	-0.108	-0.110	-0.117
	(0.058)	(0.065)	(0.070)	(0.056)	(0.064)	(0.070)
Controls						
Year=2018	0.012	-0.002	0.022	0.013	0.000	0.024
	(0.026)	(0.028)	(0.031)	(0.026)	(0.028)	(0.031)
Top 5	0.110	0.128	0.172	0.121	0.145	0.177
	(0.065)	(0.082)	(0.112)	(0.072)	(0.091)	(0.118)
Experience	-0.006	-0.011	0.015	-0.006	-0.011	0.014
	(0.007)	(0.008)	(0.010)	(0.007)	(0.008)	(0.010)
Experience^2	0.019	0.037	-0.043	0.019	0.038	-0.042
	(0.023)	(0.025)	(0.035)	(0.023)	(0.025)	(0.035)
Top Institution	-0.020	0.004	-0.038	-0.022	0.000	-0.039
	(0.041)	(0.043)	(0.048)	(0.041)	(0.044)	(0.048)
PhD Top Institution	-0.030	-0.006	0.045	-0.028	-0.001	0.050
	(0.037)	(0.040)	(0.045)	(0.038)	(0.041)	(0.046)
Replication Characteristics						
Direct Access to Data & Code	-0.011	-0.033	-0.057			
	(0.034)	(0.038)	(0.042)			
Provision of (at least) Code				-0.025	-0.051	-0.056
				(0.043)	(0.048)	(0.061)
Other Controls						
Reporting Method	Y	Y	Y	Y	Y	Y
Solo Authored	Y	Y	Y	Y	Y	Y
Share Female Authors	Y	Y	Y	Y	Y	Y
Editor	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y
Observations	2,904	2,109	1,244	2,904	2,109	1,244
Window	[1.96±0.50	[1.96±0.35	[1.96±0.2]	[1.96±0.50	[1.96±0.35	[1.96±0.2]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Column (1) and (4) display our baseline results already shown in Table 2.5. In column (2) and (5), we restrict to $z \in [1, 61, 2.31]$, while column (3) and (6) restricts the sample to $z \in [1, 76, 2.16]$ Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table B.14: Caliper Test, Significant at the 10 percent level, different thresholds: Unweighted Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.014	0.029	0.012	0.010	0.032	0.028
	(0.039)	(0.047)	(0.063)	(0.039)	(0.046)	(0.063)
hand collected	-0.028	-0.009	-0.041	-0.035	0.001	-0.013
	(0.054)	(0.064)	(0.073)	(0.054)	(0.062)	(0.071)
other	0.034	0.016	0.003	0.029	0.016	0.010
	(0.042)	(0.050)	(0.061)	(0.042)	(0.050)	(0.060)
Estimation Method: (omitted RCT)						
DID	-0.029	-0.017	0.027	-0.023	-0.014	0.026
	(0.045)	(0.054)	(0.066)	(0.045)	(0.054)	(0.067)
IV	-0.062	-0.047	-0.031	-0.055	-0.044	-0.034
	(0.045)	(0.053)	(0.062)	(0.045)	(0.053)	(0.061)
RDD	-0.112	-0.064	-0.001	-0.107	-0.063	-0.004
	(0.053)	(0.064)	(0.075)	(0.053)	(0.064)	(0.075)
Controls						
Year=2018	-0.011	0.003	0.035	-0.010	0.001	0.032
	(0.024)	(0.027)	(0.033)	(0.024)	(0.026)	(0.033)
Top 5	0.098	0.006	-0.010	0.079	-0.015	-0.020
	(0.087)	(0.082)	(0.121)	(0.085)	(0.081)	(0.120)
Experience	0.006	0.004	0.014	0.006	0.004	0.014
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)
Experience^2	-0.027	-0.020	-0.040	-0.027	-0.021	-0.042
	(0.020)	(0.019)	(0.021)	(0.020)	(0.019)	(0.021)
Top Institution	-0.048	-0.020	-0.066	-0.046	-0.016	-0.057
	(0.036)	(0.038)	(0.046)	(0.036)	(0.038)	(0.046)
PhD Top Institution	-0.070	-0.066	-0.094	-0.070	-0.070	-0.102
	(0.032)	(0.035)	(0.042)	(0.032)	(0.036)	(0.042)
Replication Characteristics						
Direct Access to Data & Code	-0.023	0.026	0.075			
	(0.032)	(0.036)	(0.046)			
Provision of (at least) Code				0.013	0.057	0.082
				(0.040)	(0.044)	(0.059)
Other Controls						
Reporting Method	Y	Y	Y	Y	Y	Y
Solo Authored	Y	Y	Y	Y	Y	Y
Share Female Authors	Y	Y	Y	Y	Y	Y
Editor	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y
Observations	2,926	1,934	1,109	2,926	1,934	1,109
Window	[1.65±0.50	[1.65±0.35	[1.65±0.2]	[1.65±0.50	[1.65±0.35	[1.65±0.2]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. Column (1) and (4) display our baseline results already shown in Table 2.6. In column (2) and (5), we restrict to $z \in [1, 3, 2]$, while column (3) and (6) restricts the sample to $z \in [1, 85, 2.45]$ Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table B.15: Caliper Test, Significant at the 1 percent level, different thresholds: Unweighted Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.054	0.014	-0.010	0.058	0.019	-0.011
	(0.046)	(0.047)	(0.064)	(0.046)	(0.047)	(0.064)
hand collected	0.034	0.046	0.010	0.040	0.052	0.010
	(0.051)	(0.054)	(0.066)	(0.051)	(0.054)	(0.065)
other	0.097	0.065	0.063	0.097	0.067	0.059
	(0.042)	(0.048)	(0.069)	(0.042)	(0.048)	(0.072)
Estimation Method: (omitted RCT)						
DID	0.062	0.075	0.021	0.065	0.074	0.031
	(0.045)	(0.051)	(0.059)	(0.044)	(0.050)	(0.056)
IV	-0.042	-0.028	-0.004	-0.039	-0.028	0.004
	(0.045)	(0.047)	(0.054)	(0.044)	(0.048)	(0.054)
RDD	0.055	0.035	-0.031	0.056	0.033	-0.022
	(0.057)	(0.061)	(0.076)	(0.056)	(0.060)	(0.074)
Controls						
Year=2018	0.022	0.044	0.022	0.020	0.043	0.023
	(0.027)	(0.031)	(0.044)	(0.027)	(0.030)	(0.043)
Top 5	0.060	0.061	0.159	0.026	0.050	0.098
	(0.103)	(0.090)	(0.099)	(0.108)	(0.102)	(0.113)
Experience	0.001	0.002	0.001	0.001	0.002	0.000
	(0.007)	(0.008)	(0.010)	(0.007)	(0.008)	(0.010)
Experience^2	-0.009	-0.011	-0.029	-0.010	-0.011	-0.030
	(0.019)	(0.023)	(0.032)	(0.019)	(0.023)	(0.031)
Top Institution	-0.036	0.023	0.035	-0.033	0.025	0.036
	(0.040)	(0.047)	(0.061)	(0.040)	(0.047)	(0.062)
PhD Top Institution	0.014	-0.040	-0.022	0.010	-0.042	-0.022
	(0.036)	(0.039)	(0.062)	(0.036)	(0.040)	(0.063)
Replication Characteristics						
Direct Access to Data & Code	0.017	0.017	-0.015			
	(0.034)	(0.039)	(0.053)			
Provision of (at least) Code				0.054	0.025	0.055
				(0.056)	(0.059)	(0.064)
Other Controls						
Reporting Method	Y	Y	Y	Y	Y	Y
Solo Authored	Y	Y	Y	Y	Y	Y
Share Female Authors	Y	Y	Y	Y	Y	Y
Editor	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y
Observations	2,247	1,524	871	2,247	1,524	871
Window	[2.58±0.50	[2.58±0.35	[2.58±0.2]	[2.58±0.50	[2.58±0.35	[2.58±0.2]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 1 percent level. Column (1) and (4) display our baseline results already shown in Table B.6. In column (2) and (5), we restrict to $z \in [1.15, 2.15]$, while column (3) and (6) restricts the sample to $z \in [1.45, 1.85]$ Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table B.16: Caliper Test, Significant at the 5 percent level, Ambiguous removed

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.021	0.017	0.032	0.028	0.029	0.031
	(0.058)	(0.057)	(0.058)	(0.058)	(0.055)	(0.057)
hand collected	-0.096	-0.091	-0.040	-0.029	-0.028	-0.024
	(0.049)	(0.045)	(0.044)	(0.068)	(0.067)	(0.068)
other	0.061	0.065	0.047	0.043	0.043	0.047
	(0.058)	(0.055)	(0.055)	(0.059)	(0.059)	(0.059)
Estimation Method: (omitted RCT)						
DID				-0.012	-0.013	-0.013
				(0.061)	(0.063)	(0.060)
IV				0.036	0.036	0.035
				(0.062)	(0.064)	(0.061)
RDD				0.013	0.013	0.017
				(0.090)	(0.091)	(0.090)
Controls						
Top 5		0.006	0.138	0.136	0.139	0.195
		(0.064)	(0.072)	(0.071)	(0.076)	(0.083)
Year=2018		-0.015	0.001	0.004	0.004	0.011
		(0.039)	(0.037)	(0.037)	(0.037)	(0.037)
Experience		-0.007	-0.009	-0.007	-0.007	-0.008
		(0.010)	(0.009)	(0.009)	(0.009)	(0.010)
Experience^2		0.008	0.021	0.015	0.016	0.019
		(0.025)	(0.025)	(0.026)	(0.026)	(0.026)
Top Institution		-0.011	0.025	0.022	0.023	0.022
		(0.070)	(0.057)	(0.058)	(0.058)	(0.058)
PhD Top Institution		-0.028	-0.043	-0.041	-0.042	-0.040
		(0.055)	(0.052)	(0.052)	(0.052)	(0.052)
Replication Characteristics						
Direct Access to Data & Code					-0.004	
					(0.048)	
Provision of (at least) Code						-0.079
						(0.051)
Other Controls						
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Journal FE			Y	Y	Y	Y
Observations	2,863	2,863	2,863	2,863	2,863	2,863
Window	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted. We remove ambiguous tests.

