Essays in Tax Policy Evaluation

by

Steven Hamilton

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Public Policy and Economics)
in the University of Michigan
2018

Doctoral Committee:

Professor Joel Slemrod, Chair
Professor Jim Hines
Professor Jeff Smith
Professor Justin Wolfers

Steven Hamilton

steveham@umich.edu

ORCID iD: 0000-0001-5660-1495

Dedication

To Bria.

Acknowledgments

Without the sacrifice and support of my wife, Bria, I'm not sure this would have been possible. And she did it all without complaint, and without ever having made me feel as though I owe her a debt of gratitude. But I really do.

I thank my mother, Karen. I've been able to achieve all of this because a mother put her son before herself.

I thank my committee, the four Js, Joel, Jim, Jeff, and Justin for their stewardship. Joel, you gave me so many opportunities throughout my six years at Michigan; that kind of generosity is very rare. Jim, you taught me so much of what I know of public finance, and you really helped shape the way I think about economics. Jeff, you were not only a great source of advice and a great mentor, but also a great friend. Justin, you taught me how to communicate, which surely is the most fundamental skill of an economist.

I thank my friends and colleagues, Ben, Jaron, Steve, and Traviss, who provided support and an outlet whenever I needed it.

I thank my mentor, Flavio, who started me on this path. From the very beginning, you motivated me to reach this point.

Last, but certainly not least, I thank the support staff at Michigan—Mim in the Ford School, and Julie and Laura in the Department of Economics, among others. Your hard work and dedication has never gone unnoticed by me.

Contents

List of Figures

List of Tables

List of Appendices

Abstract

This dissertation contains a collection of new theories, empirical methodologies, and data analyses for the evaluation of tax policy.

In the first chapter, I investigate the degree to which tax deductions respond to the tax rate, and the implications this has for tax policy. I define a new tax instrument, the 'deductibility rate', which specifies the proportion of eligible expenses a taxpayer may deduct when preparing her taxes. If the utilities of gross income and deductions are separable, then the deduction elasticity reflects the revenue leakage caused by greater deductibility. To identify this elasticity, I develop the first method to decompose bunching in taxable income into its constituent parts, exploiting the removal of a notch in the tax schedule. This setting also generates an observed counterfactual density, obviating the parametric assumptions routinely made in bunching studies. Applying this method to new administrative tax data from Australia, I find that while deductions account for just 5% of taxable income, they account for 35% of the response of taxable income to the tax rate. Based on an elasticity of taxable income of 0.06, the deduction elasticity is -0.45, and the gross-income elasticity is 0.04. Consistent with standard optimal-tax logic, the sensitivity of deductions to the tax rate suggests that restricting deductions could raise welfare.

In the second chapter, I develop a new empirical method for addressing a common problem afflicting regression discontinuity (RD) designs. This design has gained wide popularity for its perceived credibility in identifying treatment effects. But there are common settings in which the necessary assumptions of the method are not satisfied. When units manipulate their value of the running variable across the treatment threshold, this can distort the average outcome near the threshold, invalidating the RD design. Some settings in which the design would otherwise be appropriate offer a comparison group for which the treatment status does not change at the threshold. In such settings, it's common to observe variables, such as lags of the outcome, that predict the outcome in the absence of manipulation. I devise a nonparametric propensity-score reweighting method that exploits these variables to correct for manipulation bias. The method relies on 'manipulation-on-observed-variables' and common-support assumptions akin to those used in standard matching and weighting exercises.

In the third chapter, I develop a new theory of tax evasion under third-party report-

ing. When a tax authority requires reports from third parties about a taxable transaction, tax evasion is feasible only if the reporters underreport collusively. I develop a model of third-party reporting to investigate its limits as an enforcement tool. Under what conditions is a third-party reporting regime robust to collusion between reporters? The deterrence effect of third-party reporting increases with the number of reporters per transaction and with uncertainty about the other reports. Under certain conditions, third-party reporting can ensure full compliance when a common report is required across transactions. Compliance also improves with the number of related transactions in which there is underreporting, such that there is a maximum number of related transactions beyond which evasion is infeasible. These findings offer insights into the settings in which third-party reporting obligations are most effective in increasing compliance.

Chapter 1

Optimal deductibility: Theory, and evidence from a bunching decomposition

1.1 Introduction

The modern literature on the behavioural response to income taxes has focused almost exclusively on the bottom line of the tax return, taxable income finding it responds fairly modestly to a change in the tax rate. I develop a new method to decompose the taxable-income response associated with bunching near discontinuities in the tax schedule into the responses available, gross income and deductions, the two principal components of taxable income. Using new administrative tax records from Australia, I find that the effect of the tax rate on deductions is an order of magnitude larger than that on gross income. The logic that the optimal tax rate is inversely related to the behavioural response suggests the anatomy of the response can be informative to policy. Namely, the large observed response indicates that limiting the ability of taxpayers to claim deductions could raise welfare.

I formalise this logic in a simple Ramsey (1927) model of optimal deductibility, in which, instead of consuming commodities, the taxpayer reports gross income and deductions in her tax return. In addition to the tax rate, the government selects the proportion of expenses that are deductible (the deductibility rate). While the elasticity of taxable income (ETI) continues to be a sufficient statistic for the deadweight loss of the tax rate, it is not sufficient for the deadweight loss of the deductibility rate. Because the latter depends on the response of taxable income to the deductibility rate, the lack of observed variation in the deductibility rate presents a challenge for empirical work.

It is more common to observe variation in the tax rate, but this won't in general induce the same response as a change in the deductibility rate. Under quasilinear, isoelastic, and separable utility, which combines the functional form used in bunching studies with the assumption often made in Ramsey models that the cross-price elasticities are zero, the two changes have equivalent effects on deductions. In that case, the response of deductions to a change in the tax rate, in the form of the deduction

elasticity, is a sufficient statistic for the deadweight loss of the deductibility rate. The validity of this assumption depends on the substitutability or complementarity of gross income and deductions in practice. As in the standard Ramsey case, if they are complementary or substitutable, then the optimal deductibility rate will still depend inversely on the deduction response, but this will be either attenuated or accentuated by the gross-income response.

Bunching methods have been widely used to estimate the response of taxable income to a change in the tax rate, but they have not been used to decompose the response. I extend the standard bunching model to include gross income and deductions, and exploit the taxpayer's optimality conditions before and after bunching to derive a simple formula for the deduction elasticity, which depends on the ETI and the relative proportional changes of deductions and taxable income in bunching. To estimate the deduction elasticity using this formula, it is necessary to observe a set of taxpayers who face the discontinuity, so have an incentive to bunch, and a set who do not.

I consider a 16% sample of Australian administrative income tax records, which to date has seen little use by academic researchers. I study the effects of a provision under which an additional 1% tax is paid by taxpayers without dependents who do not have private health insurance coverage, and whose taxable income exceeds AU$50,000. This is a notch, in public-finance parlance, because tax liability jumps by $500 at the threshold. This generates a strong incentive for certain taxpayers to reduce their taxable incomes to below the threshold, which requires them to decrease gross income or increase deductions. In 2009, the government raised the threshold to $70,000, which generates a treatment group (those near the $50,000 threshold in 2008) and a comparison group (those in the same region in 2009) whose tax returns can be compared.

I identify the range of taxpayers affected by the policy (the manipulation region) by comparing the densities of the treated and nontreated groups. Considering all taxpayers in the manipulation region with and without the treatment avoids bias due to selection into bunching, because all those who bunch and who don't bunch are always included. A simple comparison of the treated and nontreated means is, however, subject to a different selection bias because, at a given taxable income, gross income and deductions vary year-to-year in the absence of treatment. To address this, I implement a difference-in-differences design, exploiting a placebo group in the region below and adjacent to the manipulation region. These taxpayers are comparable to the treated but never receive the treatment. I address the possibility of non-parallel pre-trends by predicting the counterfactual outcome in the nontreatment period based on apparently linear pre-trends, and use these in place of the observed nontreatment outcomes.

I find that, in absolute terms, the average response in deductions accounts for around a third ($187) of the average response in taxable income ($527), with gross income accounting for the remaining two thirds ($340). But because deductions constitute only

5% ($2,380) of taxable income in the absence of treatment ($50,535), their proportional response is an order of magnitude greater than that of gross income. For every 1% decrease in taxable income, deductions increase by 7.5%, while gross income decreases only by 0.6%. Given an estimated ETI of 0.06, this translates to a deduction elasticity of -0.45, and a gross-income elasticity of 0.04.

Based on the optimal-tax formulas I derive in the chapter, and given reasonable parameter values, each $1 of deductions would have to generate 68¢ in external benefits, over and above the benefits to the taxpayer claiming it, in order for it to be optimal to allow taxpayers to fully deduct their expenses. If the external benefit of deductions were even as high as 30¢, the optimal deductibility rate would be just 34%. These results are driven by the large observed response of deductions to a change in the tax rate. They reflect the standard logic of the Ramsey inverse-elasticity rule that goods with high elasticities should be taxed less.

This chapter informs three strands of the public-finance literature. The first is the literature on the behavioural response to income taxes, which, following Feldstein (1999), has come to focus on the ETI. Saez, Slemrod and Giertz (2012) offer a useful review of this literature. Slemrod (1998), Chetty (2009), and Doerrenberg, Peichl and Siegloch (2015) propose conditions under which the ETI is not a sufficient statistic for the welfare impact of a change in the tax rate. I follow them by proposing conditions under which the ETI is not a sufficient statistic for the optimal setting of a different tax instrument. In that regard, my work is related to that of Slemrod and Kopczuk (2002), who consider the ability of the government to set the optimal tax base as do I with the deductibility rate.

The second strand is the literature on empirical bunching methods. While the ETI literature historically used panel data to observe variation over time following tax reforms (Feldstein, 1995; Gruber and Saez, 2002), the efficacy of these methods has been questioned (Weber, 2014). More recently, scholars have relied on bunching to identify the ETI, which is seen to offer more credible identification (Saez, 2010; Chetty, Friedman, Olsen and Pistaferri, 2011; Kleven and Waseem, 2013). Kleven (2016) offers a useful review of this literature. I extend this literature by proposing a method to decompose the bunching response.

The last strand is the developing literature on deductions, to which my main results contribute. Doerrenberg et al. (2015) show that the response of deductions to the tax rate is a necessary statistic for the optimal tax rate when deductions generate external benefits, and apply the traditional panel-data methods to estimate the deduction elasticity in Germany. Based on a higher ETI of 0.6, they find a deduction elasticity of -0.9.[1] Also using German data, Schächtele (2016) notes that, when considering a measure of taxable income that excludes deductions, there is no bunching at a particular kink in

[1] In the present case, if the ETI were 0.6, then the measured deduction elasticity would be -4.52.

the German tax schedule, which suggests that those without deductions are unresponsive to the tax rate. Paetzold (2017) estimates the deduction elasticity by applying a regression-kink design to a change in the probability of claiming deductions at a discontinuity in the marginal tax rate in Austria. Given an ETI of 0.1, he finds a deduction elasticity of -0.6. A comparable decomposition for firms has been considered by Best, Brockmeyer, Kleven, Spinnewijn and Waseem (2015) and Bachas and Soto (2017). The methods I develop in this chapter are easily adaptable to the corporate taxation setting.

In section 1.2, I present a model of optimal deductibility, setting out the conditions under which the deduction elasticity is a sufficient statistic for the optimal deductibility rate. In section 1.3, I present a bunching decomposition method, which I later rely on to estimate the deduction elasticity. In section 1.4, I describe the Australian tax system, the policy that I consider, and the data on which I rely. In section 1.5, I estimate the deduction elasticity. In section 1.6, I conclude.

1.2 A Ramsey model of optimal deductibility

I develop a model of optimal deductibility to determine the conditions under which the observed response of deductions to a change in the tax rate, which I observe, is informative to policy. In doing so, I am motivated by two questions. How does the optimal deductibility rate depend on the responsiveness of deductions to the deductibility rate? And, as the deductibility rate seldom varies in practice, under what conditions can variation in the tax rate empirically be relied upon instead?

I apply the Ramsey (1927) model of optimal commodity taxation to a taxpayer who reports gross income and deductions in her tax return. I represent the tax return similarly to Feldstein (1999), but he assumes the government chooses only the tax rate applicable to taxable income. I allow the government to choose also the proportion of eligible expenditures that is deductible (the deductibility rate). The tax return can be disaggregated in many ways, but gross income and deductions are of interest to a developing literature.[2] The expenses are assumed to have some external value, which is why the government might permit their deductibility. In the empirical exercise later in the paper, I rely on the values of gross income and deductions reported in the tax return, so the correct interpretation of the model is that the *reporting* of the expenses generates external value. If the observed expenses data are misreported, then their true external value may be less than is represented in the model.

The taxpayer chooses consumption, c, gross income, y, and deductions, d to maximise utility, $u(c, y, d)$, subject to her budget constraint. She does so subject to the constant marginal tax rate, τ, which applies to taxable income, $z = y - \rho d$, with ρ

[2]This literature includes Doerrenberg et al. (2015), Schächtele (2016), and Paetzold (2017) for personal taxation, but also Best et al. (2015) and Bachas and Soto (2017) for corporate taxation.

the deductibility rate. Generating gross income reduces utility at an increasing rate ($u_y < 0$ and $u_{y,y} > 0$), while deductions increase utility at a decreasing rate ($u_d > 0$ and $u_{d,d} < 0$). This is a reduced-form representation of utility, similar to that of Feldstein (1999), in which gross income and deductions enter the utility function individually. Income generation involves disutility, for example due to the supply of labour. Deductible expenses often have consumption value, for example charitable contributions, mortgage payments, medical expenses, health-insurance premiums, or certain work-related expenses. But any portion of deductible expenses undertaken solely to generate income, so that it has no consumption value and is reported truthfully, cannot be represented as deductions in this model.[3]

Given the taxpayer's optimal gross income and deductions, indirect utility is:

$$v(\tau, \rho) = \max_{c,y,d} \; u(c, y, d) + \lambda^c \left[y - d - \tau \cdot (y - \rho d) - c \right],$$

with her first-order conditions yielding:

$$\frac{u_y}{u_d} = -\frac{1-\tau}{1-\rho\tau}.$$

As shown in Figure 1.1, under full deductibility the tax rate does not affect the relative prices of the items, but a change in deductibility does.

The government chooses τ and ρ to maximise social welfare, which includes indirect utility and the external value of deductions, $\Phi(d)$, subject to a revenue requirement, R:

$$\max_{\tau,\rho} w(\tau, \rho) = v(\tau, \rho) + \Phi(d) + \lambda^g \cdot \left[\tau \cdot (y - \rho d) - R \right].$$

The government's first-order conditions are:

$$\frac{\partial w}{\partial \tau} = \left(1 - \frac{u_c}{\lambda^g}\right) \cdot (y - \rho d) + \frac{\Phi'(d)}{\lambda^g} \cdot \frac{\partial d}{\partial \tau} + \tau \cdot \left(\frac{\partial y}{\partial \tau} - \rho \cdot \frac{\partial d}{\partial \tau}\right) = 0 \tag{1.1}$$

$$\frac{\partial w}{\partial \rho} = -\left(1 - \frac{u_c}{\lambda^g}\right) \cdot \tau d + \frac{\Phi'(d)}{\lambda^g} \cdot \frac{\partial d}{\partial \rho} + \tau \cdot \left(\frac{\partial y}{\partial \rho} - \rho \cdot \frac{\partial d}{\partial \rho}\right) = 0. \tag{1.2}$$

The first term in equations 1.1 and 1.2 represents a transfer of funds between taxpayer and government, and the last term represents a behavioural distortion.[4]

The distortion in equation 1.2 is due to a change in the deductibility rate, which

[3] In a standard model of optimal profit taxation, these kinds of expenses optimally will be fully deductible, with any limit on deductibility introducing a productive inefficiency. Best et al. (2015) note, however, that the revenue leakage caused by the misreporting of business expenses reduces welfare, meaning that less-than-full deductibility could be optimal.

[4] When deductions generate an external benefit, the elasticity of deductions with respect to the net-of-tax rate is a necessary statistic for determining the optimal tax rate, as previously observed by Doerrenberg et al. (2015).

presents a challenge for empirical work because this seldom is observed. Because we instead observe variation in the tax rate, it is useful to consider the conditions under which such variation can be relied upon for the normative analysis of deductibility. Using the implicit function theorem, the response of deductions to the deductibility rate can be written as:

$$\frac{\partial d}{\partial \rho} = -\frac{\tau}{1-\rho} \cdot \left[\frac{\partial d}{\partial \tau} - \left(\frac{u_{dy} - u_{yy}}{u_{dd} - u_{yd}} \right) \cdot \frac{\partial y}{\partial \tau} \right]. \tag{1.3}$$

As can be seen in equation 1.3, while the ETI is a sufficient statistic for the deadweight loss of the tax rate, as per Feldstein (1999), in general it is neither necessary nor sufficient for the deadweight loss of the deductibility rate. When variation is observed only in the tax rate, one needs to know also the second-order terms in the taxpayer's utility function. Without a functional-form assumption, it is unclear what the effect on deductions of a change in the tax rate implies for optimal deductibility.