Table B.17: Caliper Test, Significant at the 10 percent level, Ambiguous removed

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.025	0.009	0.023	0.024	0.023	0.025
	(0.044)	(0.042)	(0.040)	(0.040)	(0.039)	(0.040)
hand collected	-0.033	-0.038	-0.005	-0.016	-0.017	-0.015
	(0.035)	(0.033)	(0.037)	(0.063)	(0.064)	(0.063)
other	0.025	0.042	0.032	0.032	0.031	0.033
	(0.041)	(0.033)	(0.040)	(0.041)	(0.041)	(0.041)
Estimation Method: (omitted RCT)						
DID				0.016	0.017	0.016
				(0.059)	(0.059)	(0.059)
IV				-0.029	-0.029	-0.030
				(0.052)	(0.052)	(0.052)
RDD				-0.027	-0.027	-0.025
				(0.064)	(0.064)	(0.064)
Controls						
Top 5		0.013	0.084	0.083	0.080	0.109
		(0.033)	(0.065)	(0.068)	(0.073)	(0.081)
Year=2018		0.018	0.032	0.031	0.031	0.033
		(0.027)	(0.028)	(0.028)	(0.028)	(0.028)
Experience		0.005	0.002	0.000	0.000	-0.000
		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Experience^2		-0.022	-0.009	-0.004	-0.005	-0.003
		(0.015)	(0.015)	(0.015)	(0.015)	(0.016)
Top Institution		0.060	0.041	0.044	0.044	0.043
		(0.036)	(0.037)	(0.037)	(0.037)	(0.036)
PhD Top Institution		-0.113	-0.113	-0.116	-0.116	-0.115
		(0.033)	(0.035)	(0.035)	(0.036)	(0.035)
Replication Characteristics						
Direct Access to Data & Code					0.004	
					(0.035)	
Provision of (at least) Code						-0.036
						(0.046)
Other Controls						
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Journal FE			Y	Y	Y	Y
Observations	2,863	2,860	2,829	2,829	2,829	2,829
Window	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted. We remove ambiguous tests.

Table B.18: Caliper Test, Significant at the 5 percent level: Unweighted Estimates (De-rounded)

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	-0.009	-0.001	0.032	0.025	0.031	0.026
	(0.039)	(0.038)	(0.042)	(0.041)	(0.042)	(0.041)
hand collected	-0.046	-0.043	-0.035	-0.042	-0.033	-0.042
	(0.025)	(0.026)	(0.028)	(0.042)	(0.045)	(0.042)
other	0.008	-0.009	-0.026	-0.031	-0.027	-0.029
	(0.035)	(0.036)	(0.041)	(0.039)	(0.039)	(0.039)
Estimation Method: (omitted RCT)						
DID				0.021	0.018	0.018
				(0.042)	(0.042)	(0.042)
IV				-0.005	-0.008	-0.008
				(0.046)	(0.046)	(0.045)
RDD				-0.073	-0.075	-0.074
				(0.051)	(0.050)	(0.050)
Controls						
Top 5		0.019	0.087	0.077	0.091	0.098
		(0.028)	(0.063)	(0.061)	(0.067)	(0.076)
Year=2018		0.022	0.022	0.023	0.023	0.023
		(0.023)	(0.023)	(0.023)	(0.022)	(0.023)
Experience		-0.005	-0.005	-0.006	-0.005	-0.006
		(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Experience^2		0.014	0.015	0.018	0.017	0.018
		(0.021)	(0.023)	(0.023)	(0.023)	(0.023)
Top Institution		-0.043	-0.028	-0.021	-0.022	-0.024
		(0.032)	(0.037)	(0.036)	(0.036)	(0.035)
PhD Top Institution		-0.001	-0.019	-0.025	-0.026	-0.023
		(0.029)	(0.032)	(0.031)	(0.031)	(0.031)
Replication Characteristics						
Direct Access to Data & Code					-0.024	
					(0.032)	
Provision of (at least) Code						-0.032
						(0.044)
Other Controls						
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Journal FE			Y	Y	Y	Y
Observations	2,691	2,691	2,686	2,686	2,686	2,686
Window	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. We rely on de-rounded z-statistics. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table B.19: Caliper Test, Significant at the 10 percent level: Unweighted Estimates (De-rounded)

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	0.050	0.057	0.045	0.036	0.040	0.035
	(0.034)	(0.035)	(0.039)	(0.040)	(0.041)	(0.040)
hand collected	-0.017	0.009	0.019	-0.010	-0.003	-0.010
	(0.028)	(0.028)	(0.031)	(0.045)	(0.048)	(0.045)
other	0.076	0.080	0.067	0.060	0.064	0.058
	(0.036)	(0.037)	(0.043)	(0.041)	(0.042)	(0.042)
Estimation Method: (omitted RCT)						
DID				-0.016	-0.018	-0.013
				(0.040)	(0.040)	(0.040)
IV				-0.024	-0.028	-0.021
				(0.040)	(0.040)	(0.040)
RDD				-0.099	-0.101	-0.098
				(0.049)	(0.049)	(0.049)
Controls						
Top 5		0.070	0.134	0.122	0.132	0.108
		(0.030)	(0.080)	(0.080)	(0.080)	(0.079)
Year=2018		-0.005	-0.004	-0.006	-0.007	-0.006
		(0.024)	(0.025)	(0.025)	(0.025)	(0.025)
Experience		0.004	0.004	0.003	0.003	0.003
		(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
Experience^2		-0.027	-0.024	-0.021	-0.021	-0.021
		(0.019)	(0.021)	(0.021)	(0.021)	(0.021)
Top Institution		-0.054	-0.043	-0.041	-0.042	-0.039
		(0.034)	(0.033)	(0.034)	(0.034)	(0.034)
PhD Top Institution		-0.053	-0.048	-0.053	-0.054	-0.054
		(0.026)	(0.031)	(0.032)	(0.031)	(0.031)
Direct Access to Data & Code					-0.020	
					(0.032)	
Provision of (at least) Code						0.023
						(0.041)
Other Controls						
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Journal FE			Y	Y	Y	Y
Observations	2,734	2,734	2,727	2,727	2,727	2,727
Window	[1.65±0.50]	[1.65±0.50]	[1.65±0.50]	[1.65±0.50]	[1.65±0.50]	[1.65±0.50]

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 10 percent level. We rely on de-rounded z-statistics. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table B.20: Caliper Test, Significant at the 5 percent level, logit: Unweighted Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Method of Data Collection: (omitted admin)						
survey	-0.001	-0.001	0.029	0.016	0.018	0.016
	(0.038)	(0.036)	(0.040)	(0.041)	(0.041)	(0.041)
hand collected	-0.076	-0.062	-0.052	-0.087	-0.084	-0.088
	(0.027)	(0.027)	(0.031)	(0.051)	(0.050)	(0.051)
other	0.025	0.019	0.016	0.007	0.009	0.009
	(0.045)	(0.047)	(0.053)	(0.051)	(0.053)	(0.051)
Estimation Method: (omitted RCT)						
DID				-0.014	-0.016	-0.017
				(0.043)	(0.044)	(0.043)
IV				-0.033	-0.035	-0.037
				(0.051)	(0.052)	(0.050)
RDD				-0.108	-0.110	-0.110
				(0.057)	(0.058)	(0.057)
Controls						
Top 5		0.029	0.117	0.103	0.110	0.121
		(0.029)	(0.060)	(0.060)	(0.066)	(0.073)
Year=2018		0.009	0.011	0.012	0.012	0.013
		(0.027)	(0.027)	(0.026)	(0.026)	(0.026)
Experience		-0.004	-0.005	-0.006	-0.006	-0.006
		(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Experience^2		0.011	0.015	0.018	0.019	0.019
		(0.021)	(0.023)	(0.023)	(0.024)	(0.024)
Top Institution		-0.037	-0.021	-0.020	-0.020	-0.022
		(0.038)	(0.042)	(0.041)	(0.041)	(0.041)
PhD Top Institution		-0.001	-0.022	-0.029	-0.030	-0.027
		(0.038)	(0.040)	(0.038)	(0.037)	(0.038)
Replication Characteristics						
Direct Access to Data & Code					-0.012	
					(0.034)	
Provision of (at least) Code						-0.026
						(0.043)
Other Controls						
Reporting Method		Y	Y	Y	Y	Y
Solo Authored		Y	Y	Y	Y	Y
Share Female Authors		Y	Y	Y	Y	Y
Editor		Y	Y	Y	Y	Y
Journal FE			Y	Y	Y	Y
Observations	2,904	2,904	2,904	2,904	2,904	2,904
Window	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50	[1.96±0.50]

Notes: This table reports marginal effects from logit regressions (equation (2.4)). The dependent variable is a dummy for whether the test statistic is significant at the 5 percent level. Robust standard errors are in parentheses, clustered by article. Observations are unweighted.