Modern estimation of the ETI exploits bunching around discontinuities in the tax schedule (Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2013). In these studies, utility takes a particular quasilinear and isoelastic form. Analogous to that approach, I assume that functional form applies to gross income and deductions separately:[5]

$$u(c, y, d) = c - \frac{n_y}{1 + 1/e_y} \cdot \left(\frac{y}{n_y} \right)^{1+1/e_y} + \frac{n_d}{1 + 1/e_d} \cdot \left(\frac{d}{n_d} \right)^{1+1/e_d}, \tag{1.4}$$

in which n_y and n_d are gross income and deductions in the absence of taxes, and e_y and e_d are elasticities. The optimal gross income and deductions are then:

$$y_0 = n_y \cdot (1-\tau)^{e_y} \qquad d_0 = n_d \cdot (1-\rho\tau)^{e_d}, \tag{1.5}$$

with elasticities:

$$e_y = \frac{dy}{d(1-\tau)} \cdot \frac{1-\tau}{y} \qquad e_d = \frac{dd}{d(1-\rho\tau)} \cdot \frac{1-\rho\tau}{d}.$$

The effects of the tax and deductibility rates on deductions are proportional to one another because all that matters for deductions is the effective net-of-tax rate, $1-\rho\tau$. Accordingly, equation 1.3 becomes:

$$\frac{\partial d}{\partial \rho} = \frac{\tau}{\rho} \cdot \frac{\partial d}{\partial \tau}. \tag{1.6}$$

This is illustrated in Figure 1.1. A one-unit change in the tax rate has the same effect

[5]Only quasilinearity and separability are necessary to obtain the desired mapping; isoelasticity is necessary to obtain explicit solutions for the optimal deductibility rate.

on deductions as a τ/ρ-unit change in the deductibility rate, but the change in the deductibility rate has no effect on gross income. Under quasilinear, isoelastic, and separable utility, the deduction elasticity is a sufficient statistic for the deadweight loss of the deductibility rate.

An increase in deductions can be achieved by increasing either of the tax or deductibility rates because both increase the implicit subsidy to deductions. This raises a question as to their relative social costs in doing so. Given the government's first-order conditions and the functional form assumption, the net effect on welfare of simultaneously raising the tax rate by one unit and lowering the deductibility rate by τ/ρ units (which leaves deductions unchanged) is:

$$\frac{\partial w}{\partial \tau} - \frac{\rho}{\tau} \cdot \frac{\partial w}{\partial \rho} = -(1 - \lambda^g) \cdot y - \lambda^g \cdot \frac{\tau}{1-\tau} \cdot e_y y. \tag{1.7}$$

The two changes on net generate revenue at the cost of a behavioural response, with deductions unaffected. The deductibility rate should be chosen to maximise the net social benefit of deductions, and the two rates may then be adjusted together based on the impact on gross income.

If the deductibility rate has been set optimally, then the optimal tax rate is:

$$\tau^*(p^*) = \frac{1 - \lambda^g}{1 - \lambda^g \cdot (1 + e_y)}, \tag{1.8}$$

which takes the usual Ramsey (1927) inverse-elasticity form with the gross-income elasticity a sufficient statistic for the optimal tax rate. As a function of the prevailing tax rate, the optimal deductibility rate is:

$$\rho^*(\tau) = \frac{1}{\tau} \cdot \frac{1 - \lambda^g - \Phi'(d) \cdot e_d}{1 - \lambda^g - \lambda^g \cdot e_d}, \tag{1.9}$$

which is explicit only under a constant marginal external benefit of deductions.

Only if a dollar in the hands of the government is worth the same as a dollar in the hands of the taxpayer ($\lambda^g = 1$) is the deduction elasticity irrelevant.[6] In that case, the only concern is correcting the deduction 'externality', so the effective deduction subsidy rate, $\rho\tau$, should be set equal to the marginal external benefit of deductions, $\Phi'(d)$, as per a Pigouvian subsidy.[7] To the extent that government revenue is raised at

[6]The term, λ^g is the 'social marginal utility cost of public funds', and captures the welfare loss associated with the marginal dollar raised by government given that revenue is raised at a social cost.

[7]Note also the relevance of this setting to the 'double dividend' debate in optimal environmental taxation (Bovenberg and de Mooij, 1994). If expenses instead generated a negative externality, then the model would stipulate a *negative* deductibility rate, implying the expenses attract an additional tax burden on top of gross income. Because deductions and gross income are separable, this tax would have no effect on gross income. This assumption implies a double dividend: the negative-externality-generating expenses could be taxed with no distortion to gross income, with the proceeds used to fund a

a social cost, which surely is the case in practice, the deduction elasticity is a necessary statistic for the optimal deductibility rate. As long as the marginal external benefit of deductions exceeds the social marginal utility cost of public funds ($1 < \lambda^g < \Phi'(d)$), the Ramsey (1927) inverse-elasticity rule applies: the more elastic are deductions with respect to the net-of-tax rate, the lower is the optimal deductibility rate.

The functional-form assumption rules out income effects for gross income and deductions, heterogeneity in the gross-income and deduction elasticities (though this can be accommodated), adjustment frictions, and any dependence of gross income on the deducibility rate. It will be necessary to assume at least quasilinearity and isoelasticity to perform the bunching analysis later on. The separability assumption sets to zero the cross-price elasticities of gross income and deductions, as is sometimes assumed in order to derive the 'inverse-elasticity' representation of the Ramsey (1927) optimal tax rate. When this doesn't hold, the optimal tax rate continues to depend inversely on the own-price elasticity, but depends also on the cross-price elasticities. To the extent that substitution to other taxed goods reduces revenue leakage, the optimal tax rate is higher, and the opposite is true for complements. Any degree of complementarity or substitutability of gross income and deductions will accentuate or attenuate the effect of the deductibility rate on welfare that I derive under separability.[8]

1.3 A bunching decomposition method

When a discontinuous increase in tax liability is introduced into the tax schedule, taxpayers just above the discontinuity have an incentive to reduce their taxable incomes and bunch just below it. I identify the deduction and gross-income elasticities by attributing the reduction in taxable income associated with bunching to changes in deductions and gross income. I extend the standard bunching model (Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2013) to include deductions and gross income, and then derive their elasticities as functions of the ETI and their proportional changes relative to income in bunching.

I assume there is full deductibility ($\rho = 1$), and taxable income is gross income less deductions, $z = y - d$. In the absence of taxes, deductions are $n_y \sim F_d(\cdot)$ and gross income is $n_y \sim F_y(\cdot)$, which are the taxpayer's 'types'. Given her tax liability, $T(z)$, the taxpayer chooses gross income and deductions to maximise the quasilinear, isoelastic,

reduction in distortive income taxation.

[8]The standard taxable-income representation (Saez et al., 2012) includes a separability assumption implicitly because taxable income in a given period is itself one of several possible 'items' upon which a taxpayer could report. Taxpayers might reclassify personal income as business income, or shift income between periods in response to relative tax rates. As noted by Slemrod (1998), the taxpayer's ability to substitute into these other items undermines the sufficiency of the ETI, just as in the present case any substitution into gross income would undermine the sufficiency of the deduction elasticity.

and separable utility function specified earlier:

$$u(y, d) = y - d - T(y - d) - \frac{n_y}{1 + 1/e_y} \cdot \left(\frac{y}{n_y}\right)^{1+1/e_y} + \frac{n_d}{1 + 1/e_d} \cdot \left(\frac{d}{n_d}\right)^{1+1/e_d}.$$

A discontinuity (the treatment) is introduced into a previously linear tax schedule (the counterfactual). Deductions and gross income, and thus taxable income, are the outcomes of treatment or nontreatment. Accordingly, I use potential-outcomes notation from the program-evaluation literature (Rubin, 1974; Rosenbaum and Rubin, 1983). Deductions under treatment and nontreatment are d_1 and d_0. I refer explicitly only to deductions, but analogous representations of taxable income and gross income are implied. For now, think of the treatment as being randomly assigned.

Under a linear tax, $T(z) = t \cdot z$, deductions are $d_0 = n_d \cdot (1 - t)^{e_d}$. Given the type distributions, this generates a deductions distribution, $h_d(d_0)$. I restrict the type distributions only insofar as they generate a smooth taxable-income distribution, $h_z(z_0) = \int h_{y,d}(z_0 - d_0, d_0)\, dd_0$. Consider the introduction of a discontinuous increase in tax liability (a notch), at $z = z^*$. The notched tax function is $T(z) = t \cdot z + \Delta t \cdot z \cdot \mathbf{1}[z > z^*]$. For a taxpayer who was located just above the threshold, the introduction of the notch will mean that reducing taxable income increases after-tax income as she avoids paying the additional tax. And the only way for her to do so is to decrease gross income or increase deductions.

The taxable income of a taxpayer who is indifferent to bunching at the threshold is $z^* + \Delta z^*$. Unlike the case of taxable income with heterogeneous elasticities, this does not define a unique buncher, but rather a set of bunchers for whom $y_0 - d_0 = z^* + \Delta z^*$. If taxpayers have the same elasticities and there are no adjustment frictions, then all taxpayers with $z_0 \in (z^*, z^* + \Delta z^*]$ will bunch at the threshold. The decision to bunch is discrete, with the taxpayer comparing her utility when she bunches to that when she doesn't bunch.

The task is to identify the deduction elasticity:

$$e_d = \frac{dd/d}{d(1-\tau)/(1-\tau)}.$$

The ETI is $e = (dz/z)/(d(1-\tau)/1-\tau)$. The decision problem of a buncher is:

$$\max_d u\,(z^* - d, d) = (1 - t) \cdot z^* - \frac{n_y}{1 + 1/e_y} \cdot \left(\frac{z^* - d}{n_y}\right)^{1+1/e_y} + \frac{n_d}{1 + 1/e_d} \cdot \left(\frac{d}{n_d}\right)^{1+1/e_d},$$

that is, the choice of one of the items is residual.

To derive the deduction and gross-income elasticities, I rely on the taxpayer's always equalising the marginal utilities of gross income and deductions. For bunchers, with

the notch in place:

$$\left(\frac{y_1}{n_y}\right)^{1/e_y} = \left(\frac{d_1}{n_d}\right)^{1/e_d}, \tag{1.10}$$

with $y_1 - d_1 = z^*$, while without the notch, gross income and deductions are:

$$y_0 = n_y \cdot (1-t)^{e_y} \qquad d_0 = n_d \cdot (1-t)^{e_d}. \tag{1.11}$$

Rearranging equations 1.11, substituting for n_y and n_d in equation 1.10, and solving for the ratio of elasticities yields:

$$\frac{e_y}{e_d} = \frac{\ln y_1 - \ln y_0}{\ln d_1 - \ln d_0}. \tag{1.12}$$

The ETI is the average of the deduction and gross-income elasticities weighted by the proportions of taxable income for which they account:[9]

$$e = \frac{y}{z} \cdot e_y - \frac{d}{z} \cdot e_d. \tag{1.13}$$

Evaluating equation 1.13 in the absence of the notch, solving for the ratio of the gross-income and deduction elasticities, and substituting into equation 1.12 yields the deduction elasticity:

$$e_d = e \cdot \frac{\ln\left(\frac{\Delta d}{d_0} + 1\right)}{\frac{y_0}{z_0} \cdot \ln\left(\frac{\Delta y}{y_0} + 1\right) - \frac{d_0}{z_0} \cdot \ln\left(\frac{\Delta d}{d_0} + 1\right)} \approx e \cdot \frac{\Delta d}{\Delta z} \cdot \frac{z}{d}, \tag{1.14}$$

where the approximate form can be verified by the chain rule, and $\Delta d = d_1 - d_0$. Given the ETI, the deduction elasticity can be estimated via the change in deductions associated with a change in taxable income in bunching. The approach is depicted in Figure 1.1a.

A helpful feature of equation 1.14 is that, outside of the ETI term, the tax-rate change is irrelevant to the deduction elasticity. For a notch, it is the *average* tax rate that changes at the threshold, but the desired elasticities are with respect to the net-of-*marginal*-tax rate. The challenge this poses has been addressed for estimating the ETI, eased by the fact that the standard bunching method exploits the taxable income responses of a set of taxpayers with the same taxable income.[10] But, as I will describe in a moment,

[9]The disaggregated model I consider generates ETI heterogeneity because the ETI depends on the proportions of gross income and deductions in the tax return, which differs across taxpayers. As noted by Kleven (2016), the standard bunching model is robust to ETI heterogeneity because the ETI estimate can be interpreted as the average ETI across taxpayers. The same is true in the present case.

[10]Kleven and Waseem (2013) assume isoelastic and quasilinear utility in order to yield a closed-form

estimation of the deduction elasticity in equation 1.14 relies on the average response of a set of taxpayers among whom the proximity to the threshold varies. Their implicit marginal tax rates differ because those depend on the proximity to the threshold. But so long as you can estimate the ETI using the existing methods, explicit consideration of the tax-rate change is unnecessary.

Equation 1.14 applies for a given taxpayer, so strictly requires estimation of an individual-level treatment effect, which in practice is not observed. Instead, I must rely on an average treatment effect (ATE) among all bunchers, which introduces two imperfections: the formula is nonlinear, so the average elasticity, which I wish to measure, differs from the elasticity of the average taxpayer, which I actually measure; and I must rely on different treatment and comparison groups to construct the ATE, which could introduce a selection bias. Little can be done about the former issue, a problem that afflicts all bunching studies in which the ETI is heterogeneous, but the small range of taxable income considered suggests the bias should be modest. The latter imperfection is a major consideration in the research design, to follow.

To estimate the deduction elasticity, it is necessary to estimate the ETI, two ATEs, and two average outcomes under nontreatment:

$$\hat{e}_d = \hat{e} \cdot \frac{\hat{\mathbb{E}}[\,d_1 - d_0]}{\hat{\mathbb{E}}[\,z_1 - z_0]} \cdot \frac{\hat{\mathbb{E}}[\,z_0]}{\hat{\mathbb{E}}[\,d_0]}, \tag{1.15}$$

in which all expectations are conditional on $z_0 \in (z^*, z^* + \Delta z^*]$. Estimating the means requires observing taxpayers who face the notch and taxpayers who don't. In practice, discontinuities in the tax schedule often apply only to those with a particular characteristic, so that those to whom the notch applies can be compared to those to whom it does not. The Earned Income Tax Credit in the U.S., for example, applies only to those with children. Another potential source of variation is the introduction or removal of the discontinuity at a point in time. This is the variation I exploit, by comparing the tax returns of those present before the change to the tax returns of those present after.

In practice, the taxpayers who bunch do not all do so precisely. Instead, there tends to be a sharp spike in mass at the threshold, with a diffusion of excess mass throughout a region below the threshold, as shown in Figure 1.2. The 'manipulation region', $\mathcal{Z} = [z_L, z_U]$, is the range of taxable income around the threshold containing the origin and destination of all bunchers. The upper bound of the manipulation region, z_U, corresponds to the counterfactual taxable income of the marginal buncher, $z^* + \Delta z^*$. The threshold bisects the manipulation region into lower ($\mathcal{Z}_L = [z_L, z^*]$) and upper ($\mathcal{Z}_U = (z^*, z_U]$) portions, with taxpayers moving from the upper to the lower portion

solution for the ETI. They also propose an alternative, non-parameterised version of the formula, which relies on an inference of the approximate MTR change implicit in the average tax rate change at the threshold.

when the notch is present.

For estimating the ETI, this setting offers an advantage over the usual one, in which taxpayers are observed only when the notch is in place. The task is to estimate the counterfactual taxable income of the marginal buncher, z_U, because bunching theory connects her proximity to the threshold with the ETI. Without a counterfactual density, this is difficult to determine ocularly because the missing mass is diffuse. Kleven and Waseem (2013) address this by determining ocularly the lower bound of the manipulation region, z_L, (which is easier because the excess mass deviates more sharply from the counterfactual density), and then exploiting the equality of the excess and missing mass to estimate the upper bound of the missing mass, z_U. This is problematic when there are large extensive-margin responses (as in the present case), where those above the threshold in the absence of the notch exit the sample, causing an imbalance in the excess and missing mass (balancing them is central to the standard method). I avoid these considerations by observing a counterfactual density in a period when the notch is absent. I can estimate z_U, and thus the ETI, by determining ocularly the convergence point of the actual and observed counterfactual densities, just as Kleven and Waseem do for z_L.