Table B.21: Excess Coefficients by Significance Region

	(1)	(2)	(3)	(4)
$[0, 1.65)$	admin	survey	hand	other
Observed	0.402	0.368	0.517	0.316
Expected	0.319	0.383	0.533	0.347
Difference	0.083	-0.016	-0.016	-0.031
Ratio to Excess to Expected	0.259	-0.041	-0.029	-0.089
$[1.65, 1.96)$	admin	survey	hand	other
Observed	0.070	0.083	0.074	0.070
Expected	0.085	0.086	0.079	0.086
Difference	-0.015	-0.003	-0.005	-0.016
Ratio to Excess to Expected	-0.177	-0.035	-0.060	-0.184
$[1.96, 2.58)$	admin	survey	hand	other
Observed	0.143	0.162	0.127	0.168
Expected	0.146	0.139	0.116	0.143
Difference	-0.003	0.022	0.011	0.025
Ratio to Excess to Expected	-0.017	0.161	0.095	0.176
$[2.58, 5)$	admin	survey	hand	other
Observed	0.204	0.239	0.179	0.278
Expected	0.280	0.249	0.180	0.266
Difference	-0.076	-0.010	-0.002	0.012
Ratio to Excess to Expected	-0.271	-0.039	-0.008	0.046
$[5, \infty)$	admin	survey	hand	other
Observed	0.171		0.093	0.158
Expected	0.170	0.143	0.093	0.158
Difference	0.000		0.000	-0.001
Ratio to Excess to Expected	0.001		0.003	-0.004
Degrees of Freedom	2	2	2	2
Non-centrality Parameter	1.970	1.720	1.140	1.860

Notes: Each panel of the table is a separate significance region. In each panel and for each type of data, we report four statistics: 1) The observed mass of test statistics 2) The expected mass informed by a calibrated t distribution 3) The difference and 4) The ratio of the observed to expected. For the difference and ratio, a negative value implies 'missing' test statistics in the region whereas a positive number implies an excess of test statistics. The degrees of freedom and the non-centrality parameter for the *t*-distribution that fit the observed data best are presented at the bottom.

Table B.22: Excess Coefficients by Significance Region in Comparison to Admin Data

	(1)	(2)	(3)
$[0, 1.65)$	survey	hand	other
Observed	0,368	0,517	0,316
Expected (admin)	0,402	0,402	0,402
Difference	-0.034	0.116	-0,086
Ratio to Excess to Expected	-0,084	0.288	-0,213
$[1.65, 1.96)$	survey	hand	other
Observed	0,083	0,074	0,070
Expected (admin)	0,070	0,070	0,070
Difference	0,012	-0,004	-0,000
Ratio to Excess to Expected	0,178	0.057	-0,003
$[1.96, 2.58)$	survey	hand	other
Observed	0,162	0,127	0,168
Expected (admin)	0,143	0,143	0,143
Difference	0,018	-0,017	0,025
Ratio to Excess to Expected	0,128	-0.115	0,176
$[2.58, 5)$	survey	hand	other
Observed	0,239	0,179	0,278
Expected (admin)	0,204	0,204	0,204
Difference	0,035	-0,025	0,074
Ratio to Excess to Expected	0,173	-0,124	0,365
$[5, \infty)$	survey	hand	other
Observed		0,093	0,158
Expected (admin)	0,171	0,171	0,171
Difference		-0.078	-0,013
Ratio to Excess to Expected		-0.456	-0,077

Notes: Each panel of the table is a separate significance region. In each panel and for each type of data, we report four statistics: 1) The observed mass of test statistics 2) The expected mass informed by a calibrated t distribution 3) The difference and 4) The ratio of the observed to expected. For the difference and ratio, a negative value implies 'missing' test statistics in the region whereas a positive number implies an excess of test statistics.

Table B.23: Excess Coefficients by Significance Region: Replication Characteristics

	(1)	(2)	(3)
$[0, 1.65)$	data and code	at least code	no data and/or code
Observed	0.456	0.433	0.429
Expected	0.540	0.487	0.350
Difference	-0.084	-0.055	0.080
Ratio to Excess to Expected	-0.155	-0.112	0.228
$[1.65, 1.96)$	data and code	at least code	no data and/or code
Observed	0.076	0.075	0.072
Expected	0.078	0.082	0.086
Difference	-0.002	-0.007	-0.013
Ratio to Excess to Expected	-0.030	-0.083	-0.155
$[1.96, 2.58)$	data and code	at least code	no data and/or code
Observed	0.151	0.149	0.137
Expected	0.114	0.124	0.143
Difference	0.037	0.025	-0.006
Ratio to Excess to Expected	0.323	0.205	-0.041
$[2.58, 5)$	data and code	at least code	no data and/or code
Observed	0.216	0.225	0.196
Expected	0.177	0.200	0.265
Ratio to Excess to Expected	0.039	0.025	-0.069
Difference	0.221	0.125	-0.260
$[5, \infty)$	data and code	at least code	no data and/or code
Observed	0.091	0.107	0.157
Expected	0.1090	0.106	0.157
Difference	0.000	0.000	0.000
Ratio to Excess to Expected	0.002	0.002	0.001
Degrees of Freedom	2	2	2
Non-centrality Parameter	1.11	1.32	1.85

Notes: Each panel of the table is a separate significance region. In each panel and for each type of data, we report four statistics: 1) The observed mass of test statistics 2) The expected mass informed by a calibrated t distribution 3) The difference and 4) The ratio of the observed to expected. For the difference and ratio, a negative value implies 'missing' test statistics in the region whereas a positive number implies an excess of test statistics.

Table B.24: Working Paper Available?

	(1)	(2)	(3)	(4)
Method of Data Collection: (omitted admin)				
survey	0.099	0.085	0.090	-0.016
	(0.081)	(0.120)	(0.112)	(0.114)
hand collected	0.031	0.050	0.033	-0.054
	(0.072)	(0.115)	(0.106)	(0.103)
other	-0.080	-0.073	0.048	-0.033
	(0.075)	(0.116)	(0.132)	(0.119)
Controls				
Top 5		-0.146	-0.549	-0.335
		(0.115)	(0.211)	(0.226)
Year=2018		0.089	0.082	0.082
		(0.087)	(0.081)	(0.073)
Experience		0.002	0.004	0.012
		(0.020)	(0.019)	(0.019)
Experience^2		-0.048	-0.043	-0.054
		(0.055)	(0.055)	(0.054)
Top Institution		-0.069	-0.065	-0.056
		(0.134)	(0.124)	(0.115)
PhD Top Institution		0.016	0.073	0.168
		(0.110)	(0.106)	(0.097)
Other Controls				
Reporting Method		Y	Y	Y
Solo Authored		Y	Y	Y
Share Female Authors		Y	Y	Y
Editor		Y	Y	Y
Field FE			Y	
Journal FE				Y
Articles	320	320	320	304

Notes: This table reports marginal effects from probit regressions (equation (2.4)). The dependent variable is a dummy that takes a value of one if a published article has a public working paper. No article weights applied.

Table B.25: Working Paper vs Published Version - 5% significance

	(1) ALL	(2) admin	(3) survey	(4) hand	(5) other
Published Version	-0.014	-0.047	0.079	-0.040	-0.031
	(0.026)	(0.054)	(0.061)	(0.036)	(0.042)
Constant	0.473	0.489	0.369	0.469	0.629
	(0.041)	(0.087)	(0.068)	(0.063)	(0.114)
Test Statistics	2,103	507	454	886	256
Articles	118	33	24	41	21
Window	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]	[1.96±0.50]

Notes: This table reports estimates from a linear regression. The dependent variable is a dummy that takes a value one if a given test statistic is significant at the 5% level (i.e. equal to 1.96). The independent variable of interest is a dummy that takes the value of one if a given test statistic is from the published version of an article. The sample is accordingly restricted to estimates from published articles that had an associated working paper. We apply no weights.