To estimate the ATEs, I compare taxpayers located in the manipulation region when the notch is in place to those when it is not. I consider the entire manipulation region because it contains the origin and destination of all bunchers. A 'local average treatment effect', estimated in a small neighbourhood of the threshold via a regression-discontinuity design, would not identify the ATEs because of selection into bunching: those who bunch might differ from those who do not, and that difference could be related to the outcome. By considering the entire manipulation region, I avoid any selection bias due to bunching, as the outcomes of all bunchers and nonbunchers are always included in the estimates. For the average outcomes under nontreatment, I focus only on the upper portion of the manipulation region under nontreatment because that is where all bunchers originate.[11]

In practice, however, a simple comparison of means in the manipulation region with and without the notch might not reflect the true ATE. Even isolated from any confounding effects of other policies, deductions and gross income grow over time, and likely at different rates (the consumer price inflation rate typically differs from the wage inflation rate, for example). As a result, the gross income and deductions distributions at a given level of taxable income will differ between years in the absence of treatment. Given the fixed window of taxable income that defines the treatment and comparison groups, these year-on-year changes must be controlled for.

[11]Consideration of the average within the lower region would capture the mean outcomes under treatment rather than nontreatment, but it also would capture the outcomes of those located below the threshold in the absence of treatment, who have no incentive to respond to the treatment.

The difference-in-differences (DiD) design is a standard way to do so. It requires the availability of a placebo group that is not exposed to the treatment in either period. If the placebo group is comparable to the treatment group, then taxpayers in the placebo group should on average exhibit the changes in gross income and deductions that those in the treatment group would have exhibited in the absence of treatment. These differences can then be subtracted from the differences observed for the treatment group to identify the ATEs. The same approach can be taken to construct counterfactual average outcomes under nontreatment for those who receive the treatment.

The placebo group can be drawn from a range of taxable income just below the manipulation region. It is important to ensure that no confounding policy affects the placebo group in the periods considered. As they are located below the threshold, taxpayers in the placebo group face no incentive to alter their tax returns in response to the treatment. And their close proximity to the manipulation region should ensure that they are comparable to those in the treatment group. The validity of this assumption can be checked by ascertaining whether there are parallel trends in outcomes between the two groups prior to treatment. If there are non-parallel pre-trends, then they can be used to predict what the counterfactual outcome would have been, with estimated average treatment effects adjusted accordingly.

1.4 Institutional settings, and data

I study the removal of a notch in the Australian personal income tax system, which has a structure similar to those of comparable countries like the U.S..[12] Taxable income is calculated as the taxpayer's income items less her deductible expenses.[13,14] The system differs from some others in that there are no limits on the amount of expenses that a taxpayer may deduct. A common category is work-related expenses, for which the only limits are in the types of items that may be claimed and that the expenses were incurred in the course of earning income. For example, a work uniform with an embroidered logo is deductible, but general business attire is not, and only the portion of a computer's value used for work purposes may be deducted against income based on depreciation over its life. In the baseline group of taxpayers considered later, average

[12] All figures are for 2009, the relevant year in which the treatment is absent for the policy that I consider. References in the chapter to '2009' refer to the 2008-09 financial year, in which taxpayers lodged their tax returns between July 1 and October 31 of 2009.

[13] Income items include wage and salary income, tips, government allowances, pension income, self-employment income, fringe benefits, business income, interest, and dividends, while deductible expenses include work-related expenses (car, travel, uniform, self-education), interest and dividend expenses (any interest paid in order to generate income, which includes interest on loans to fund investments, but not for a mortgage on the primary residence), gifts and donations, business expenses, and the cost of managing one's tax affairs.

[14] Among all taxpayers, 65.3% claimed work-related deductions (Australian Taxation Office, 2009).

deductions are $1,897, of which 4% are for charitable giving, 6% are for the cost of managing tax affairs, and 86% are for work-related expenses. The largest work-related expenses are car expenses (37% of total deductions), clothing expenses (11%), travel expenses (6%), and education expenses (5%).

The Australian Commissioner of Taxation, who is responsible for administering the Australian tax system, has claimed based on a recent increase in the number of audits that work-related expenses are widely misreported (Jordan, 2017). "[The prevalence of work-related clothing expense deduction claims] would mean that almost half of the individual taxpayer population was required to wear a uniform—suits are not uniforms—or protective clothing, or had some special requirements for things like sunglasses and hats and a variety of other things. Half the population," the Commissioner said in a recent speech. Work-related expenses for a car and other travel are said also to have been widely misreported. The claim is not that many of the expenses were not made, but rather that they were undertaken for the purposes of consumption rather than to generate earnings.

The income tax schedule features a tax-free threshold (AU$6,000), and increasing marginal tax rates (15, 30, 40, and 45% at $6,000, $34,000, $80,000, and $180,000, respectively). Average full-time adult total earnings are $64,662 (Australian Bureau of Statistics, 2009), and average taxable income is $46,462 (Australian Taxation Office, 2009). Unlike the U.S., spouses must file separately, but provide some of the details of the spouse's tax return for the purposes of means tests that depend on family income. As in the U.S., it is common for taxpayers to use a tax preparer (71.2%), though the government provides a free online system for submitting tax returns, including pre-filling of government payments and third-party reported income.[15]

Australia has a government-funded universal healthcare system, similar to the U.K. 'National Health System', called 'Medicare'. This system provides subsidies for general practitioner visits, and free hospital care for most treatments. Patients seeking elective procedures often are subject to a waiting period. On top of the publicly-provided system lies a voluntary private health insurance system. Private health insurance premiums are controlled and subsidised by the government. Private health insurers incur the healthcare costs that otherwise would be borne by the public system. Patients with private health insurance might be able to receive elective procedures sooner, and receive higher-quality amenities such as a private room. Around half of Australian adults are covered by private health insurance (Australian Bureau of Statistics, 2013).

In order to encourage people to take out private health insurance, in 1998 the Australian government introduced the 'Medicare Levy Surcharge' (MLS), which is an additional 1% tax (applicable to all of taxable income) on those without private health

[15] 18.8% of tax returns were submitted by the taxpayer online (Australian Taxation Office, 2009).

insurance coverage whose taxable income is above a threshold.[16,17] It is a notch because tax liability increases discontinuously at the threshold.[18] The threshold depends on whether a taxpayer has dependents (either a dependent spouse or children). The threshold is $50,000 for those without dependents, and starts at $100,000 for those with dependents, and increases incrementally for each additional child. For a single taxpayer without children and without private health insurance, an increase in taxable income from $50,000 to $50,001 would increase tax liability by around $500, implying an effective-marginal tax rate of 50,000%.

I use the Australian Taxation Office unit record file, which is a 16% random sample of all taxpayers (around 4 million returns per year), including all items in the income tax return.[19] The data are partly self reported and partly third-party reported. To condition on the characteristics of the taxpayers to whom the policy applies, I exclude those with a spouse (must be living with the taxpayer), dependent children, private health insurance, or eligibility for a 'Medicare levy exemption category'. I consider the threshold for those without dependents because the data are not disaggregated for the spouse.

The threshold for those without dependents remained at $50,000 for 10 years. As the mass of taxpayers moved upward due to wage inflation, the tax—originally intended only for those on high incomes—came to apply to a large portion of the population. To address this concern, in 2009 the government raised the threshold from $50,000 to $70,000, and indexed it based on wage inflation. The policy otherwise was unaltered. The historical path of the threshold is displayed in Figure 1.3, with the threshold change highlighted in grey. There is no change in the marginal tax rate near the original threshold of $50,000 in the years leading up to the change, and no other policy changes apply near the threshold in the years considered.

The movement in the MLS threshold provides the necessary variation in treatment status to identify the deduction and gross-income elasticities using the bunching decomposition method. The notch is removed rather than introduced, so to minimise confusion in interpreting the results, I set the treatment period to 2009 (when the notch is absent) and the nontreatment period to 2008 (when the notch is present). The estimates should therefore be interpreted as the effects of removing the notch, which have

[16]Stavrunova and Yerokhin (2014) study the effect of this policy on private health insurance take up.

[17]The tax does not strictly apply to taxable income, but rather to 'income for MLS purposes'. This is taxable income plus fringe benefits, 'the amount on which family trust distribution tax has been paid', and 'any element of a superannuation [similar to an IRA] lump sum for which the tax rate is zero'. For the vast majority of taxpayers, the two concepts are the same.

[18]It is a 'proportional notch' because both average and marginal tax rates are discontinuous at the threshold (put differently, the policy combines a 'pure notch' with a kink). The 1% increase in the marginal tax rate should cause a leftward shift in the taxable-income density above the threshold, but, as is common in studies of proportional notches, I ignore this shift in the empirical analysis later on. The small change in the marginal tax rate suggests any movement in the density will be small.

[19]The data are not publicly available.

opposite signs to the effects of introducing the notch.

I consider the \$50,000 threshold rather than the \$70,000 threshold. The policy encourages those earning above the threshold to take up private health insurance, and this incentive increases with income. In 2009, taxpayers earning \$70,000 without private health insurance pay a tax of \$700, compared to the \$500 tax paid by those earning \$50,000. The large extensive-margin response this induces for taxpayers without private health insurance earning around \$70,000 in 2008, when the tax applied from \$50,000, means they might be poor counterfactuals for those in 2009, when the threshold was raised to \$70,000. There is no such problem with those without private health insurance earning around \$50,000 in 2009, as the policy no longer applies to them in any way. Those earning \$50,000 are also closer to the peak of the taxable-income density, which increases statistical power. The elasticity estimates apply to those with incomes around \$50,000; their external validity will depend on the degree to which the responses change with income.

The estimates apply to those who have an incentive to bunch, so it is necessary to determine who they are. With the notch, the bunchers locate below, rather than above, the notch, causing the taxable-income densities of the treated and nontreated taxpayers to diverge within a 'manipulation region', shown in Figure 1.2. I describe in appendix A.1 how I determine this region. Determining the range in which the treatment- and nontreatment-period densities diverge is equivalent to determining that in which their ratio diverges from one. A local-logit estimate of the ratio of the two densities is displayed in Figure 1.4.[20] The manipulation region is located where the slope of the density ratio diverges from (approximately) zero. I determine this ocularly as the taxable incomes between \$49,150 and \$51,400, and assess the implications of this choice in a sensitivity analysis. This defines the treatment group. For the placebo group, I include all taxpayers with a taxable income between \$42,000 and \$47,000, which is chosen to be close to the treatment group so as to be comparable, but not so close as to be tainted by the treatment, as well as far away from any confounding policy variation further below the threshold. I denote the placebo region $\mathcal{Z}^0 = [z_L^0, z_U^0]$.

Taxpayers in 2008 are indicated by $T = 0$, those in 2009 by $T = 1$, those with taxable incomes between \$49,150 and \$51,400 by $S = 1$, and those with taxable incomes between \$42,000 and \$47,000 by $S = 0$. A visual comparison of the densities of taxable income for the placebo and treatment groups across the treatment and nontreatment periods is presented in Figure 1.5. The region of taxable income occupied by the placebo group appears to be free of manipulation either due to the treatment or due to any confounding policies, with no difference between the treatment-period and

[20] There appears to be an extensive-margin response to the right of the threshold, with the treated mass less than the nontreated mass. This likely reflects the choice of some taxpayers to take up private health insurance. I estimate the effect of any extensive-margin response to be minor, and detail my approach in appendix A.2.

nontreatment-period densities for the placebo group. The nontreatment-period density also appears to offer an appropriate counterfactual for the treatment-period density. Of note is a complete absence of bunching for the treatment group immediately after the removal of the notch, suggesting taxpayers are highly responsive to the policy change.

Summary statistics are displayed in Table 1.1 for the variables of interest across the four groups. The deductions distribution has a long right tail. While median deductions are around $1,000, some taxpayers have deductions in excess of $200,000. The 99th percentile of deductions for all groups combined is $16,469. To address the effect of outliers in the deductions distribution on the results, I present the results for both all taxpayers and excluding the top 1% of deduction claimers.

1.5 Empirical analysis

1.5.1 Probability of bunching

Some taxpayers near the threshold might in practice not respond to the treatment because of adjustment frictions or ETI heterogeneity (Kleven and Waseem, 2013).[21] The bunching decomposition method relies on the standard bunching method to estimate the ETI, which is then decomposed into the item responses. This ETI estimate is for the marginal buncher, who is representative of the bunchers rather than the nonbunchers (the ETI of whom we do not observe). Consistent with this, it is appropriate to condition the item responses on the bunchers, which won't affect the relative proportional changes of the items and taxable income (as both the numerator and denominator would be attenuated equally by the prevalence of nonbunchers), but might affect the estimated average outcomes under nontreatment (if the bunchers and nonbunchers differ in that dimension), and thus the deduction and gross-income elasticities.

As shown in Figure 1.2, three groups of taxpayers are located in the manipulation region before and after treatment: those below the threshold in the absence of treatment for whom there will be no treatment effect; those above the threshold in the absence of treatment but who do not respond so for whom there will be no treatment effect; and those above the threshold in the absence of treatment but who do respond so for whom there will be a treatment effect. By default, an average in the manipulation region is for the three groups combined. But, as there is no response among two of the groups, the response among the bunchers can be determined using the proportion of taxpayers who bunch.

Bunching is indicated by $B = 1$. Given that $\mathbb{E}[\, d_1 - d_0 \mid B = 0] = 0$, the ATEs can be

[21] This may be likened to 'compliance', in the program evaluation parlance, in which not all units assigned to the treatment comply with their assignment.

inflated so as to be conditional on bunching:

$$\mathbb{E}[\, d_1 - d_0 \mid z_0 \in \mathcal{Z}, B = 1] = \frac{\mathbb{E}[\, d_1 - d_0 \mid z_0 \in \mathcal{Z}]}{\mathbb{P}[B = 1 \mid z_0 \in \mathcal{Z}]}. \tag{1.16}$$

For the average outcomes under nontreatment, the approach is similar. Given that $\mathbb{E}[\, d_1] = \mathbb{E}[\, d_0 \mid B = 0]$, the effect of the nonbunchers can be subtracted from the nontreated mean, which can be inflated in the same way:

$$\begin{aligned}&\mathbb{E}[\, d_0 \mid z_0 \in \mathcal{Z}_U, B = 1] \\ &\qquad = \frac{\mathbb{E}[d_0 \mid z_0 \in \mathcal{Z}_U] - (1 - \mathbb{P}[B = 1 \mid z_0 \in \mathcal{Z}_U]) \cdot \mathbb{E}[\, d_1 \mid z_0 \in \mathcal{Z}_U]}{\mathbb{P}[B = 1 \mid z_0 \in \mathcal{Z}_U]}.\end{aligned} \tag{1.17}$$

Equations 1.16 and 1.17 require estimates of the bunching probability, which can be obtained from the densities of the treated and nontreated taxpayers. If the distribution under nontreatment is a valid counterfactual for that under treatment, then the bunching probability is:

$$\mathbb{P}[B = 1 \mid z_0 \in \mathcal{Z}] = \frac{H_0(z_U) - H_0(z^*)}{H_0(z_U) - H_0(z_L)} - \frac{H_1(z_U) - H_1(z^*)}{H_0(z_U) - H_0(z_L)}, \tag{1.18}$$

where $H_0(\cdot)$ is the cumulative distribution function of taxable income under nontreatment. This corresponds to the missing mass above the threshold in the treated density. The probability in the upper portion of the manipulation region, to be used to condition the average outcomes under nontreatment on bunching, is the same as that in equation 1.18, except the denominators are conditional on $z_0 \in \mathcal{Z}_U$. Equation 1.18 can be estimated via the corresponding empirical distributions:

$$\hat{\mathbb{P}}[B = 1] \mid T = 1, S = 1] = \frac{\sum_{i=1}^{n} \mathbf{1}[z_i > z^*] \cdot (1 - T_i) \cdot S_i}{\sum_{i=1}^{n} (1 - T_i) \cdot S_i} - \frac{\sum_{i=1}^{n} \mathbf{1}[z_i > z^*] \cdot T_i \cdot S_i}{\sum_{i=1}^{n} (1 - T_i) \cdot S_i}, \tag{1.19}$$

with standard errors computed via a bootstrap procedure.[22,23]

If the density of taxable income in the treatment and nontreatment periods were to differ in the absence of treatment, then a DiD approach can be used in which the distribution of the treatment group in the nontreatment period is adjusted for any

[22] In the bootstrap procedure, I draw with replacement 10,000 random samples (each with a sample size equal to the full sample size), then perform the estimation procedure in equation 1.19 on each of the samples. The reported standard errors are the standard deviations of the estimated probabilities across the 10,000 samples.