Table B.26: Working Paper vs Published Version - 10% significance

	(1) ALL	(2) admin	(3) survey	(4) hand	(5) other
Published Version	-0.023	-0.032	0.030	-0.062	0.019
	(0.028)	(0.054)	(0.057)	(0.044)	(0.041)
Constant	0.522	0.504	0.466	0.529	0.637
	(0.046)	(0.094)	(0.078)	(0.078)	(0.123)
Test Statistics	2,119	473	457	934	255
Articles	120	31	27	42	21
Window	[1.65±0.50]	[1.65±0.50]	[1.65±0.50]	[1.65±0.50]	[1.65±0.50]

Notes: This table reports estimates from a linear regression. The dependent variable is a dummy that takes a value one if a given test statistic is significant at the 1% level (i.e. equal to 1.65). The independent variable of interest is a dummy that takes the value of one if a given test statistic is from the published version of an article. The sample is accordingly restricted to estimates from published articles that had an associated working paper. We apply no weights.

Table B.27: Working Paper vs Published Version - 1% significance

	(1) ALL	(2) admin	(3) survey	(4) hand	(5) other
Published Version	0.010	0.031	-0.036	0.001	-0.053
	(0.026)	(0.034)	(0.039)	(0.034)	(0.079)
Constant	0.324	0.178	0.432	0.282	0.574
	(0.042)	(0.053)	(0.086)	(0.062)	(0.094)
Test Statistics	1,701	440	417	599	245
Articles	118	32	25	39	23
Window	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]	[2.58±0.50]

Notes: This table reports estimates from a linear regression. The dependent variable is a dummy that takes a value one if a given test statistic is significant at the 1% level (i.e. equal to 2.58). The independent variable of interest is a dummy that takes the value of one if a given test statistic is from the published version of an article. The sample is accordingly restricted to estimates from published articles that had an associated working paper. We apply no weights.

Appendix to Chapter 4

C.1 Appendix Figures and Tables

Figure C.1: Example of outside earnings public disclosure on website of the German federal parliament

Entgeltliche Tätigkeiten neben dem Mandat

Compamedia GmbH, Überlingen,
Vortrag, 2015, Stufe 3 (Deutscher Mittelstands-Summit)

CSA Celebrity Speakers GmbH, Düsseldorf,
Vortrag, 2015, Stufe 4 (AGRAVIS-Vortragsveranstaltung, AGRAVIS Raiffeisen AG, Münster)

Econ Referenten-Agentur, München,
Vortrag, 2014, Stufe 4 (Haspa-Branchen Treff „Wirtschaftsfaktor Russland", Hamburger Sparkasse AG, Hamburg)
Vortrag, 2016, Stufe 4 (Optimum Asset Management-Event 2016, Optimum Asset Management SA, Berlin)

Forum Executive AG, Zürich, Schweiz,
Vortrag, 2016, Stufe 4 (Funds Expert Forum)

GUILLOT Referenten-Kommunikation-Speakers Bureau, Ralingen,
Podiumsdiskussion, 2014, Stufe 4 (Das Freihandelsabkommen TTIP - Chance oder Schreckensvision für Europa, Deutscher Zigarettenverband e.V., Berlin)
Vortrag, 2015, Stufe 4 (Immobilien Investment Forum 2015, Savills Investment Management, Frankfurt/Main)
Vortrag, 2015, Stufe 4 (Tacheles 2015 - Das Investmentgespräch, Drescher & Cie Gesellschaft für Wirtschafts- und Finanzinformation mbH, St. Augustin)

Hoffmann & Campe Verlag GmbH, Hamburg,
Publizistische Tätigkeit, 2014, Stufe 8; 2015, Stufe 8
Vortrag, 2015, Stufe 3 (Lesereise)
Vortrag, 2016, Stufe 2 (Lesereise)

iGZ - Interessenverband Deutscher Zeitarbeitsunternehmen e.V., Berlin,
Vortrag, 2015, Stufe 4 (iGZ-Bundeskongress)

Internationales Steuerseminar Schweiz (ISt S), Zürich, Schweiz,
Vortrag, 2016, Stufe 3 (Internationales Steuerseminar 2016)

marcus evans Germany Ltd., Berlin,
Vortrag, 2015, Stufe 4 (9. CMO-Gipfel)
Vortrag, 2016, Stufe 4 (9. CEO-Gipfel)

MMM-Club (Moderne Markt-Methoden) e.V., Wettenberg,
Vortrag, 2016, Stufe 4 (54. MMM-Kongress)

pilot München GmbH, München,
Podiumsdiskussion, 2016, Stufe 4 (pilot Business-Lounge: „Zukunft gestalten")

Schweizerisches Institut für Auslandforschung (SIAF), Zürich,
Vortrag, 2014, Stufe 3 (Veranstaltungsreihe „Die Zukunft der Demokratie")

The London Speaker Bureau Germany, Karlsruhe,
Vortrag, 2015, Stufe 4 (UniCredit Wirtschaftsgespräch)
Vortrag, 2015, Stufe 4 (beim Industriebeirat der Triton Beratungsgesellschaft GmbH, Frankfurt/Main)

Vodafone Institute for Society and Communications GmbH, Berlin,
Vortrag, 2016, Stufe 2 (Veranstaltungsreihe „AusZeit")

WBMG - Unternehmensberatung GmbH, Landshut,
Beratung, 2014, Stufe 5; 2015, Stufe 7; 2016, Stufe 6

Zeitverlag Gerd Bucerius GmbH & Co. KG, Hamburg,
Publizistische Tätigkeit, 2014, Stufe 1

Funktionen in Unternehmen

Borussia Dortmund GmbH & Co. KGaA, Dortmund,
Mitglied des Aufsichtsrates, 2015, Stufe 4

ThyssenKrupp AG, Essen,
Mitglied des Aufsichtsrates (bis 31.12.2012), 2014, Stufe 3 (für 2012)

Funktionen in Vereinen, Verbänden und Stiftungen

Deutsche Nationalstiftung, Hamburg,
Mitglied des Senats

Helmut und Loki Schmidt-Stiftung, Hamburg,
Mitglied des Kuratoriums

Stiftung Berliner Schloss - Humboldtforum, Berlin,
Mitglied des Kuratoriums

ZEIT-Stiftung Ebelin und Gerd Bucerius, Hamburg,
Mitglied des Kuratoriums, jährlich, Stufe 3

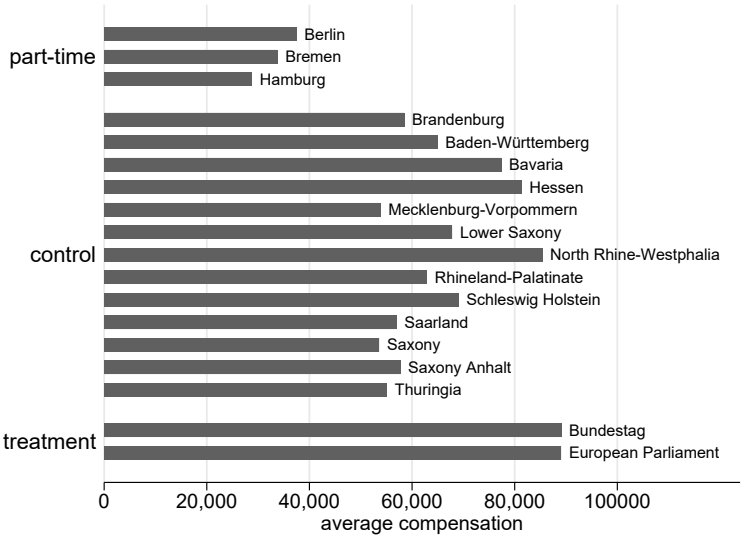
Notes: This figure is a screen shot of Peer Steinbrück's published outside earnings in election period 18. Source: Website of the Bundestag `https://www.bundestag.de/abgeordnete/biografien18/S/steinbrueck_peer/259022`

Table C.1: Public disclosure rules and measures of reported outside earnings

	election period 16 & 17				election period 18			
level	from	to	baseline	lower bound	from	to	baseline	lower bound
0	0	1,000	500	500	0	1,000	500	500
1	1,000	3,500	2,250	2,250	1,000	3,500	2,250	2,250
2	3,500	7,000	5,250	5,250	3,500	7,000	5,250	5,250
3	7,000		9,500	7,000	7,000	15,000	9,500	7,000
4					15,000	30,000	9,500	7,000
5					30,000	50,000	9,500	7,000
6					50,000	75,000	9,500	7,000
7					75,000	100,000	9,500	7,000
8					100,000	150,000	9,500	7,000
9					150,000	250,000	9,500	7,000
10					250,000		9,500	7,000

Notes: All values are in Euros. Public disclosure rules for election period 16, 17 and 18 as well as our two different measures that are used in the reported data. See Section 3.3.2 for details of the construction of the baseline and lower bound measures.

Figure C.2: Average compensation of MPs in each parliament



Notes: This figure plots the distribution of average compensation for a MP in each parliament (federal, state or EU). These values refer to the average for the years 2001 to 2014.