[23] To obtain the probability of bunching among taxpayers in the upper portion of the manipulation region, which is used to condition the levels of the outcomes under nontreatment among bunchers, the same procedure may be used but with an indicator function included in the denominators given by $\mathbf{1}[z_i > z^*]$.

change over time in the placebo distribution. For identification, it must be the case that the growth rates of the distributions of the placebo and treatment groups would have been the same in the absence of treatment:

$$\frac{H_0(z_U \mid T=0, S=1) - H_0(z_L \mid T=0, S=1)}{H_0(z_U \mid T=1, S=1) - H_0(z_L \mid T=1, S=1)} = \frac{H_0(z_U^0 \mid T=0, S=0) - H_0(z_L^0 \mid T=0, S=0)}{H_0(z_U^0 \mid T=1, S=0) - H_0(z_L^0 \mid T=1, S=0)}. \tag{1.20}$$

The assumption for the upper portion of the manipulation region of bunching is the same, except that the distributions of the treatment group ($S = 1$) are conditional on $z_0 \in \mathcal{Z}_U$. The corrected bunching probability can then be estimated by multiplying the second fraction in equation 1.19 by:

$$\frac{\hat{H}(z_U^0 \mid T=0, S=0) - \hat{H}(z_L^0 \mid T=0, S=0)}{\hat{H}(z_U^0 \mid T=1, S=0) - \hat{H}(z_L^0 \mid T=1, S=0)} = \frac{\sum_{i=1}^{n}(1-T_i)\cdot(1-S_i)}{\sum_{i=1}^{n} T_i \cdot (1-S_i)}.$$

The bunching probability estimates are presented in Table 1.2. The 'DiD' columns adjust for the observed change in the placebo density.

1.5.2 Changes in the items when bunching

In the DiD design, identification of the ATEs requires the difference in nontreated outcomes between the treatment and placebo groups to be constant over time:

$$\mathbb{E}[\,d_0 \mid T=1, S=1] - \mathbb{E}[\,d_0 \mid T=1, S=0] = \mathbb{E}[\,d_0 \mid T=0, S=1] - \mathbb{E}[\,d_0 \mid T=0, S=0]. \tag{1.21}$$

A potential threat to identification is non-parallel pre-trends in the average outcomes of the treatment and comparison groups. In the context I consider, data are available only for the three years prior to treatment. Observe in Figure 1.6 that, for taxable income, the within-year difference between the treatment and placebo groups appears to be constant prior to treatment, consistent with parallel pre-trends. This is unsurprising, as the same band of taxable income is chosen in each year. The differences for gross income and deductions, however, appear to narrow over time.

As the placebo group is located in a lower region of the taxable income distribution, those taxpayers also have on average lower deductions and gross income, both of which are increasing functions of taxable income. Even if the inflation rates of the items are independent of taxable income, they are being applied to a lower base for the placebo group. This means that the year-on-year changes in gross income and deductions would be higher for the placebo group than for the treatment group, leading to a

convergence of the two over time. The secular nature of this process suggests it should be straightforward to address, aided by the apparent linearity of the time trend.

I estimate the ATEs under three specifications. First, I compute a simple difference in average outcomes in the manipulation region between the treatment and nontreatment periods. As there is reason to doubt the exogeneity of the treatment, I also compute standard DiD estimates. Then, to address non-parallel pre-trends, I use a linear regression of the outcomes in the three years prior to the treatment period to predict the nontreatment-period outcomes, which I substitute for the observed outcomes when calculating the ATEs. The latter generates my preferred estimates.

The three specifications rely on two regression models. The first is a standard DiD regression of the form:

$$d_i = \beta_0 + \beta_1 \cdot T_i + \beta_2 \cdot S_i + \beta_3 \cdot (T_i \times S_i) + \varepsilon_i. \tag{1.22}$$

For the first specification, the estimate of the ATE is $\hat{\beta}_1$, and, for the second specification, it is $\hat{\beta}_3$. For the third specification, let the year an observation is observed, among the three prior to treatment, be $Y_i \in \{2006, 2007, 2008\}$, and then let:

$$y_i = \begin{cases} -1 & \text{if } Y_i = 2008 \\ -2 & \text{if } Y_i = 2007 \\ -3 & \text{if } Y_i = 2006. \end{cases}$$

The second regression is:

$$d_i = \delta_0 + \delta_1 \cdot S_i + \delta_2 \cdot y_i + \delta_3 \cdot (y_i \times S_i) + \gamma_i. \tag{1.23}$$

For the third specification, the estimate of the ATE is then $\hat{\beta}_2 + \hat{\beta}_3 - \hat{\delta}_1$.

Regression output for the three specifications is presented in Table 1.3, and the estimates for the DiD specifications are depicted in Figure 1.6. The columns titled 'Difference', 'DiD (1)', and 'DiD (2)' contain estimates under the first, second, and third specifications, respectively. To address outliers in the deductions distribution, I compute the first two both on all taxpayers, and on a group excluding the top 1% of deduction claimers, and the third specification only on the latter. In the final column of the table are estimates computed under the third specification but conditional on the bunchers. For these, the standard errors are estimated via a bootstrap procedure.[24] I find the effect on the results of any extensive-margin response to be modest, and

[24]In the bootstrap procedure, I draw with replacement 10,000 random samples (each with a sample size equal to the full sample size), then perform the estimation procedure in equation 1.19 on each of the samples. I then divide the ATE estimate by the bunching probability estimate. The reported standard errors are the standard deviations of the ATE conditional on bunching across the 10,000 samples.

describe my approach in appendix A.2.

The DiD estimates are similar across the trimmed and full samples, but the standard errors for the former are substantially lower. The simple differences appear to misrepresent substantially the treatment effect. As can be seen in Figure 1.6, there is some variation in outcomes across years, but the variation of the treatment and placebo groups is highly correlated, which supports the DiD design. The presence of converging pre-trends appears to bias the estimates, overstating the role of deductions and understating that of gross income.

The differences-in-differences of the three outcomes, with deductions represented in terms of their contribution to taxable income (that is, negative deductions), are depicted in Figure 1.7. Under parallel pre-trends, the differences in differences in the years prior to treatment would be zero. As shown in Figure 1.7a, this is not the case for gross income and deductions, but controlling for linear pre-trends shifts the plots for gross income and deductions vertically so that they are centered at zero, as shown in Figure 1.7b. Deductions account for around a third of the response in taxable income, with gross income accounting for the remaining two thirds.

1.5.3 Levels of the items before bunching

For the average outcomes under nontreatment, the identification assumptions and regression specifications are the same as those for the ATEs. The analysis focuses, among the treatment group, only on those taxpayers with a taxable income above the threshold. The counterfactuals I construct are estimates of what the outcomes would have been in the treatment period in the absence of treatment.

For the first specification, the estimate is $\hat{\beta}_0 + \hat{\beta}_1$ from the model in equation 1.22, but, for those in the treatment group, only among those above the threshold. For the second specification, the estimate is $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$ from the same model. For the third specification, the estimate is $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 - \hat{\delta}_3$ from the models in equations 1.22 and 1.23. In this specification, the difference in slopes between the treatment and placebo groups over the three-year period prior to treatment is deducted from the treatment effect to account for pre-trends.

Regression output for the three specifications is presented in Table 1.4, and construction of the counterfactuals is depicted in Figure 1.8. The columns titled 'Actual', 'C.f. (1)', and 'C.f. (2)' contain estimates under the first, second, and third specifications, respectively. In the final column estimates are computed under the third specification, but conditional on the bunchers. For these, the standard errors are estimated via a bootstrap procedure.

There are large differences in outcomes under nontreatment between the bunchers and nonbunchers, which is indicative of selection into bunching. Bunchers are esti-

mated to have 12% higher deductions in the absence of treatment, for example. This highlights the importance of focusing on changes in average outcomes within the entire manipulation region, rather than locally at the threshold or above the threshold, where any estimated treatment effects would be biased by selection into bunching. The counterfactuals are displayed in Figure 1.8. As one would expect, the first counterfactual remains parallel to the levels in the placebo group between 2006 and 2009. This appears to be inappropriate, as the levels of both gross income and deductions converge prior to treatment. The second counterfactual, which adds the convergence in trends back to the counterfactual, appears to perform better.

1.5.4 Deduction and gross-income elasticities

The key results of the chapter are presented in Table 1.5.[25] The sufficient statistics of interest are the deduction and gross-income elasticities with respect to the net-of-tax rate, which consist of the elasticities with respect to taxable income, driven by the relative proportional changes of deductions and gross income, and taxable income, multiplied by the ETI. The first column displays the proportion of deductions and gross income in taxable income in the absence of treatment, and the second the proportion of the change in taxable income that can be attributed to them. Dividing the second column by the first yields the item elasticity with respect to taxable income. Then multiplying by the ETI yields the deduction and gross-income elasticity with respect to the net-of-tax rate.

Deductions constitute 5% of taxable income, but account for 35% of the response of taxable income to the tax rate. Accordingly, a 1% decrease in taxable income is achieved via a 7.5% increase in deductions, and only a 0.62% decrease in gross income. The ETI is a scalar, having no effect on the relative item elasticities. Using the reduced-form approximation of Kleven and Waseem (2013), the upper bound of the manipulation region ($51,400) implies an ETI of 0.06.[26] This yields a gross-income elasticity of 0.04, and a deduction elasticity of -0.45.

While deductions account for only a small fraction of taxable income, they account for a large fraction of the response of taxable income to the tax rate, making the deduction elasticity more than an order of magnitude larger than the gross-income

[25] These results are computed using: the trimmed sample, which excludes the top 1% of deduction claimers; the 'DiD' specification for the bunching probability estimation, which accounts for the change over time in the placebo density; the pretrend-corrected estimates for both the changes and levels of the outcome; and the estimates conditioned on bunching, which differ only due to the different levels in the absence of treatment between the bunchers and nonbunchers.

[26] From Kleven and Waseem (2013), the reduced-form approximation of the ETI is:

$$e \approx \frac{\left(\left(z_U - z^*\right)/z^*\right)^2}{\Delta t/\left(1-t\right)} = \frac{(\$1,400/\$50,000)^2}{0.01/0.685} = 0.057.$$

elasticity. This suggests that a reduction in deductibility could substantially reduce the ETI, and thus the impact of income taxation on welfare. Whether the ETI would fall all the way to equal the gross-income elasticity depends on the validity of the separability assumption.

Using the formula for the optimal deductibility rate (equation 1.9), it is possible to determine the external benefit that deductions would need to generate in order for full deductibility to be optimal. Assuming that the marginal tax rate is 0.315, as it is where the notch is located, and that the marginal social utility cost of public funds (that is, the opportunity cost of $1 of funds held by government) is 1.2 (implying a marginal efficiency cost of revenue collection of 20%), then the marginal dollar of deductions would need to generate at least 68¢ in external value over and above the benefit to the taxpayer claiming them.[27] If $1 of deductions were to generate, for example, 30¢ in external benefits, the optimal deductibility rate would be just 34%.

1.6 Conclusion

The composition of the response of taxable income to the tax rate, not just its magnitude, is informative for policy. While many scholars have studied the ETI, and generally found it to be modest, there has not been an appreciation of the sources of the response. My results suggest that, if gross income and deductions are separable, or close to it, then a substantial lowering of the deductibility rate would substantially lower the ETI. This is similar to the argument of Slemrod and Kopczuk (2002), who show that the ETI is endogenous to the size of the tax base. That such a large proportion of the ETI is driven by something as small as deductions suggests a policy change is in order. Indeed, it seems inconceivable that the same marginal tax rate should apply to two items for which the behavioural responses are more than an order of magnitude apart.

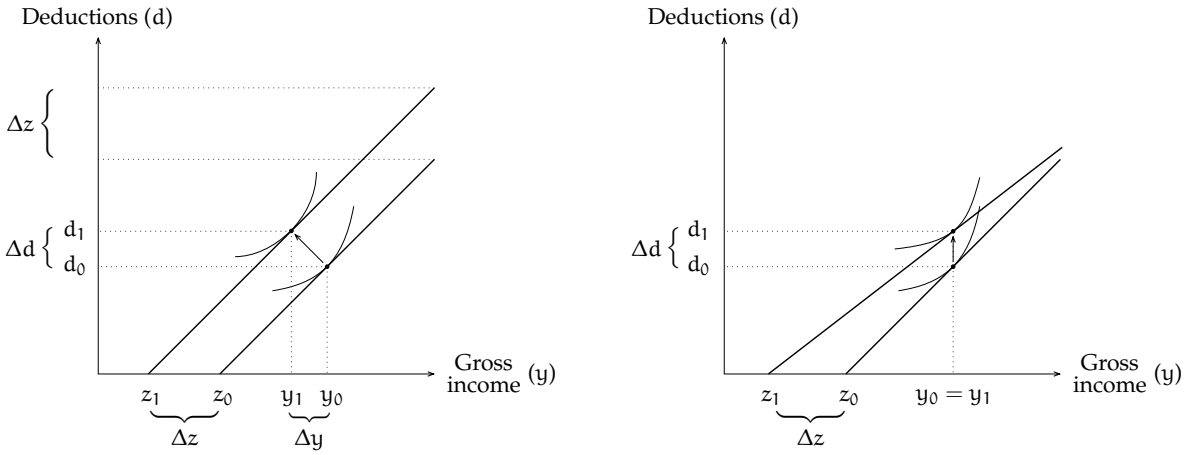
My results of course apply to the Australian tax system. The observed elasticities are a function of the setting—the tax law, enforcement regime, local tax morale, among countless other things. The relatively liberal conditions governing the deductions that can be claimed are sure to play a role, and the results will differ in other countries, such as the U.S., where the rules are tighter. That a fairly large deduction elasticity has been estimated too in Germany and Austria lends some weight to the external validity of the findings. But I offer a set of tools that can be applied wherever a notch is introduced or removed, supporting replicability in other settings.

The recent focus on the ETI has been motivated by Feldstein's (1999) argument that it summarises the effect of the income tax rate on welfare. In reality, however, the government has more tax instruments at its disposal than just the income tax rate. It

[27] Even if the efficiency cost of revenue collection were only 10%, deductions would still need to generate 48¢ in external value in order for full deductibility to be optimal.

is within the government's remit to decide the extent to which taxpayers may claim deductions. But it could just as easily choose the extent to which the tax rate applies to all of the other items in the tax return. For this richer set of choices, the ETI is not sufficient. In this chapter, I propose a method for decomposing the ETI into the item elasticities. Armed with estimates of more of the item elasticities, governments could make changes to the effective item-specific tax rates that raise welfare.

Figures



(a) 1-unit increase in τ.

(b) τ/ρ-unit increase in ρ.

Figure 1.1: *Increase in the tax versus deductibility rate.* Taxable income is $z = y - d$. The responses assume quasilinear, isoelastic, and separable utility, so the the effect of the tax and deductibility rates on deductions is the same, but only the tax rate affects gross income.

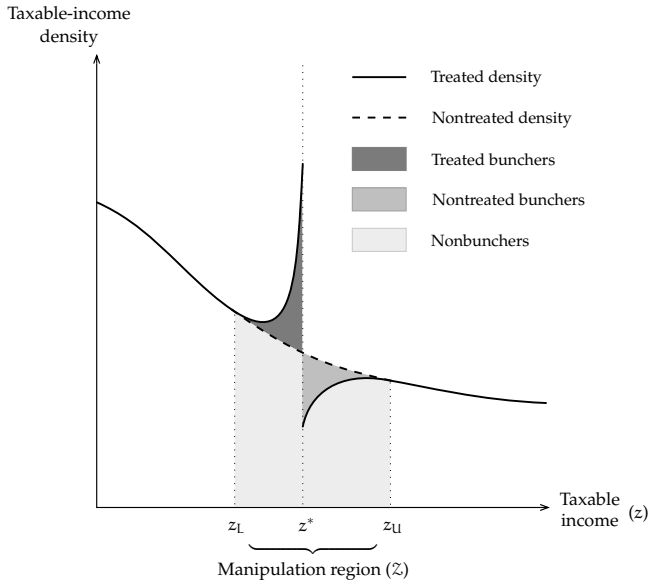


Figure 1.2: *Effect of manipulation on taxable-income density.* The 'bunchers' relocate below the threshold when treated, while the 'nonbunchers' don't alter their taxable income when treated.