Table C.2: Details of election periods in federal parliament

	Election Period 16	Election Period 17	Election Period 18
Election Details			
election date	18.09.2005	27.09.2009	22.09.2013
duration	18.10.2005 - 27.10.2009	27.10.2009 - 22.10.2013	22.10.2013 - 24.10.2017
seats	614	622	631
Party			
CDU/CSU	226	239	311
SPD	222	146	193
FDP	61	93	0
The Left	54	76	64
Greens	51	68	63

Notes: This table consists of information of each election period in federal parliament under study.

Figure C.3: Reporting brackets

Notes: This figures visualizes the imputed values for each bracket. The blue dots are the average value for the respective bracket through which we fit a polynomial (blue dotted line). We then extrapolate the polynomial function to the highest bracket and impute the predicted value (orange dots).

Table C.3: Descriptive statistics: demographics (reported data)

variable	**mean**	**sd**	**N**
female	0.34	0.47	1952
age below 50	0.39	0.49	1952
age between 50 and 60	0.37	0.48	1952
age 60 and above	0.24	0.43	1952
East Germany	0.17	0.37	1952
married	0.72	0.45	1952
# children	1.60	1.37	1952
title: doctor	0.19	0.39	1952
title: professor	0.01	0.09	1952
occupation: other	0.32	0.47	1952
occupation: lawyer	0.19	0.39	1952
occupation: economist/MBA	0.16	0.36	1952
occupation: farmer	0.03	0.16	1952
occupation: teacher	0.09	0.28	1952
occupation: civil servant	0.02	0.15	1952
occupation: doctor	0.02	0.12	1952
occupation: journalist	0.03	0.16	1952
occupation: academic	0.08	0.28	1952
occupation: self-employed	0.07	0.26	1952
party: left-wing	0.50	0.50	1952
party: CDU/CSU	0.41	0.49	1952
party: SPD	0.30	0.46	1952
party: Greens	0.10	0.30	1952
party: The Left	0.10	0.30	1952
party: FDP	0.08	0.28	1952
terms: newcomer	0.31	0.46	1952
terms: 2 - 3	0.38	0.49	1952
terms: > 3	0.30	0.46	1952
early dropout	0.03	0.18	1952
late entry	0.04	0.20	1952

Source: Reported data for election periods 16, 17 and 18.

Table C.4: Descriptive statistics: political and electoral variables (reported data)

variable	mean	sd	N
entry: direct ballot	0.46	0.50	1952
entry: list ranking	10.55	12.31	1733
vote margin: candidate	6.78	16.51	1866
vote margin: party	12.09	10.11	1866
leadership	0.11	0.32	1952
committee chair	0.07	0.25	1952
committee: interior	0.06	0.23	1952
committee: digital	0.01	0.09	1952
committee: social	0.06	0.23	1952
committee: family	0.05	0.22	1952
committee: health	0.05	0.22	1952
committee: culture	0.03	0.18	1952
committee: human rights	0.03	0.16	1952
committee: justice	0.05	0.23	1952
committee: environment	0.05	0.22	1952
committee: election	0.03	0.16	1952
committee: development	0.03	0.18	1952
committee: exterior	0.06	0.23	1952
committee: budget	0.10	0.29	1952
committee: petition	0.04	0.19	1952
committee: accounting	0.02	0.13	1952
committee: sports	0.03	0.16	1952
committee: agriculture	0.05	0.22	1952
committee: tourism	0.03	0.16	1952
committee: traffic	0.06	0.24	1952
committee: defense	0.05	0.22	1952
committee: economics	0.06	0.24	1952
committee: science	0.05	0.22	1952
committee: EU	0.05	0.22	1952

Source: Reported data for election periods 16, 17 and 18.

Table C.5: Composition of outside activities across all activities (reported data)

	EP 16		EP 17		EP 18		**Total**	
	N	in %	N	in %	N	in %	N	in %
remunerated activity	1,335	32.7	1,469	32.45	1,621	34.15	**4,425**	**33.13**
type of activity[1]								
law	654	51.01	763	52.26	718	44.35	2,135	48.96
speech	353	27.54	362	24.79	320	19.77	1,035	23.73
management and consulting	183	14.7	183	12.53	177	10.93	543	12.45
other	92	7.18	152	10.41	404	24.95	648	14.86
functions in enterprises	512	12.54	442	9.76	630	13.27	**1,584**	**11.86**
type of function[2]								
public office	0	0	6	1.36	15	2.39	21	1.33
consult	175	34.38	150	33.94	166	26.43	491	31.10
control	287	56.39	252	57.01	385	61.31	924	58.52
lead	47	9.23	34	7.69	62	9.06	143	9.06
type of membership								
regular Member	428	84.09	370	83.71	535	85.19	1,333	84.42
chairman	81	15.91	72	16.29	93	14.81	246	15.58
functions in public corporations	670	16.41	695	15.35	837	17.63	**2,202**	**16.49**
type of function[2]								
public office	347	51.95	372	53.53	419	50.12	1,138	51.75
consult	151	22.6	147	21.15	169	20.22	467	21.24
control	112	16.77	106	15.25	187	22.37	405	18.42
lead	55	8.23	68	9.78	60	7.18	183	8.32
type of membership								
regular Member	601	89.97	623	89.64	753	90.07	1,977	89.9
chairman	67	10.03	72	10.36	83	9.93	222	10.10
functions in clubs	1,447	35.44	1,786	39.45	1,544	32.53	**4,777**	**35.76**
shareholdings in private corporations	119	2.91	135	2.98	115	2.42	**369**	**2.76**
Total	**4.083**	**100**	**4.527**	**100**	**4747**	**100**	**13.357**	**100**

Notes: This table provides an overview about the composition of outside activities. We consider every single activity. Activities are reported such that they belong to category: remunerated activity, functions in enterprises, functions in public corporations, functions in clubs or shareholdings in private corporations. We broadly categorize remunerated activities into law, speech, management and consulting and other. We classify type of function if a respective activity belongs to public office, consulting, control or lead. Notes: For 1.44% of all remunerated activities and for 0.32% of all functions in enterprises, no information about the type of activity is available. Functions in clubs are often voluntary work. The information 'voluntary' is optional and added in more than 85% of all functions in clubs. In some cases, the name of clients are not revealed due to existence of lawyer-client-confidentiality. We ignore the information of occupational activities pre-dating membership (e.g. lawyer). Shareholdings in private corporations need to be reported if a MP holds more than 25%.

[1] We classify remunerated activities as follows: (a) law (e.g. lawyer, judge), (b) speech (e.g. speech, publishing books), (c) management and consulting (e.g. business consultant, notary, manager) and (d) other (e.g. farmer, doctor).

[2] We classify type of function as follows: (a) public office (e.g. position in local politics/ church), (b) consult (e.g. advisory board), (c) control (e.g. supervisory board) and (d) lead (e.g. committee, management board, board of trustees).

Table C.6: Composition of outside activities per MP (reported data)

	EP 16		EP 17		EP 18		**Total**	
	N	in %	N	in %	N	in %	N	in %
remunerated activity	195	31.76	219	35.21	195	30.90	**609**	**32.62**
type of activity								
law	61	9.93	65	10.45	58	9.19	184	9.86
speech	52	8.57	55	8.84	45	7.13	152	8.14
management and consulting	61	9.93	70	11.25	57	9.03	188	10.07
other	37	6.03	42	6.75	38	6.02	117	6.27
functions in enterprises	240	39.09	223	35.85	291	46.12	**754**	**40.39**
type of function								
public office	0	0	4	0.64	11	1.74	15	0.80
consult	94	15.31	76	12.22	87	13.79	257	13.77
control	144	23.45	144	23.15	197	31.22	197	25.98
lead	25	4.07	24	3.86	32	5.07	81	4.34
type of membership								
regular member	216	35.18	201	32.32	269	42.63	686	36.74
chairman	33	5.37	34	5.47	42	6.66	109	5.84
functions in public corporations	359	58.47	357	57.40	385	61.01	**1,001**	**58.97**
type of function								
public office	226	36.81	247	39.71	264	41.84	737	39.48
consult	95	15.47	90	14.47	89	14.10	274	14.68
control	70	11.40	71	11.41	104	16.48	245	13.12
lead	35	5.70	39	6.27	31	4.91	105	5.62
type of membership								
regular member	339	55.21	341	54.82	372	58.95	372	56.35
chairman	37	6.03	37	5.95	35	5.55	109	5.84
functions in clubs	437	71.17	469	75.40	446	70.68	**1,352**	**72.42**
shareholdings in private corporations	69	11.24	76	12.22	67	10.62	**212**	**11.36**
Total # MPs	**614**		**622**		**631**			