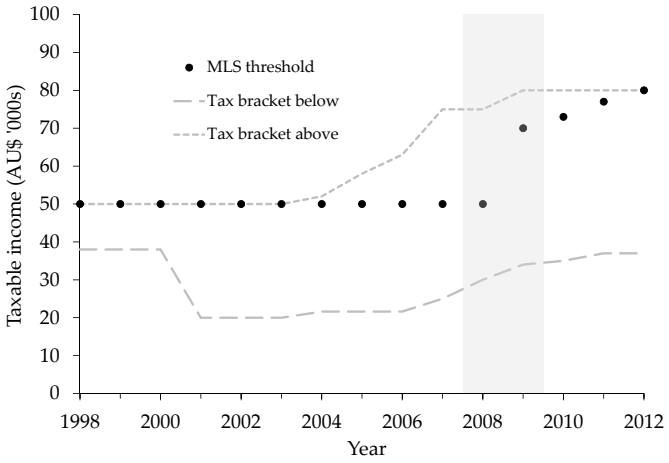
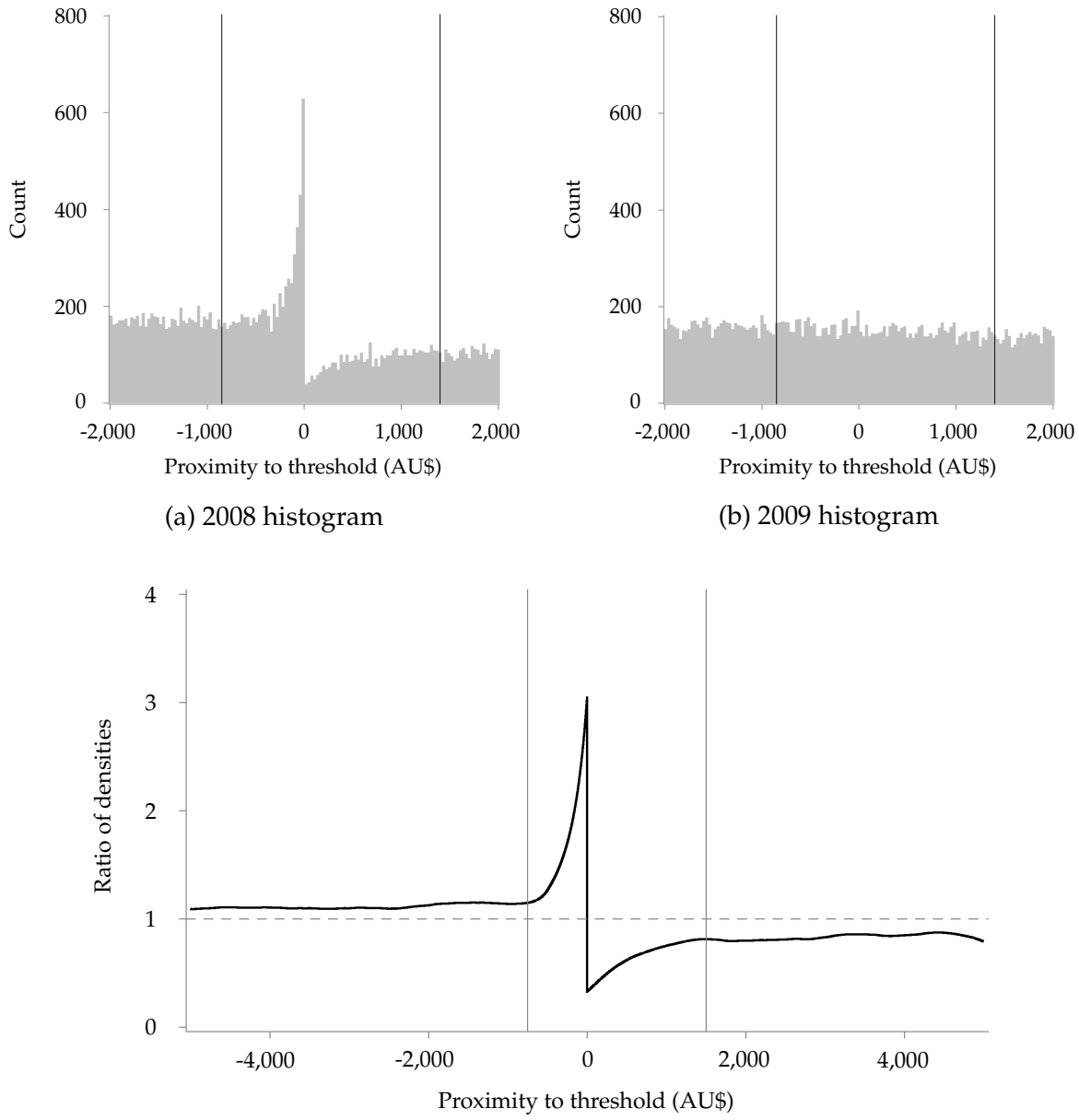


Figure 1.3: *Medicare Levy Surcharge threshold over time.* The surcharge was introduced in 1998, with the threshold constant at $50k until 2009, when it was raised to $70k, then indexed.

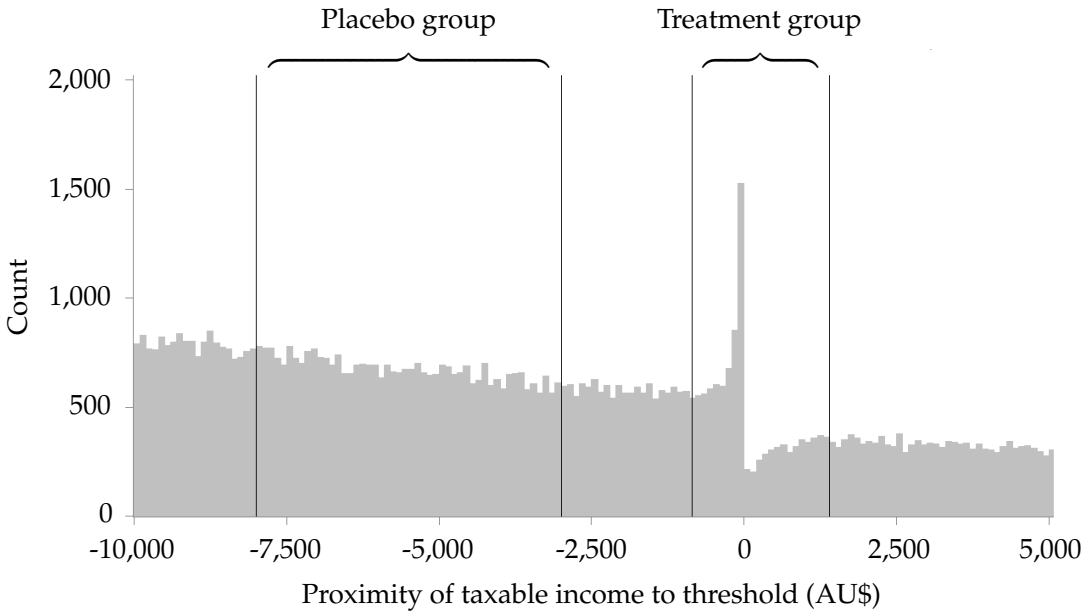


(a) 2008 histogram

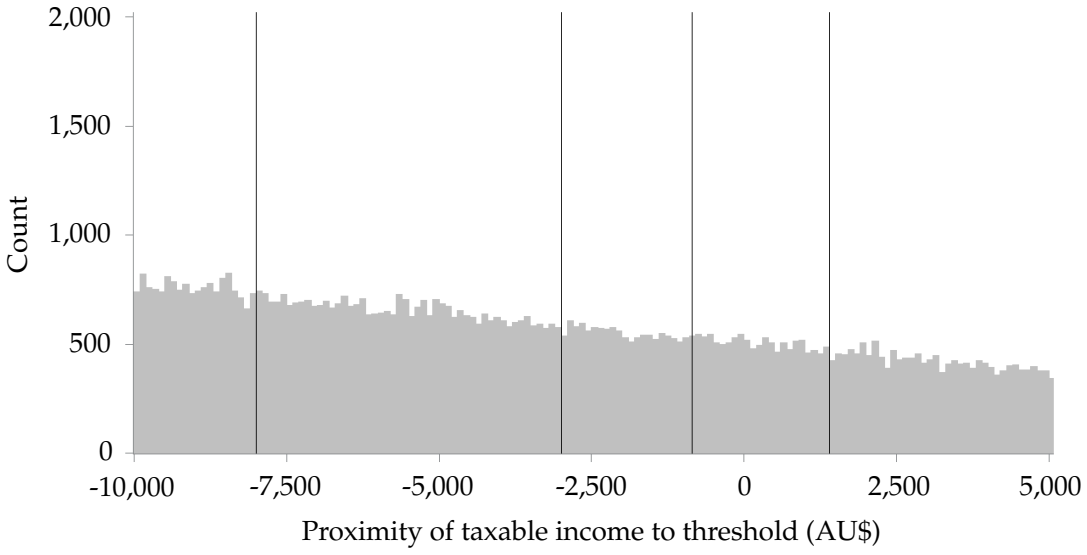
(b) 2009 histogram

(c) Estimated ratio of 2008 to 2009 densities

Figure 1.4: *Determining the manipulation region.* The top two subfigures show the histograms of taxable income under treatment and nontreatment. As the missing mass above the threshold is diffuse, it is difficult to determine ocularly the convergence point of the two densities when observing only the treated density. This is not a problem for me because I observe the nontreated density. In the present case, this is eased by observation of the nontreatment density. The bottom subfigure plots an estimate of the ratio of the treated density in subfigure 1.4a and the nontreated density in subfigure 1.4b. The vertical axis is the multiple of taxpayers in the treatment group relative to the nontreatment group. The convergence and divergence points are readily identifiable. The details of the estimation procedure are in appendix A.1. The vertical bars indicate the chosen manipulation region, [-850, 1,400].

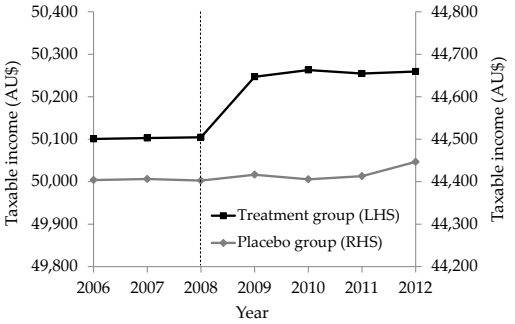


(a) Treatment period (2008).

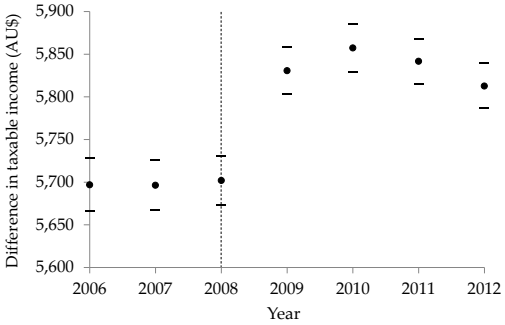


(b) Nontreatment period (2009).

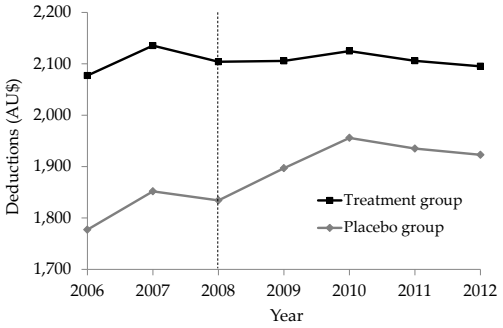
Figure 1.5: *Histograms of the proximity of taxable income to the threshold.* The histograms cover taxable incomes from \$40,000 to \$55,000, with a bin size of \$100. The four groups displayed are the placebo and treatment groups in the treatment and nontreatment periods. The removal of the treatment immediately eliminates the distortion present under treatment. The treatment appears to have no effect on the placebo-group density in either period.



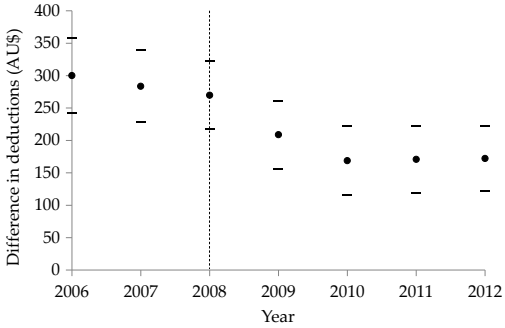
(a) Levels of taxable income



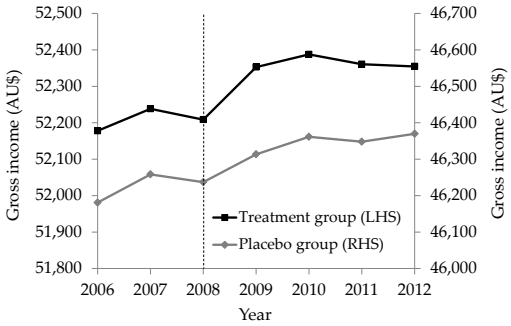
(b) Difference in taxable income



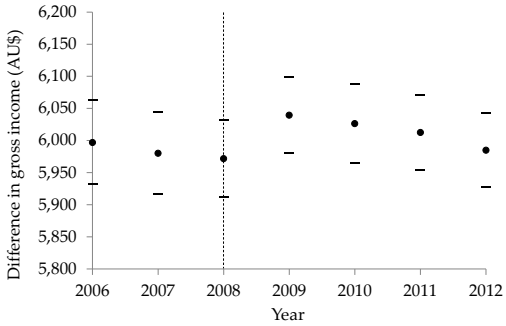
(c) Levels of deductions



(d) Difference in deductions

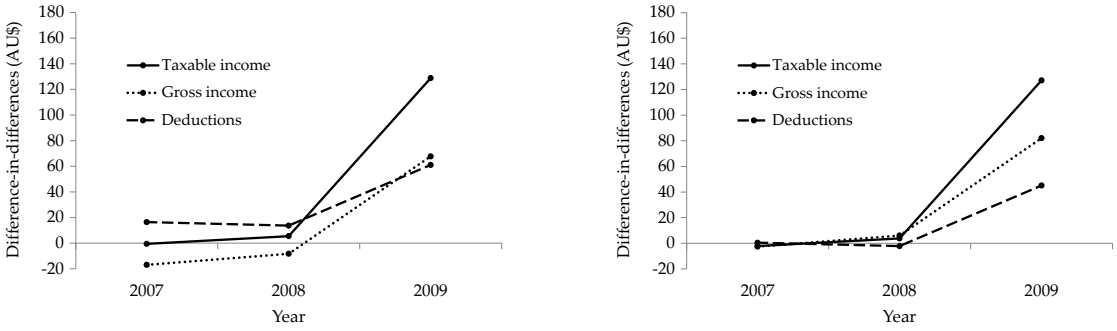


(e) Levels of gross income



(f) Difference in gross income

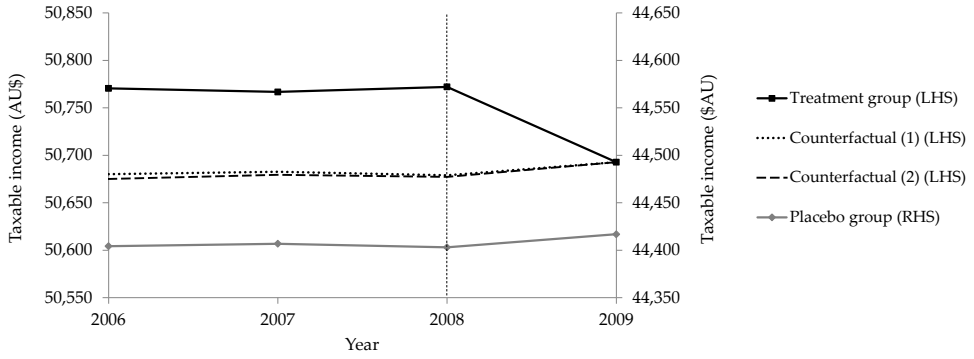
Figure 1.6: *Checking for parallel pre-trends.* The notch is removed in 2009. Presented in figures 1.6a, 1.6c, and 1.6e are the average levels of the outcomes from 2006 to 2012 in the treatment and placebo groups. In figures 1.6b, 1.6d, and 1.6f are the differences in the average levels of the outcomes between the treatment and placebo groups, with 95% confidence interval bands. The figures exclude the top 1% of deduction claimers.



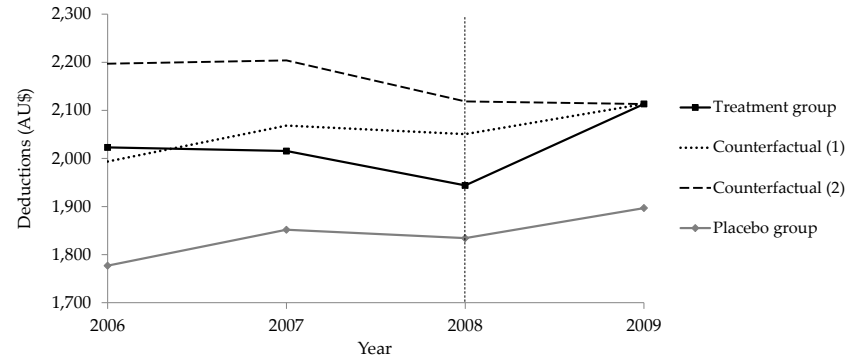
(a) Observed (DiD (1))

(b) Linear-pretrend corrected (DiD (2))

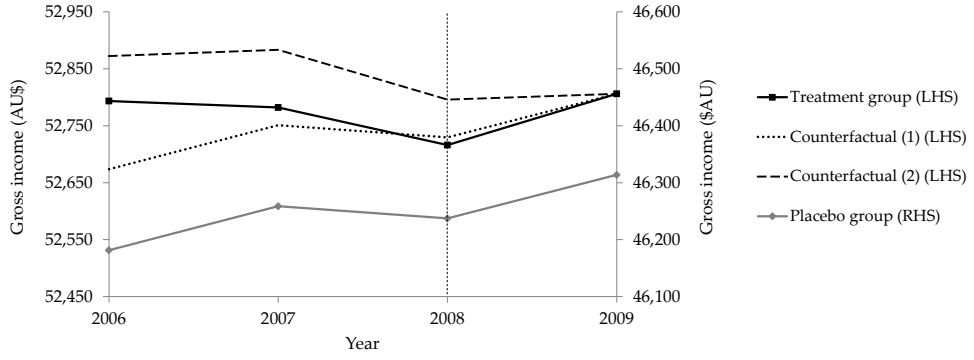
Figure 1.7: *Differences-in-differences over time.* The figures display the year-on-year differences-in-differences for the two items and taxable income for two years with the treatment and one year without. The negative of deductions is displayed to reflect their contribution to taxable income. In figure 1.7a, the displayed difference is between the observed year-on-year differences between the averages for the two groups. In figure 1.7b, the year-on-year differences have been corrected for linear pre-trends, which induces a parallel upward or downward shift in each of the lines compared to those displayed in figure 1.7a.



(a) Taxable income



(b) Deductions



(c) Gross income

Figure 1.8: *Counterfactual outcomes.* The grey and black solid lines indicate the observed outcomes in the placebo and treatment groups. The dotted line is the 2008 level estimated using a standard DiD, assuming the change in levels between 2008 and 2009 would have been the same in the treatment and placebo groups in the absence of treatment. The dashed line is the same as the dotted line, but corrected for linear pre-trends. The figures exclude the top 1% of deduction claimers.

Tables

	Median	p90	p99	Max.	Mean	St. dev.	N
Deductions							
Placebo, nontreatment	872	4,937	15,182	230,400	2,028	3,559	33,983
Placebo, treatment	900	5,107	14,959	137,153	2,072	3,475	33,019
Treatment, nontreatment	1,118	5,507	17,537	105,477	2,370	3,865	10,615
Treatment, treatment	1,058	5,687	16,725	110,041	2,348	3,788	11,486
Gross income							
Placebo, nontreatment	45,867	49,857	59,818	272,522	46,432	3,856	33,983
Placebo, treatment	45,902	50,088	59,827	179,511	46,490	3,780	33,019
Treatment, nontreatment	51,455	55,736	67,685	155,993	52,475	3,913	10,615
Treatment, treatment	51,537	56,096	67,317	160,807	52,595	3,841	11,486
Taxable income							
Placebo, nontreatment	44,370	46,434	46,946	47,000	44,404	1,443	33,983
Placebo, treatment	44,393	46,442	46,946	47,000	44,417	1,435	33,019
Treatment, nontreatment	49,947	51,109	51,370	51,400	50,105	632	10,615
Treatment, treatment	50,237	51,159	51,376	51,400	50,247	650	11,486

Table 1.1: *Summary statistics.* The four groups are: the placebo group (taxable incomes between $42,000 and $47,000) in the nontreatment period (2009); the placebo group in the treatment period (2008); the treatment group (taxable incomes between $49,150 and $52,250) in the nontreatment period; and the treatment group in the treatment period.

	Manipulation region		Above threshold	
	Difference	DiD	Difference	DiD
Bunching probability	0.2311	0.2413	0.3836	0.4005
	(0.0080)	(0.0083)	(0.0119)	(0.0125)
N	15,596	82,082	11,071	77,549

Table 1.2: *Estimated probabilities of bunching.* 'Difference' is the ratio of missing mass above the threshold in the treatment period to the total mass in the manipulation region in the nontreatment period. 'DiD' controls for the difference in mass for the placebo group between the treatment and nontreatment periods.

	Bunchers and nonbunchers					Bunchers
	Full sample		Top 1% deductions trimmed			
	Diff.	DiD (1)	Diff.	DiD (1)	DiD (2)	DiD (2)
Deductions	-22.03	-66.62	1.58	-61.03	-45.05	-186.70
	(51.50)	(55.83)	(34.43)	(37.97)	(49.74)	(215.55)
Gross income	120.44	62.79	144.16	67.73	82.05	340.03
	(52.18)	(59.51)	(35.45)	(43.25)	(56.74)	(230.60)
Taxable income	142.47	129.40	142.58	128.76	127.10	526.73
	(8.63)	(20.00)	(8.69)	(20.11)	(26.61)	(79.75)
N	22,101	89,103	21,860	88,338	166,111	166,111

Table 1.3: *Estimated average treatment effects on the treated.* 'Bunchers and nonbunchers' refers to estimates among all taxpayers in the manipulation region, and 'Bunchers' among only those who respond to the treatment. 'Diff.' is the average difference for the treatment group between 2008 and 2009. 'DiD (1)' is the difference between the treatment and placebo groups' average differences between 2008 and 2009. 'DiD (2)' is corrected for linear pre-trends between 2006 and 2008. In the final column, the second-last column is divided by the probability estimate in the second of four columns in Table 1.2.

	Bunchers and nonbunchers					Bunchers
	Full sample		Top 1% deductions trimmed			
	Actual	C.f. (1)	Actual	C.f. (1)	C.f. (2)	C.f. (2)
Deductions	2,353.25	2,308.66	2,113.27	2,050.65	2,118.53	2,379.74
	(46.52)	(97.07)	(30.86)	(65.60)	(71.65)	(154.06)
Gross income	53,045.85	52,988.20	52,806.04	52,729.61	52,795.82	52,915.09
	(46.66)	(104.02)	(31.16)	(75.64)	(82.74)	(160.42)
Taxable income	50,692.60	50,679.53	50,692.78	50,678.95	50,677.29	50,535.35
	(4.84)	(36.25)	(4.86)	(36.45)	(40.07)	(44.32)
N	6,917	73,919	6,849	73,327	178,211	178,211

Table 1.4: *Estimated average outcomes among the treated under nontreatment.*'Bunchers and nonbunchers' refers to estimates among all taxpayers in the manipulation region, and 'Bunchers' among only those who respond to the treatment. 'Actual' is the average level for the treatment group in 2009. 'C.f. (1)' is adjusted by the change in outcomes observed for the placebo group between 2008 and 2009. 'C.f. (2)' is corrected for linear pre-trends between 2006 and 2008. In the final column, the second-last column is divided by the probability estimate in the final column of Table 1.2.

	% of TI	% of ΔTI	Item elasticity w.r.t.	
			Taxable income	Net-of-tax rate
Deductions	4.71	35.45	−7.53	−0.45
Gross income	104.71	64.55	0.62	0.04

Table 1.5: *Estimated deduction and gross-income elasticities.*The first column contains the estimated percentage of taxable income accounted for by deductions and gross income in the treatment period in the absence of treatment. The second column contains the estimated average treatment effects on the treated as a percentage of the estimated change in taxable income. The estimates in the third column are calculated by dividing the second column by the first. This yields the elasticities of deductions and gross income with respect to taxable income, which is their percentage change given a 1% change in taxable income. The estimates in the final column are calculated by multiplying the third column by the ETI estimate of 0.06. These results exclude the top 1% of deduction claimers, using the 'DiD' specification for the bunching probability, the pretrend-corrected estimates for both the changes and levels of the outcome, and conditional on the bunchers (which differ only due to the different levels in the absence of treatment between the bunchers and nonbunchers).

	Implied bunching effect	Estimated differences in differences	Implied extensive margin effect
Deductions	-9.68	-20.17	-10.49
		(19.36)	
Gross income	17.64	5.718	-11.92
		(21.90)	
Taxable income	25.89	27.32	-1.43
		(9.92)	

Table 1.6: *Estimated extensive-margin effect.* In the first column, the final column of Table 1.3 (the estimated ATEs among bunchers) is multiplied by the estimated probability of bunching in the exempt group, 5.19%. This is the expected effect on the exempt group due to bunching alone, provided the responses of the bunchers in the exempt and non-exempt groups are the same. The second column contains the difference-in-difference estimates for the non-exempt group, which are the observed effects. By subtracting the implied bunching effect in the first column from the observed effect in the second column, I obtain the implied effect due to the extensive-margin response, in the final column.

Chapter 2

How do you solve a problem like manipulation? A nonparametric propensity-score reweighting method for RD designs

2.1 Introduction

The regression-discontinuity (RD) design has been heralded as the 'gold standard' of quasi-experimental design, and found to emulate well what has been considered the 'gold standard' of all designs, the randomised controlled trial (Chapman, Cook, Zurovac, Coopersmith, Finucane, Vollmer and Morris, 2018). Its perceived good performance and the large number of settings for which the method seems appropriate, have lead it to great popularity. But metaanalyses of the internal validity of the RD design can only ever consider studies that actually have been performed. Potential problems with the RD design can be found not in the studies that have been performed, but in those that have not been performed because the necessary assumptions were judged not to be satisfied.

In a RD design, the treatment is assigned to a unit when its value of a 'running variable' exceeds some threshold. The method involves comparing the average outcomes of those with values of the running variable just above and just below the threshold to observe the effect of the treatment on the outcome at the threshold. The validity of doing so rests on the assumed continuity across the threshold of the nonntreated outcome, which is not observed. In many settings, this assumption, though strictly unverifiable, is defensible on the basis of the institutional settings in place and diagnostic testing.

But there are typical cases in which a defence cannot credibly be mounted. It is not uncommon for units to have both precise control over the value of the running variable (in the sense that they can choose without error to locate on one side of the threshold or the other), and a strategic interest in moving from one side of the threshold to the other. The treatment that applies on one side of the threshold is often desirable or undesirable, encouraging units to select into or out of the treatment group. When the units that do

so have on average different values of the outcome from those who do not (and from those already located at the destination), manipulation biases the estimated treatment effect in a case of classical selection bias.

Until now, practitioners who have encountered manipulation have tended either to discard all data within the range of manipulation and perform an extrapolation in place (Barreca, Lindo and Waddell, 2015), or to abandon the endeavour entirely. In this paper, I offer a potential solution to this problem. Common situations in which a RD design otherwise would be appropriate offer a comparison group that does not face the policy discontinuity, and thus is not party to manipulation. In some cases, the policy discontinuity is introduced or removed at a particular point in time, and data are observed before and after the change. In others, often due to policy design, the discontinuity only applies to a particular category of person.

The key to the method is that the distribution of certain characteristics of these units can serve as an anchor for the comparable distribution among those who *do* face the discontinuity, the two having drifted apart due to manipulation. Lags of the outcome are a typical example of such a characteristic, and their suitability depends on how well they predict what the outcome would have been in the absence of the policy discontinuity. The method requires one to observe the necessary variables for a relevant comparison group, which won't be satisfied in every case. Under assumptions about the observed characteristics and the comparability of the two groups, correction for the divergence in the observed covariates enables recovery of the average treatment effect absent manipulation bias.

In section 2.2, I characterise manipulation bias, reviewing the identification framework underpinning RD designs, recasting the treatment-assignment mechanism to accommodate manipulation, and defining manipulation bias in that framework. In section 2.3, I identify the not-directly-observed, but desired treatment effect using propensity-score weights that summarise the aforementioned characteristics. I then propose procedures for nonparametrically estimating these propensity-score weights, and using these weights to estimate the desired average treatment effect. In section 2.4, I conclude.

2.2 Manipulation bias in RD designs

2.2.1 Identification in RD designs

RD designs are used to exploit a treatment assignment mechanism in which the treatment, which is expected to affect the outcome of interest, is assigned only to units for which the value of a 'running variable' is either less than or greater than some threshold. The goal is to estimate an average treatment effect at the threshold, z_0 (ATE(z_0)).

I shall use a running example for clarity. Suppose that, beyond a certain income level, a levy is payable by taxpayers who do not have health insurance. The policy is designed to encourage health insurance takeup, but only for those with at least the threshold income level. Those who face the prospect of the levy can be expected to be more likely to take out private health insurance coverage. A RD design is a natural choice for measuring this effect by exploiting the discontinuous change in incentives, and thus probability of having health insurance, at the threshold.

Using standard potential-outcomes notation and the RD design notation of Hahn, Todd and Van der Klaauw (2001), let $Y_i(1)$ and $Y_i(0)$ be the potential value of the outcome for unit $i \in \{1, \ldots, n\}$ under treatment and nontreatment, Z_i the value of the running variable for unit i, and z_0 the threshold value of the running variable. In the running example: the treated outcome indicates whether the taxpayer takes up health insurance in the state of the world in which the levy will apply if she does not take it up; the nontreated outcome indicates whether the taxpayer takes up health insurance in the state of the world in which the levy does not apply if she does not take it up; and the running variable is her income. The ATE one seeks to estimate with a RD design is $\mathbb{E}\left[Y_i(1) - Y_i(0) \mid Z_i = z_0\right]$, which represents the effect of the levy on health insurance take up on average at the threshold.

The challenge, as in the 'fundamental problem of causal inference' (Holland, 1986), is that one does not, for any given unit, observe both the treated and non-treated outcomes. Let the actual treatment status be indicated by T_i, which is assigned according to the rule, $T_i = \mathbb{1}\left[Z_i \geqslant z_o\right]$. In the example, the treatment status indicates whether or not the taxpayer must pay the levy when she does not have health insurance. The observed value of the outcome is given by $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. For no unit with a value of the running variable in a neighbourhood of the threshold are both the treated and nontreated outcomes observed. As with other methods of causal inference, RD designs can be seen as solving what effectively is a missing-data problem.

To do so, Hahn et al. (2001) prove that, if $\mathbb{E}\left[Y_i(0) \mid Z_i = z\right]$ and $\mathbb{E}\left[Y_i(1) \mid Z_i = z\right]$ are continuous in z at z_0, then one may identify the ATE as follows:

$$\mathrm{ATE}(z_0) = \lim_{\varepsilon \to 0^+} \mathbb{E}\left[Y_i \mid Z_i = z_0 + \varepsilon\right] - \lim_{\varepsilon \to 0^-} \mathbb{E}\left[Y_i \mid Z_i = z_0 + \varepsilon\right].$$

These limits are the estimands in an RD design. The consequence of the continuity assumption is that it validates units in a neighbourhood but on either side of the threshold serving as counterfactuals for one another. Continuity—though fundamentally unverifiable—is critical to causal identification in an RD design. A stylised representation of a RD design is displayed in Figure 2.1.

2.2.2 A two-stage treatment assignment mechanism

To formalise the problem of manipulation bias, it is useful to reconsider the treatment assignment mechanism that is assumed to operate in RD designs. Consider instead that a 'two-stage' treatment assignment mechanism operates. In the first stage, units are *exposed* to the rules governing treatment assignment, and are able to respond to them prior to treatment. The treatment remains the application of the levy. When the taxpayer chooses her income, she may take account of the fact that if it exceeds the threshold level and she does not have health insurance, then she will have to pay the levy. In the second stage, units are then *treated* on the basis of their value of the running variable following their first-stage response to exposure. If the taxpayer's income after exposure is above the threshold, then a levy will apply if she does not have insurance, which might affect her decision about whether or not to buy insurance. With two stages, one can separate the response of the running variable to exposure (which can be seen as its own kind of 'treatment effect') and the ordinary response of the outcome to treatment.

Let $Y_i(1)$ and $Y_i(0)$ be potential outcomes in the treated and non-treated states of the world in the second stage. Let $Z_i(1)$ and $Z_i(0)$ be potential values of the running variable in the exposed and nonexposed states of the world in the first stage. Let the actual exposure status in the first stage be indicated by S_i, which for now I assume is exogenous. For example, taxpayers would not be exposed in the year prior to the introduction of the policy, but would be exposed in the year the policy is introduced. In that case, exposure is exogenous if the policy introduction is unanticipated and if no confounding concomitant event occurs. The observed value of the running variable then is given by $Z_i = S_i Z_i(1) + (1 - S_i)Z_i(0)$. Let T_i indicate treatment in the second stage, which now is assigned according to the rule, $T_i = S_i \mathbb{1}[Z_i \geqslant z_0]$. The observed value of the outcome then is given by $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$.

Implicit in this setup is an assumption about the stability of the potential outcomes under changes in the running variable, akin to the 'Stable Unit Treatment Value Assumption' (or 'SUTVA') typically made under the potential-outcomes framework (and which of course I also assume herein). I have assumed that each unit is endowed with a set of just four values: $\{Z_i(0), Z_i(1), Y_i(0), Y_i(1)\}$. This means that $Y_i(0)$ and $Y_i(1)$ do not depend, for any particular unit, on $Z_i(0)$ or $Z_i(1)$. The observed outcome, Y_i, may only depend on the observed value of the running variable via the treatment status, T_i, and thus may not depend directly on the exposure status, S_i. The implication of this assumption is that, among the treated (and thus exposed) units, the (unobserved) value of the outcome under nontreatment is what the value of the outcome would have been had the unit not only not been treated but also had not been exposed.