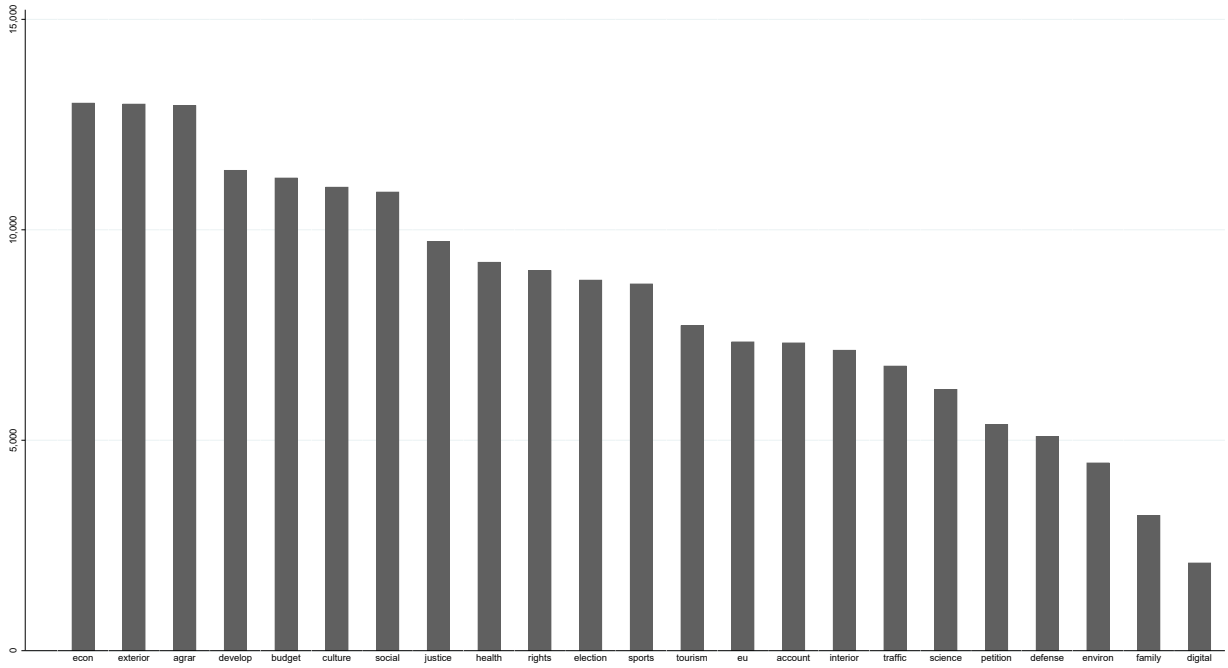
Notes: This table provides an overview about the composition of outside activities per MP, meaning how many MPs pursue a certain activity. The percentages define the share of MPs who persue a certain activity. For example, 32.62% of all MPs report a remunerated activity, while 58.97% of all MPs hold a function in a club. Activities are reported such that they belong to one of the following categories: remunerated activity, functions in enterprises, functions in public corporations, functions in clubs or shareholdings in private corporations. We broadly categorize remunerated activities into (a) law (e.g. lawyer, judge), (b) speech (e.g. speech, publishing books), (c) management and consulting (e.g. business consultant, notary, manager) and (d) other (e.g. farmer, doctor). We classify the type of function into (a) public office (e.g. position in local politics/ church), (b) consult (e.g. advisory board), (c) control (e.g. supervisory board) and (d) lead (e.g. committee, management board, board of trustees). For 1.44% of all remunerated activities and for 0.32% of all functions in enterprises, no information about the type of activity is available. Functions in clubs are often voluntary work. The information 'voluntary' is optional and added in more than 85% of all functions in clubs. In some cases, the name of clients are not revealed due to existence of lawyer-client-confidentiality. We ignore the information of occupational activities pre-dating membership (e.g. lawyer). Shareholdings in private corporations need to be reported if a MP holds more than 25%.

Table C.7: Distribution of levels and frequency by activity (reported data)

	EP 16		EP 17		EP 18		Total	
	N	**in %**	**N**	**in %**	**N**	**in %**	**N**	**in %**
Level								
0	1314	48	1395	48	1745	54	4454	50
1	696	26	780	27	721	22	2197	25
2	206	8	218	8	226	7	650	7
3 and higher	**497**	**18**	**512**	**18**	**519**	**18**	**1528**	**18**
3	497	18	512	18	235	7	1244	14
4					115	4	115	1
5					52	2	52	1
6					31	1	31	0
7					18	1	18	0
8					23	1	23	0
9					21	1	21	0
10					24	1	24	0
Frequency								
once	2559	94	2721	94	3032	94	8312	94
yearly	67	2	59	2	53	2	179	2
monthly	86	3	126	4	129	4	341	4

Notes: Levels and frequencies are reported for the following categories of activities: remunerated activities, functions in enterprises, functions in public corporations and functions in clubs. For functions in clubs, MPs can optionally indicate whether is is voluntary work or not. Source: Reported Data, own calculations.

Figure C.4: Outside earnings by committee membership



Notes: This graphs displays the average outside earnings as defined in Section 3.3.2 for each committee in the German federal parliament.
Source: Reported Data EP 16 - 18

Table C.8: Average number of MPs in federal and state parliaments

	number of MPs	election years
Treatment Group	**722**	
Federal Parliament	623	2002, 2005, 2009, 2013
Control Group	**776**	
Baden Württemberg	134	2001, 2006, 2011
Mecklenburg-Vorpommern	71	2002, 2006, 2011
North Rhine Westphalia	210	2005, 2010, 2012
Rhineland-Palatinate	101	2001, 2006, 2011
Schleswig-Holstein	83	2005, 2009, 2012
Saarland	51	2004, 2009, 2012
Saxony	126	2004, 2009, 2014
Control Group ***(excluded in 2013 & 2014)***	**187**	
Bavaria	187	2003, 2008, 2013
Control Group ***(excluded in 2014)***	**557**	
Hessia	112	2003, 2008, 2013
Lower Saxony	163	2003, 2009, 2013
Brandenburg	88	2004, 2009, 2014
Saxony-Anhalt	106	2002, 2006, 2011
Thuringia	88	2004, 2009, 2014
Part-time parliament ***(excluded in all years)***	**352**	
Berlin	146	2001, 2006, 2011
Bremen	85	2003, 2007, 2011
Hamburg	121	2001, 2004, 2008, 2011

Notes: This table consists of information of each parliament under study. The number denotes the average number of MPs in each parliament for the years 2001 to 2014. Germany consists of 16 states (*Länder*). We entirely exclude Berlin, Bremen and Hamburg from our analysis (part-time Parliament (*Feierabendparliament*)). Bavaria, Hessen, Lower Saxony, Brandenburg, Saxony-Anhalt, and Thuringia introduced public disclosure rules in 2013/2014 and are excluded from our sample for these years.

Table C.9: Tightening of the disclosure law: channels (lower bound)

	(1) EP 16 outside earnings	(2) EP 17 outside earnings	(3) EP 18 outside earnings
Panel A: directly elected			
D^{direct}	-7,870*	-5,108	-12,328***
	(4,697)	(8,512)	(4,488)
electoral district FE	Yes	Yes	Yes
controls	Yes	Yes	Yes
N	318	238	404
# politicians	318	238	404
Panel B: unsafe party rank			
$D^{unsaferank}$	-2,466	-417	-4,996**
	(2,130)	(3,473)	(2,044)
party-state FE	Yes	Yes	Yes
controls	Yes	Yes	Yes
N	562	578	593
# politicians	562	578	593

Notes: The outcome variable is outside earnings as described in Section 3.3.2. In Panel A, the sample contains only MPs from districts, where both the first- and second-placed candidate entered parliament to estimate equation 3.4. In Panel B, we use only MPs that were ranked on a party list to estimate equation 3.5. Controls refer to all variables in Tables C.3 and C.4. Robust standard errors. * p<0.1, ** p<0.05, *** p<0.01 Source: reported data EP 16 - 18

APPENDIX C. APPENDIX: DISCLOSURE

Bibliography

Abadie, A. (2020), 'Statistical Nonsignificance in Empirical Economics', *American Economic Review: Insights* **2**(2), 193–208.

Akcigit, U., Baslandze, S. and Lotti, F. (2018), 'Connecting to power: Political connections, innovation, and firm dynamics', *National Bureau of Economic Research Working Paper 25136* .

Alvaredo, F., Atkinson, A. B., Piketty, T. and Saez, E. (2013), 'The top 1 percent in international and historical perspective', *Journal of Economic perspectives* **27**(3), 3–20.

Andrews, I. and Kasy, M. (2019), 'Identification of and Correction for Publication Bias', *American Economic Review* **109**(8), 2766–94.

Angrist, J. D. and Pischke, J.-S. (2010), 'The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics', *Journal of Economic Perspectives* **24**(2), 3–30.

Arnold, F., Kauder, B. and Potrafke, N. (2014), 'Outside earnings, absence, and activity: Evidence from German parliamentarians', *European Journal of Political Economy* **36**, 147–157.

Ashenfelter, O., Harmon, C. and Oosterbeek, H. (1999), 'A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias', *Labour Economics* **6**(4), 453–470.

Auten, G. and Carroll, R. (1999), 'The Effect of Income Taxes on Household Income', *Review of Economics and Statistics* **81**(4), 681–693.

Bakos, P., Benczúr, P. and Benedek, D. (2010), 'The elasticity of taxable income: Estimates and flat tax predictions using the hungarian tax changes in 2005', *Working Paper* .

Becker, J., Peichl, A. and Rincke, J. (2009), 'Politicians' outside earnings and electoral competition', *Public Choice* **140**(3), 379–394.

Blanco-Perez, C. and Brodeur, A. (2019), 'Transparency in Empirical Economic Research', *IZA World of Labor* p. 467.

Blanco-Perez, C. and Brodeur, A. (2020), 'Publication Bias and Editorial Statement on Negative Findings', *Economic Journal* **130**(629), 1226–1247.

Bø, E., Slemrod, J. and Thoresen, T. O. (2015), 'Taxes on the internet: Deterrence effects of public disclosure', *American Economic Journal: Economic Policy* **7**(1), 36–62.

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A. and Olds, J. L. (2015), 'Social, behavioral, and economic sciences perspectives on robust and reliable science'. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, National Science Foundation.

Brodeur, A., Cook, N. and Heyes, A. (2020), 'Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics', *American Economic Review* **110**(11), 3634–60.

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y. (2016*a*), 'Star wars: The empirics strike back', *American Economic Journal: Applied Economics* **8**(1), 1–32.

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y. (2016*b*), 'Star Wars: The Empirics Strike Back', *American Economic Journal: Applied Economics* **8**(1), 1–32.

Bruns, S. B., Asanov, I., Bode, R., Dunger, M. et al. (2019), 'Reporting Errors and Biases in Published Empirical Findings: Evidence from Innovation Research', *Research Policy* **48**(9), 103796.

Bundestag (2013), 'Code of conduct for members of the German Bundestag', `https://www.bundestag.de/resource/blob/195006/a1232d4a394f7cdee1b9bccc2f374880/code_of_conduct-data.pdf`.

Burlig, F. (2018), 'Improving transparency in observational social science research: A pre-analysis plan approach', *Economics Letters* **168**, 56 – 60.

Burns, S. K. and Ziliak, J. P. (2017), 'Identifying the elasticity of taxable income', *The Economic Journal* **127**, 297–329.

Bursztyn, L., González, A. L. and Yanagizawa-Drott, D. (2020), 'Misperceived social norms: Women working outside the home in Saudi Arabia', *American Economic Review* **110**(10), 2997–3029.

Bursztyn, L. and Jensen, R. (2017), 'Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure', *Annual Review of Economics* **9**, 131–153.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T. et al. (2016), 'Evaluating replicability of laboratory experiments in economics', *Science* **351**(6280), 1433–1436.

Campbell, R. and Cowley, P. (2015), 'Attitudes to moonlighting politicians: Evidence from the United Kingdom', *Journal of Experimental Political Science* **2**(1), 63–72.

Card, D. and DellaVigna, S. (2020), 'What do Editors Maximize? Evidence from Four Economics Journals', *Review of Economics and Statistics* **102**(1), 195–217.

Card, D., DellaVigna, S., Funk, P. and Iriberri, N. (2020), 'Are Referees and Editors in Economics Gender Neutral?', *Quarterly Journal of Economics* **135**(1), 269–327.

Card, D., Kluve, J. and Weber, A. (2010), 'Active labour market policy evaluations: A meta-analysis', *The Economic Journal* **120**(548), F452–F477.

Card, D., Kluve, J. and Weber, A. (2018), 'What works? a meta analysis of recent active labor market program evaluations', *Journal of the European Economic Association* **16**(3), 894–931.

Card, D. and Krueger, A. B. (1995), 'Time-series minimum-wage studies: a meta-analysis', *American Economic Review* **85**(2), 238–243.

Carrell, S., Figlio, D. and Lusher, L. (2020), 'Clubs and Networks in Economics Reviewing'. working paper.

Chang, A. C. and Li, P. (Forthcoming), 'Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Often Not"', *Critical Finance Review* .

Chetty, R. (2009), 'Is the Taxable Income Elasticity Sufficient to Calculate Deadweight Loss? The Implications of Evasion and Avoidance', *American Economic Journal: Economic Policy* **1**(2), 31–52.

Chetty, R. (2012), 'Time Trends in the Use of Administrative Data for Empirical Research'. Cambridge, MA. NBER Summer Institute, July 2012.

Christensen, G., Dafoe, A., Miguel, E., Moore, D. A. and Rose, A. K. (2019), 'A Study of the Impact of Data Sharing on Article Citations Using Journal Policies as a Natural Experiment', *PloS one* **14**(12), e0225883.

Christensen, G. and Miguel, E. (2018), 'Transparency, reproducibility, and the credibility of economics research', *Journal of Economic Literature* **56**(3), 920–80.

DellaVigna, S. and Linos, E. (2020), 'RCTs to Scale: Comprehensive Evidence from Two Nudge Units'. UC Berkeley Mimeo.

Diamond, P. A. (1998), 'Optimal Income Taxation: An Example with a U-Shaped Pattern of Optimal Marginal Tax Rates', *American Economic Review* **88**(1), 83–95.

Djankov, S., La Porta, R., Lopez de Silanes, F. and Shleifer, A. (2010), 'Disclosure by politicians', *American Economic Journal: Applied Economics* **2**(2), 179–209.

Doucouliagos, C. and Stanley, T. D. (2013), 'Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity', *Journal of Economic Surveys* **27**(2), 316–339.

Dwenger, N. and Treber, L. (2018), 'Shaming for tax enforcement: Evidence from a new policy', *Working Paper* .

Egger, M., Smith, G. D., Schneider, M. and Minder, C. (1997), 'Bias in meta-analysis detected by a simple, graphical test', *Bmj* **315**(7109), 629–634.

Eggers, A. C. and Hainmueller, J. (2009), 'MPs for sale? Returns to office in postwar British politics', *American Political Science Review* **103**(4), 513–533.

Feige, E. L. (1975), 'The Consequences of Journal Editorial Policies and a Suggestion for Revision', *Journal of Political Economy* **83**(6), 1291–1296.

Feldstein, M. (1995), 'The Effect of Marginal Tax Rates on Taxable Income: A Panel Study of the 1986 Tax Reform Act', *Journal of Political Economy* **103**(3), 551–572.

Feldstein, M. (1999), 'Tax Avoidance and the Deadweight Loss of the Income Tax', *Review of Economics and Statistics* **81**(4), 674–680.

Firpo, S., Fortin, N. M. and Lemieux, T. (2009), 'Unconditional quantile regressions', *Econometrica* **77**(3), 953–973.

Fisman, R., Schulz, F. and Vig, V. (2021), 'Financial disclosure and political selection: Evidence from India', *Working Paper* .

Franco, A., Malhotra, N. and Simonovits, G. (2014), 'Publication Bias in the Social Sciences: Unlocking the File Drawer', *Science* **345**(6203), 1502–1505.

Furukawa, C. (2019), 'Publication Bias under Aggregation Frictions: From Communication Model to New Correction Method'. MIT Mimeo.

Gagliarducci, S., Nannicini, T. and Naticchioni, P. (2010), 'Moonlighting politicians', *Journal of Public Economics* **94**(9), 688–699.

Gerber, A. and Malhotra, N. (2008*a*), 'Do Statistical Reporting Standards Affect what is Published? Publication Bias in Two Leading Political Science Journals', *Quarterly Journal of Political Science* **3**(3), 313–326.

Gerber, A. S. and Malhotra, N. (2008*b*), 'Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?', *Sociological Methods & Research* **37**(1), 3–30.

Giertz, S. H. (2010), 'Panel Data Techniques and the Elasticity of Taxable Income', *National Tax Association 2009 Annual Conference Proceedings* .

Gneezy, U. and Rustichini, A. (2000), 'A fine is a price', *The Journal of Legal Studies* **29**(1), 1–17.

Goodman-Bacon, A. (2021), 'Difference-in-differences with variation in treatment timing', *Journal of Econometrics* (forthcoming).

Gruber, J. and Saez, E. (2002), 'The Elasticity of Taxable Income: Evidence and Implications', *Journal of Public Economics* **84**(1), 1–32.

Hamermesh, D. S. (2017), 'Replication in Labor Economics: Evidence from Data, and What it Suggests', *American Economic Review* **107**(5), 37–40.

Hargaden, E. P. (2020), 'Taxpayer responses in good times and bad', *Journal of Economic Behavior & Organization* **176**, 653–690.
URL: *https://www.sciencedirect.com/science/article/pii/S0167268120301487*

Harju, J. and Matikka, T. (2016), 'The elasticity of taxable income and income-shifting: what is "real"and what is not?', *International Tax and Public Finance* **23**(4), 640–669.

Havránek, T. (2015*a*), 'Measuring intertemporal substitution: The importance of method choices and selective reporting', *Journal of the European Economic Association* **13**(6), 1180–1204.

Havránek, T. (2015*b*), 'Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting', *Journal of the European Economic Association* **13**(6), 1180–1204.

Havránek, T. and Sokolova, A. (2020), 'Do Consumers Really Follow a Rule of Thumb? Three Thousand Estimates from 144 Studies Say Probably Not', *Review of Economic Dynamics* **35**, 97–122.

Havránek, T., Stanley, T., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W. R., Rost, K. and Van Aert, R. (2020), 'Reporting Guidelines for Meta-Analysis in Economics', *Journal of Economic Surveys* **34**(3), 469–475.