The two-stage treatment assignment mechanism reveals an additional missing-data

problem. In addition to the fact that one only observes $Y_i(0)$ for those units with $Z_i < z_0$, and $Y_i(1)$ for those units with $Z_i \geqslant z_0$, one also only observes $Z_i(0)$ for those units with $S_i = 0$, for which one never observes $Y_i(1)$ as those units never receive the treatment. That is, one does not observe the effect of the levy on the taxpayer's decision to take up health insurance in the year prior to the introduction of the policy.

2.2.3 The problem of manipulation bias

The continuity assumption guarantees identification of the ATE because it ensures that units on either side but within a neighbourhood of the threshold are valid counterfactuals for one another. There are two cases in which this continuity assumption would be violated. The first is one in which, even in the absence of the policy of interest, units on one side of the threshold differ from those on the other side in a way that is related to the outcome. If a separate health insurance subsidy applies beyond the threshold, for example, then any effect at the threshold of that subsidy on health insurance takeup would invalidate the continuity assumption. The second is one in which some units manipulate, and without error, their value of the running variable from one side of the threshold to the other, perhaps induced by the attractiveness or otherwise of the treatment. A taxpayer could both avoid paying for health insurance and avoid paying the levy by reducing her income to below the threshold.

The latter case is the focus of this paper.[1] When some units manipulate the value of the running variable following exposure, the potential value of the running variable under exposure diverges from that under nonexposure, formalised in the following definition.

Definition 1 (Manipulation)**.** *Let* $\Delta Z_i \equiv Z_i(1) - Z_i(0)$ *and* $N(z_0) \equiv [z_0 - \varepsilon, z_0 + \varepsilon]$ *for some small* $\varepsilon > 0$*. Suppose that* $\{Z_j(0), Z_j(1)\} \cap N(z_0) \neq \emptyset$ *for* $i = j$*. There has been manipulation by unit* j *in a neighbourhood of the threshold if* $\Delta Z_j \neq 0$*.*

There are three ways in which manipulation in a neighbourhood of the threshold might occur. The first is if those just above (below) the threshold move to just below (above) it. This is an example of nonrandom self selection, as the treatment effect is estimated only for those units in a neighbourhood of the threshold. The second is if those just above (below) the threshold move to below (above) but not within a neighbourhood of the threshold. This is an example of nonrandom attrition, as the relevant units exit the population for which the treatment effect is estimated. The third is if those above (below) but not within a neighbourhood of the threshold move to just

[1] The former is considered by Grembi, Nannicini and Troiano (2016), who with their 'difference-in-discontinuities' design exploit a pre-treatment comparison group to 'difference out' the confounding effect of a coincident policy discontinuity.

below (above) it. This is an example of nonrandom participation, as the relevant units enter the population for which the treatment effect is estimated.

Manipulation in a neighbourhood of the threshold does not necessarily violate the continuity assumption. Two additional conditions are necessary for that. First, as noted by Lee (2008), the manipulation must be without error (or 'precise') at the threshold rather than smoothly distributed across it.[2] If the manipulation is precise, then the density of the running variable will be discontinuous at the threshold, which is straightforward to verify.[3] Second, the manipulation must be related to the outcome, which could occur in two ways. The first is one in which there is a relationship between the running variable and either potential outcome spanning the threshold. For example, if those with higher incomes in the absence of exposure are more likely to have health insurance, then manipulation from above to below the threshold would raise the probability of having health insurance below the threshold, and *vice versa*. The second is one in which the propensity to manipulate is related to the outcome. That is, even if health insurance take up is invariant to income in the absence of exposure, if those who manipulate are less likely to have health insurance, then manipulation by those taxpayers would lower the probability of having health insurance below the threshold.

If there is precise manipulation of the running variable in a neighbourhood of the threshold, and this manipulation is related to the outcome, then the continuity assumption will be violated. But note that the continuity assumption made by Hahn et al. (2001) is sufficient rather than necessary for identification of the ATE as they define it. To see this, suppose that there is manipulation only by units i, with $Z_i(0) \in [z_0, z_0 + \delta]$ for all i and for some $\delta \geqslant 0$, with $Z_i(0) = z_0$ for some i, and $Z_i(1) \notin [z_0 - \varepsilon, z_0]$ for all i and some small $\varepsilon > 0$. Suppose that, naturally, this results in $\mathbb{E}[Y_i(0) \mid Z_i]$ being continuous at $Z_i = z_0$, and in $\mathbb{E}[Y_i(1) \mid Z_i]$ being only right-continuous, and thus discontinuous, at $Z_i = z_0$. This scenario is depicted in Figure 2.2.

But in such a case, in which all of the conditions set out earlier are met, the ATE still is identified because $\lim_{\varepsilon \to 0^-} \mathbb{E}[Y_i \mid Z_i = z_0 + \varepsilon] = \mathbb{E}[Y_i(0) \mid Z_i = z_0]$ and also $\lim_{\varepsilon \to 0^+} \mathbb{E}[Y_i \mid Z_i = z_0 + \varepsilon] = \mathbb{E}[Y_i(1) \mid Z_i = z_0] = \mathbb{E}[Y_i(1) \mid Z_i(1) = z_0]$.[4] That is:

$$\begin{aligned} &\lim_{\varepsilon \to 0^+} \mathbb{E}[Y_i \mid Z_i = z_0 + \varepsilon] - \lim_{\varepsilon \to 0^-} \mathbb{E}[Y_i \mid Z_i = z_0 + \varepsilon] \\ &= \mathbb{E}[Y_i(1) \mid Z_i = z_0] - \mathbb{E}[Y_i(0) \mid Z_i = z_0] = \mathrm{ATE}(z_0). \end{aligned}$$

[2] See Lee and Lemieux (2010) for more on this distinction.

[3] See McCrary (2008) for a test of precise manipulation via a discontinuity in the density at the threshold.

[4] As mentioned earlier, there are three possible cases of manipulation bias, of which this is one. The others are where: units enter a neighbourhood above (below) the threshold but do not enter that below (above); and units exit a neighbourhood above (below) and enter that below (above). Under these alternative cases, identification of the ATE would fail.

This is paradoxical: there has been manipulation, and this has generated a discontinuity in the density at the threshold, yet the observed conditional mean in a neighbourhood of the threshold continues to identify the ATE.

There is a simple reason for this that goes to the importance of the two-stage treatment assignment mechanism I described earlier. When there is manipulation, this ATE—defined in terms of the observed, manipulated value of the running variable—no longer in fact is the object of interest. When there is manipulation, the old ATE is $\mathbb{E}[Y_i(1) - Y_i(0) \mid Z_i = z_0] = \mathbb{E}[Y_i(1) - Y_i(0) \mid Z_i(1) = z_0]$, but the object we actually seek to identify to determine the causal impact of the treatment is $\mathbb{E}[Y_i(1) - Y_i(0) \mid Z_i(0) = z_0]$. This is the ATE that would have prevailed in the absence of manipulation. Without manipulation, the two objects would coincide; manipulation bias is their divergence.[5] This ATE, alongside the observed conditional expectation subject to manipulation bias, is displayed in Figure 2.3. The identification challenge is that, when there is manipulation, the ATE is not directly observable, as for no unit are both $Y_i(1)$ and $Z_i(0)$ observed.

2.3 Correcting for manipulation bias

2.3.1 Intuition

Manipulation of the running variable is neither a necessary nor sufficient condition for the failure of identification in an RD design. As noted earlier, precise manipulation of the running variable (that is, manipulation without error across the threshold) that is related to the outcome is sufficient, however. As such, the McCrary (2008) test of a discontinuity in the density of the running variable at the threshold, commonly performed in studies with RD designs, is simply a diagnostic test, indicating only the potential for manipulation bias. In addition to the occurrence of manipulation, it is necessary also to show it is related to the outcome. To do so, Lee (2008) suggests looking for discontinuities at the threshold in (ideally pretreatment) covariates suspected of being related to the outcome.

Instead of using these covariates to *test* for manipulation, I propose using them to *correct* for manipulation. If a comparison group (the 'nonexposed') without an incentive to manipulate is observed at each value of the running variable, then one would expect the distributions of the covariates proposed by Lee (2008) to be undisturbed among those in the comparison group. These can then serve as an anchoring point for the corresponding distributions among those with an incentive to manipulate (the 'exposed'), which have diverged due to manipulation. The trick is to find the right set of

[5] Barreca et al. (2015) refer to this as 'heaping-induced bias', where 'heaping' is an accumulation in the density of the running variable resulting from manipulation. In the public finance literature, this phenomenon is known as 'bunching' (Kleven, 2016).

variables, conditioning upon which will recover the outcome distribution undistorted by manipulation but still affected by the treatment. It is necessary for these covariates among those in the comparison group to be unaffected by the treatment among those in the comparison group, and for manipulation not to be complete in a neighbourhood of the threshold for any value of the covariates of the nonexposed.

Formally, exposure induces a response in the distributions of the outcome variable, Y_i, the running variable, Z_i, and a set of covariates, $\mathbf{X}_i$, among only the exposed units. In a neighbourhood of the threshold, the distributions of Y_i and $\mathbf{X}_i$ are altered by the presence of manipulators. This means that, in a neighbourhood of the threshold, the joint distributions of $\mathbf{X}_i$ differ among the exposed between the manipulators and nonmanipulators (those who would have been there in the absence of exposure), but do not differ between the nonmanipulators and nonexposed. By conditioning, within a neighbourhood of the threshold, the conditional distribution of the outcome among the exposed on the covariate values among the nonexposed, one recovers the conditional distribution of the outcome (and thus the true ATE) in the absence of manipulation.

2.3.2 Identification

The task is to identify the ATE, $\mathbb{E}[Y_1(1) - Y_i(0) \mid Z_i(0) = z_0]$. This can be achieved in two steps. The approach is first to construct $\mathbb{E}[Y_i \mid Z_i(0) = z, S_i = 1]$, not directly observed, and second to take its limits at the threshold. One then may invoke the continuity argument of Hahn et al. (2001) to identify the ATE. To that end, I make an analogous continuity assumption.

Assumption 1 (Continuity)**.** $\mathbb{E}[Y_i(1) \mid Z_i(0) = z]$ *and* $\mathbb{E}[Y_i(0) \mid Z_i(0) = z]$ *are continuous in* z *at* z_0*.*

This assumption implies that, aside from the policy of interest and manipulation, there is no feature at the threshold that generates a discontinuity in the conditional expected potential outcomes; for example, other policies operating at the threshold that could affect the outcome, or the threshold holding broader significance in a way that affects the outcome. This assumption differs from that made by Hahn et al. (2001) only in that it explicitly is conditional on the potential value of the running variable in the absence of exposure.

Under continuity, and the exogeneity of S_i, one can derive the ATE as follows:

$$\begin{aligned}\lim_{\varepsilon \to 0^+} \mathbb{E}[Y_i \mid Z_i(0) = z_0 + \varepsilon, S_i = 1] - \lim_{\varepsilon \to 0^-} \mathbb{E}[Y_i \mid Z_i(0) = z_0 + \varepsilon, S_i = 1] \\ = \mathbb{E}[Y_i(1) \mid Z_i(0) = z_0] - \mathbb{E}[Y_i(0) \mid Z_i(0) = z_0] = \text{ATE}(z_0).\end{aligned}$$

It remains to identify the nonobserved function, $\mathbb{E}[Y_i \mid Z_i(0) = z, S_i = 1]$. Recall that one observes some units not exposed to the threshold, for which $S_i = 0$ and thus

$Z_i = Z_i(0)$ and $Y_i = Y_i(0)$. One also observes some units exposed to the threshold, for which $S_i = 1$ and thus $Z_i = Z_i(1)$ and $Y_i = \mathbb{1}[Z_i \geqslant z_0]\, Y_i(1) + (1 - \mathbb{1}[Z_i \geqslant z_0])Y_i(0)$. For both groups, one observes units with values of the running variable in a neighbourhood of the threshold, defined earlier as $N(z_0) \equiv [z_0 - \varepsilon, z_0 + \varepsilon]$ given some small $\varepsilon > 0$.

The approach will be to reweight the observed conditional expectation function among the exposed, $\mathbb{E}[Y_i \mid Z_i = z, S_i = 1] = \mathbb{E}[Y_i \mid Z_i(1) = z, S_i = 1]$, which is subject to manipulation bias at the threshold, so that it equals the nonobserved object, $\mathbb{E}[Y_i \mid Z_i(0) = z, S_i = 1]$, which is not subject to manipulation bias. To perform this reweighting, suppose that one observes the values of a set of random variables $\mathbf{X}_i$ such that the following assumption holds.

Assumption 2 (Manipulation on observed variables). *Let* $Y_i \sim F_Y(\cdot)$ *with* $f_Y(\cdot) = F'_Y(\cdot)$, *then* $f_Y(y \mid Z_i(1) = z, \mathbf{X}_i = \mathbf{x}, S_i = 1) = f_Y(y \mid Z_i(0) = z, \mathbf{X}_i = \mathbf{x}, S_i = 1)$ *for* $z \in N(z_0)$ *and* $Y_i \in \{Y_i(0), Y_i(1)\}$.

This is akin to the standard 'selection-on-observed-variables' assumption, so in the present context it could be termed one of 'manipulation on observed variables'. It states that, given a set of observed covariates, the distributions of the observed outcome among the exposed units conditional on the values of the running variable are equal in the exposed and nonexposed states of the world. The validity of the assumption relies on observing a set of variables that accounts fully for the effect of manipulation on the observed outcome. The assumption is local as the reweighting is strictly required only in a neighbourhood of the threshold for the purpose of estimating the ATE.

In addition, one must be sure to reweight the observed object to match the correct values of the observed variables. That is, for each value of the running variable under nonexposure in the neighbourhood of the threshold, the nonexposed units must have covariate values that match those of the exposed units. If they do, then the non-exposed units offer valid counterfactuals for the exposed units. If the treatment were to affect the covariates as well as the outcome, then reweighting the exposed units to match the nonexposed would not only unwind the effect on the outcome of manipulation but also some of the effect of the treatment, inducing bias. Random assignment of exposure is sufficient but not necessary to address these problems. In a practical sense, the assumption would be satisfied if one were to use pretreatment covariate values with the policy introduction unanticipated, and if no other relevant change occurs simultaneously. This is reflected in the following assumption.

Assumption 3 (As-good-as-random assignment). $\mathbf{X}_i \perp\!\!\!\perp S_i \mid Z_i(0) = z$ *for* $z \in N(z_0)$.

In order to construct the weights, note that one observes $f_{\mathbf{X}}(\mathbf{x} \mid Z_i = z, S_i = 1) = f_{\mathbf{X}}(\mathbf{x} \mid Z_i(1) = z, S_i = 1)$ and $f_{\mathbf{X}}(\mathbf{x} \mid Z_i = z, S_i = 0) = f_{\mathbf{X}}(\mathbf{x} \mid Z_i(0) = z, S_i = 0)$. By

Bayes' rule:

$$\begin{aligned}
&f_{\mathbf{X}}(\mathbf{x} \mid Z_i(0) = z, S_i = 0) \\
&\quad = f_{\mathbf{X}}(\mathbf{x} \mid Z_i(1) = z, S_i = 1) \frac{1 - \mathbb{P}[S_i = 1 \mid Z_i = z, \mathbf{X}_i = \mathbf{x}]}{\mathbb{P}[S_i = 1 \mid Z_i = z, \mathbf{X}_i = \mathbf{x}]} \frac{\mathbb{P}[S_i = 1 \mid Z_i = z]}{1 - \mathbb{P}[S_i = 1 \mid Z_i = z]} \\
&\quad \equiv f_{\mathbf{X}}(\mathbf{x} \mid Z_i(1) = z, S_i = 1)\omega,
\end{aligned}$$

where ω are the standard 'propensity-score weights', but conditional on the running variable.[6] In order for these weights to be well defined, it is necessary to make the following 'common-support' assumption.

Assumption 4 (Common support)**.** *Let* $\mathbb{P}[S_i = 1 \mid Z_i = z, \mathbf{X}_i = \mathbf{x}] > 0$ *and* $\mathbb{P}[S_i = 1 \mid Z_i = z] < 1$*, both for* $z \in N(z_0)$.

Common support means that, for given values of the observed variables, manipulation in a neighbourhood of the threshold must not be complete; that is, there must be both manipulators and non-manipulators on either side of and arbitrarily close to the threshold. This is a testable assumption.