Höffler, J. H. (2017), 'Replication and Economics Journal Policies', *American Economic Review: Papers and Proceedings* **107**(5), 52–55.

Ioannidis, J., Stanley, T. D. and Doucouliagos, H. (2017), 'The power of bias in economics research', *The Economic Journal* **127**(605).

Kapteyn, A. and Ypma, J. Y. (2007), 'Measurement Error and Misclassification: A Comparison of Survey and Administrative Data', *Journal of Labor Economics* **25**(3), 513–551.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S. and Saez, E. (2011), 'Unwilling or Unable to Cheat? Evidence from a Randomized Tax Audit Experiment in Denmark', *Econometrica* **79**(3), 651–692.

Kleven, H. J., Kreiner, C. T. and Saez, E. (2016*a*), 'Why can modern governments tax so much? an agency model of firms as fiscal intermediaries', *Economica* **83**(330), 219–246.

Kleven, H. J., Kreiner, C. T. and Saez, E. (2016*b*), 'Why can modern governments tax so much? An agency model of firms as fiscal intermediaries', *Economica* **83**(330), 219–246.

Kleven, H. J. and Schultz, E. A. (2014), 'Estimating Taxable Income Responses Using Danish Tax Reforms', *American Economic Journal: Economic Policy* **6**(4), 271–301.

Kopczuk, W. (2005), 'Tax Bases, Tax Rates and the Elasticity of Reported Income', *Journal of Public Economics* **89**(11), 2093–2119.

Kreiner, C. T., Leth-Petersen, S. and Skov, P. E. (2016), 'Tax Reforms and Intertemporal Shifting of Wage Income: Evidence from Danish Monthly Payroll Records', *American Economic Journal: Economic Policy* **8**(3), 233–57.

Künn, S. (2015), 'The Challenges of Linking Survey and Administrative Data', *IZA World of Labor* .

Le Maire, D. and Schjerning, B. (2013), 'Tax bunching, income shifting and self-employment', *Journal of Public Economics* **107**, 1–18.

Lichter, A., Peichl, A. and Siegloch, S. (2015), 'The own-wage elasticity of labor demand: A meta-regression analysis', *European Economic Review* **80**, 94–119.

Malik, R. (2020), 'Transparency, elections, and Pakistani politicians' tax compliance', *Comparative Political Studies* **53**(7), 1060–1091.

Maniadis, Z., Tufano, F. and List, J. A. (2017), 'To Replicate or not to Replicate? Exploring Reproducibility in Economics Through the Lens of a Model and a Pilot Study', *Economic Journal* **127**(605).

Mas, A. (2016), 'Does disclosure affect CEO pay setting? Evidence from the passage of the 1934 Securities and Exchange Act', *Working Paper* .

Mas, A. (2017), 'Does transparency lead to pay compression?', *Journal of Political Economy* **125**(5), 1683–1721.

McCullough, B., McGeary, K. A. and Harrison, T. D. (2008), 'Do Economics Journal Archives Promote Replicable Research?', *Canadian Journal of Economics* **41**(4), 1406–1420.

Merritt, A. C., Effron, D. A. and Monin, B. (2010), 'Moral self-licensing: When being good frees us to be bad', *Social and Personality Psychology Compass* **4**(5), 344–357.

Mertens, K. and Montiel Olea, J. L. (2018), 'Marginal Tax Rates and Income: New Time Series Evidence', *The Quarterly Journal of Economics* **133**(4), 1803–1884.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D., Humphreys, M., Imbens, G. et al. (2014), 'Promoting Transparency in Social Science Research', *Science* **343**(6166), 30–31.

Mirrlees, J. A. (1971), 'An Exploration in the Theory of Optimum Income Taxation', *The Review of Economic Studies* **38**(2), 175–208.

Moffitt, R. and Wilhelm, M. O. (2000), 'Taxation and the labor supply decisions of the affluent', *Does Atlas Shrug?: The Economic Consequences of Taxing the Rich* p. 193.

Mueller-Langer, F., Fecher, B., Harhoff, D. and Wagner, G. G. (2019), 'Replication studies in economics: How many and which papers are chosen for replication, and why?', *Research Policy* **48**(1), 62–83.

Perez-Truglia, R. and Troiano, U. (2018), 'Shaming Tax Delinquents', *Journal of Public Economics* **167**, 120–137.

Piketty, T. and Saez, E. (2013), 'Optimal Labor Income Taxation', *Handbook of Public Economics* **5**, 391–474.

Romer, C. D. and Romer, D. H. (2010), 'The macroeconomic effects of tax changes: estimates based on a new measure of fiscal shocks', *American Economic Review* **100**(3), 763–801.

Rosenthal, R. (1979*a*), 'The File Drawer Problem and Tolerance for Null Results.', *Psychological Bulletin* **86**(3), 638.

Rosenthal, R. (1979*b*), 'The File Drawer Problem and Tolerance for Null Results', *Psychological Bulletin* **86**, 638.

Saez, E. (2001), 'Using Elasticities to Derive Optimal Income Tax Rates', *Review of Economic Studies* **68**(1), 205–229.

Saez, E. (2010), 'Do Taxpayers Bunch at Kink Points?', *American Economic Journal: Economic Policy* **2**(3), 180–212.

Saez, E. (2017), 'Taxing the rich more: Preliminary evidence from the 2013 tax increase', *Tax Policy and the Economy* **31**(1), 71–120.

Saez, E., Slemrod, J. and Giertz, S. H. (2012), 'The Elasticity of Taxable Incomewith Respect to Marginal Tax Rates: A Critical Review', *Journal of Economic Literature* **50**(1), 3–50.

Slemrod, J. (2016), 'Caveats to the Research Use of Tax-Return Administrative Data', *National Tax Journal* **69**(4), 1003.

Slemrod, J. and Kopczuk, W. (2002), 'The Optimal Elasticity of Taxable Income', *Journal of Public Economics* **84**(1), 91–112.

Slemrod, J., Rehman, O. U. and Waseem, M. (2020), 'How do taxpayers respond to public disclosure and social recognition programs? Evidence from Pakistan', *The Review of Economics and Statistics* pp. 1–44.

Spiegel (2010), 'Parlamentsschwänzer mit Spitzenverdienst', `https://www.spiegel.de/politik/deutschland/ex-minister-steinbrueck-parlamentsschwaenzer-mit-spitzenverdienst-a-712225.html`.

Stanley, T. D. (2008), 'Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection', *Oxford Bulletin of Economics and Statistics* **70**(1), 103–127.

Stanley, T. D. and Doucouliagos, H. (2014), 'Meta-regression Approximations to Reduce Publication Selection Bias', *Research Synthesis Methods* **5**(1), 60–78.

Stanley, T., Doucouliagos, C., Giles, M., Heckemeyer, J., Johnston, R., Laroche, P., Nelson, J., Paldam, M., Poot, J., Pugh, G. et al. (2013), 'Meta Analysis of Economics Research Reporting Guidelines', *Journal of Economic Surveys* **27**(2), 390–394.

Stanley, T. and Doucouliagos, H. (2010), 'Picture this: a simple graph that reveals much ado about research', *Journal of Economic Surveys* **24**(1), 170–191.

Swanson, N., Christensen, G., Littman, R., Birke, D., Miguel, E., Paluck, E. L. and Wang, Z. (2020), Research Transparency Is on the Rise in Economics, *in* 'AEA Papers and Proceedings', Vol. 110, pp. 61–65.

Vilhuber, L. (2020), 'Reproducibility and Replicability in Economics', *Harvard Data Science Review* **2**(4).

Vivalt, E. (2019), 'Specification Searching and Significance Inflation Across Time, Methods and Disciplines', *Oxford Bulletin of Economics and Statistics* **81**(4), 797–816.

Weber, C. (2014), 'Toward Obtaining a Consistent Estimate of the Elasticity of Taxable Income Using Difference-in-Differences', *Journal of Public Economics* **117**, 90–103.

Carina Neisser

Current Position

[since 10.2020] Post-doc, University of Cologne and member of the ECONtribute: Markets & Public Policy Cluster of Excellence

Education

[since 04.2015] Ph.D. candidate, University of Mannheim, Germany

[08.2016—10.2016] Visiting Student Researcher, UC Berkeley

[08.2014] Master of Science in Economics, University of Cologne

[09.2010] Bachelor of Science in Economics, University of Bonn

[09.2009—06.2010] Toulouse School of Economics (Exchange Student)

Affiliations and Previous Academic Positions

[since 07.2016] Research Affiliate, IZA Institute of Labor Economics

[07.2014–10.2020] Researcher, Social Policy and Redistribution, ZEW - Leibniz Center for European Economic Research, Mannheim, Germany (Maternity Leave, 08.2018—06.2019)

[10.2013—07.2014] Student Research Assistant, Institute for the Study of Labor (IZA)

[11.2011—10.2013] Student Research Assistant, Cologne Institute for Economic Research

[06.2008—09.2009] Student Research Assistant, Max Planck Institute for Research on Collective Goods