Under Assumptions 1 through 4, and using the propensity-score weights just derived, the nonobserved conditional expectation can be obtained by reweighting the observed conditional expectation as follows:[7]

$$\begin{aligned}
\mathbb{E}[Y_i \mid Z_i(0) = z, S_i = 1] &\equiv \int y f_Y(y \mid Z_i(0) = z, S_i = 1)\, dy \\
&= \int\int y f_{Y,\mathbf{X}}(y, \mathbf{x} \mid Z_i = z, S_i = 1)\, \omega d\mathbf{x} dy \\
&\equiv \tilde{\mathbb{E}}[Y_i \mid Z_i = z, S_i = 1].
\end{aligned}$$

Using this reweighted conditional expectation, the ATE is obtained as follows:

$$\begin{aligned}
&\lim_{\varepsilon \to 0^+} \tilde{\mathbb{E}}[Y_i \mid Z_i = z_0 + \varepsilon, S_i = 1] - \lim_{\varepsilon \to 0^-} \tilde{\mathbb{E}}[Y_i \mid Z_i = z_0 + \varepsilon, S_i = 1] \\
&= \mathbb{E}[Y_i(1) \mid Z_i(0) = z_0] - \mathbb{E}[Y_i(0) \mid Z_i(0) = z_0] = \mathrm{ATE}(z_0).
\end{aligned}$$

2.3.3 Estimation

The propensity-score weights could be estimated via logit or probit regression (Rosenbaum and Rubin, 1983), or the 'generalised boosted regression' method used in the machine learning literature (McCaffrey, Ridgeway and Morral, 2004). Care must be taken when selecting an estimator because both conditional probabilities that constitute

[6] Weights of this form are sometimes referred to as 'inverse propensity-score weights' or 'inverse probability-of-treatment weights'.

[7] See Appendix A for a proof of this result.

the weights are discontinuous at the threshold, and the focus is on a small neighbourhood of the threshold. The local-likelihood logit regression (a standard nonparametric local regression but with a logistic rather than simple linear specification for the local regressions and estimation via maximum likelihood) is ideally suited to such a setting, exhibiting the standard characteristic of local regressions of performing well at endpoints (Frölich, 2006).

The estimated probability conditional on $Z_i = \tilde{z}$ (but not on $\mathbf{X}_i$) is $\hat{\mathbb{P}}[S_i = 1 \mid Z_i = \tilde{z}] = 1/(1 + \exp(-g(\tilde{z}; \hat{\boldsymbol{\alpha}}_{\tilde{z}})))$, where:

$$\hat{\boldsymbol{\alpha}}_{\tilde{z}} = \underset{\boldsymbol{\alpha}_{\tilde{z}}}{\arg\max} \sum_{i=1}^{n} \left(S_i \ln \left(\frac{1}{1 + e^{-g(Z_i;\hat{\boldsymbol{\alpha}}_{\tilde{z}})}} \right) + (1 - S_i) \ln \left(\frac{1}{1 + e^{g(Z_i;\hat{\boldsymbol{\alpha}}_{\tilde{z}})}} \right) \right) \cdot K(Z_i - \tilde{z}, h),$$

with kernel function $K(Z_i - \tilde{z}, h)$ and bandwidth h, and:

$$g(Z_i; \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 \mathbf{1}[Z_i > z_0] + \alpha_2 Z_i + \alpha_3 \left(Z_i \cdot \mathbf{1}[Z_i > z_0] \right).$$

For the estimator of the probability conditional also on the covariates, this function includes $\mathbf{X}_i$ fully interacted.

These estimates can be combined with the estimated conditional treatment probabilities to generate the desired unit-level weights. As with the inverse-propensity-score reweighting of a simple mean, these are then used to reweight the units when computing sample averages for the conditional mean estimates. In this way, they perform the same function as the sampling weights sometimes used in survey methods.

Because of the discontinuity at the threshold, it is standard in RD designs to estimate the ATE via local polynomial regression (Calonico, Cattaneo and Titiunik, 2014). As the true ATE identified earlier is simply a reweighted version of the standard ATE estimated in RD designs, the true ATE can be estimated by reweighting the standard local polynomial estimator. The estimate of the upper-limit is $\hat{\beta}_0$ from the following weighted least-squares regression:

$$\left(\hat{\beta}_0, \ldots, \hat{\beta}_p\right) = \underset{\beta_0,\ldots,\beta_p}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j \left(Z_i - z_0 \right)^j \right)^2 \cdot \hat{\omega}_i \cdot K(Z_i - z_0, h) \cdot \mathbb{1}\left[Z_i \geqslant z_0 \right],$$

for a polynomial of order p, kernel function $K(\cdot)$, bandwidth h, and consistent estimates of the propensity-score weights $\hat{\omega}_i$. The the lower limit can be estimated analogously. Standard errors can be estimated using a bootstrap procedure.

A local polynomial regression is a weighted least-squares regression centered on the threshold, where the chosen kernel function assigns less weight to observations with a value of the running variable further from the threshold. The 'propensity-score reweighted local polynomial regression' estimator I propose additionally weights the

observations according to their probability of treatment conditional on the running variable and the value of observed covariates. Given the right covariates, the units subject to manipulation will, at a given value of the running variable, have lower weights than those not subject to manipulation, and will thus have a greater impact on estimation of the conditional mean function. This is similar to the propensity-score reweighted kernel density estimation procedure performed by DiNardo, Fortin and Lemieux (1996).

2.3.4 Alternative approaches

Keele, Titiunik and Zubizarreta (2015) propose a method that incorporates covariate balance into a RD design to address manipulation across a geographic boundary, but which conceivably could be applied in other, non-geographic contexts.[8] In doing so, they appeal to the 'randomization' interpretation of RD designs of Lee (2008), rather than the 'continuity' interpretation of Hahn et al. (2001) to which I appeal. Their approach is to balance covariates across the threshold, and via matching rather than weighting.

The key difference between their approach and mine is in the groups between which covariate balance is sought, and among which the treatment effect applies. While they seek to equate the covariate values of units on one side of the threshold with those on the other, I seek to equate the covariate values of units on both sides of the threshold with the corresponding units (those with the same value of the running variable) in a separate comparison group. The outputs of these two procedures could differ substantially in particular contexts.

To understand why, note that under the usual continuity assumption, the conditional expectations of all characteristics related to the outcome must be continuous (in the running variable) at the threshold. This means that the characteristics of units located in a neighbourhood but on opposite sides of the threshold must have been approximately equal on average prior to exposure. Consequently, in the absence of manipulation, the 'Average Treatment Effect on the Treated' (ATET) must equal the 'Average Treatment Effect on the Untreated' (ATEU), with both equal to the 'Average Treatment Effect' (ATE).

When there is manipulation, however, the values of these objects diverge along with the divergence in the values of covariates related to the outcome across but within a neighbourhood of the threshold. The nature of this divergence depends on the pattern of manipulation. As I described earlier, manipulation comes in one of three forms: units may exit a neighbourhood above the threshold, they may enter a neighbourhood below the threshold, or they may do both.

[8] A similar approach is taken by Linden and Adams (2012).

In the standard setup, if there is no nonrandom attrition or participation (units nonrandomly exiting or entering the treatment or control groups), then the ATE ordinarily can be obtained via a weighted average of the ATET and ATEU conditional on observed variables. In a RD design with manipulation, the feasibility of doing so depends on the form of manipulation that has occurred. Only when the units that have exited a neighbourhood above the threshold correspond, one-to-one, to the units that have entered a neighbourhood below the threshold can one construct the ATE via a weighted average of the ATET and ATEU estimated by balancing covariate values across the threshold. This is because a weighted average of the covariate values of the treated and untreated units in a neighbourhood of the threshold no longer equal the average covariate value which prevailed there prior to exposure, and these differences are related to the outcome.

Moreover, if the nature of the manipulation is 'extreme' in the sense that it is concentrated among a large number of highly idiosyncratic units, then the ATE(T/U) estimated via covariate balancing across the threshold could diverge dramatically from the ATE. This is because the extreme manipulation substantially distorts the conditional distribution of the covariates (thus the outcome) in a neighbourhood of the threshold. Equating the covariate values on one side of the threshold to the very different covariate values that prevail on the other would generate a ATE(T/U) that is very different from the ATE, which might therefore be of little practical usefulness.

My approach avoids these problems by achieving covariate balance at every value of the running variable between those exposed and not exposed to the policy settings of interest. This means that those above and below but within a neighbourhood of the threshold have their covariates balanced seperately from one another. As such, the ability to generate a ATE, or the usefulness of a ATE(T/U) estimated via covariate balancing, do not depend on the nature of the manipulation. And I return the covariate values to what they were prior to manipulation, not the more extreme values of the units now present on the other side of the threshold. This is of great usefulness in many settings.

But of course this comes at a cost. My approach requires a comparison group to be observed. This means that not only must one observe units in the absence of the policy, but the observed covariates must support the manipulation-on-observed-variables assumption, and exposure to the policy settings must be as-good-as-random.

In situations in which neither approach is feasible—perhaps because no covariates are observed for which the manipulation-on-observed-variables assumption is satisfied—some researchers identify and discard the units potentially subject to manipulation, and then extrapolate the relevant conditional expectations in their absence, on the basis of which they estimate a ATE (Barreca et al., 2015). This is known as the 'doughnut-hole' RD method.

The method is reasonably widely used, owing to the frequency with which manipulation is encountered in settings for which an RD design otherwise would be appropriate. Similar methods are used in the public finance literature to estimate the counterfactual density of taxable income when estimating the elasticity of taxable income (Chetty et al., 2011). The doughnut-hole RD method has the potential to work well when the ATE does not vary with the running variable, when the conditional expectations to be estimated are close to linear, or when only a narrow range of the running variable is subject to manipulation. Otherwise, a doughnut-hole estimate of the ATE is likely to be subject to a high degree of uncertainty.

2.4 Conclusion

The RD design offers great potential for credibly uncovering treatment effects in many settings. But a method is only as sound as its strongest assumption. Where precise manipulation of the running variable occurs, and that manipulation is related to the outcome, standard estimation procedures generate bias. Where one is observed, a comparison group for which the discontinuity is absent can serve as an anchor for uncovering the true treatment effect. The method recovers the true treatment effect by balancing covariates across the treatment and comparison groups, just as does matching or weighting in the case of a simple mean. The method relies on observing a suitable comparison group, and a set of variables that adequately predict the counterfactual outcome, conditions that won't be met in all settings.

Figures

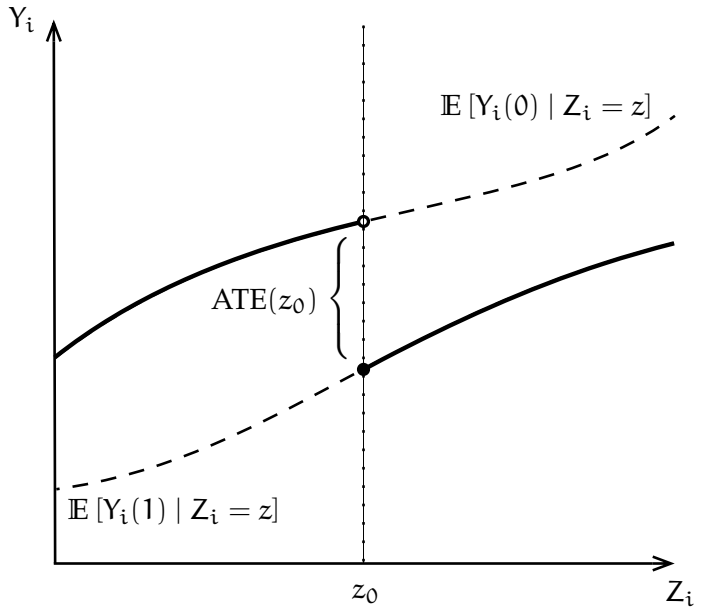


Figure 2.1: *RD design without manipulation.* In bold is the observed conditional expectation, $\mathbb{E}[Y_i \mid Z_i = z]$

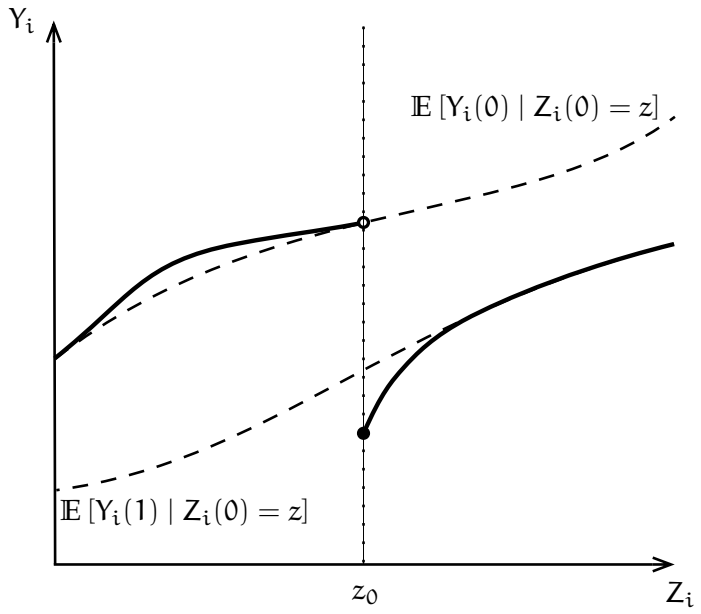


Figure 2.2: *RD design with manipulation in a neighbourhood above but not below the threshold.* In bold is the observed conditional expectation, $\mathbb{E}[Y_i \mid Z_i = z]$

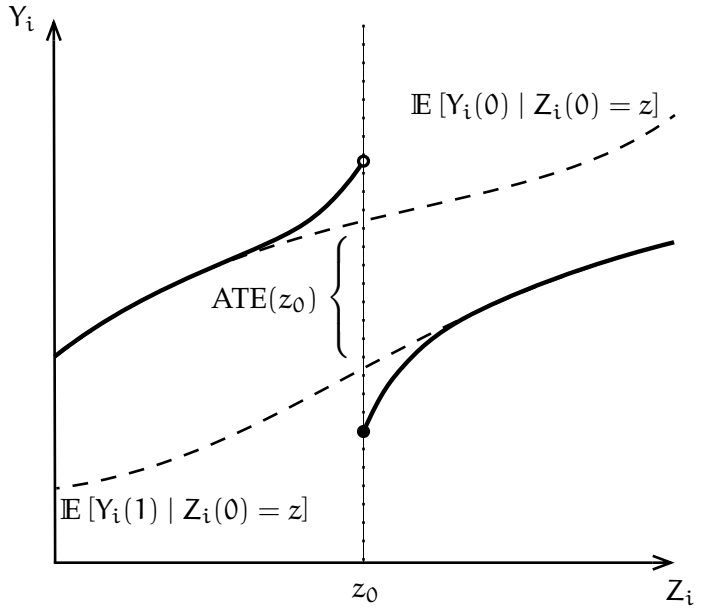


Figure 2.3: *RD design with manipulation in a neighbourhood above and below the threshold.* In bold is the observed conditional expectation, $\mathbb{E}[Y_i \mid Z_i = z]$

Chapter 3

United we evade:
A theory of tax evasion under third-party reporting

3.1 Introduction

Consider three seemingly unrelated observations. A Chinese restaurant in Ann Arbor offers a free egg roll if you don't require a receipt. Over the past few decades, the value-added tax (V.A.T.) has been adopted by governments all over the world, in part because of its desirable enforcement properties. In Scandinavia, rates of evasion of personal income tax have been found to be lower than traditional theories predict; almost non-existent for wage and salary income, and higher only for self-reported income. What do these observations have in common? Third-party reporting. Though a near-universal feature of modern tax systems, there is no canonical theory to explain how third-party reporting works. In this chapter, I offer such a theory.

Taxes are levied typically on the basis of a transaction between parties. To collect the taxes owed on the transaction, the tax authority needs to know its value. It could ask one of the participants to report this, but it cannot be sure the report will be true. It could audit the reporter to try and ascertain the true value, but doing so for every taxable transaction would be prohibitively costly. Auditing a subset of transactions would be less costly, but would tolerate some degree of evasion. Under this traditional model of the tax system, the costliness of audits means that the social optimum will always include some underreporting.

Third-party reporting is said to offer relief from this tradeoff. Under third-party reporting, the tax authority requires both participants in a transaction (and potentially others) to declare the transaction value, and any inconsistency between the reports triggers an audit. The promise is that full compliance will be achieved at a fraction of the cost of the universal audits that otherwise would be required to achieve full compliance. For this reason, third-party reporting is used by tax authorities all over the world. In the U.S.A., for example, all employers engaged in a trade or business who pay remuneration of $600 or more for the year for services performed by an employee must

submit a W-2 form to the Internal Revenue Service (I.R.S.) detailing the employee's income and tax withheld. Employees also must declare this income to the I.R.S. when filing their taxes.

The conventional wisdom is that the advent of third-party reporting has lead to the all-but-elimination of evasion on third-party-reported income, with any income-tax evasion that remains associated with non-third-party-reported income. The conventional wisdom is driven by the essential property of third-party reporting regimes: that non-matching reports can trigger an audit with certain probability, and thus ensure for certain that evasion is detected and penalised. While this is taken for granted, it is in fact not inevitable. This is because ruling out non-matching reports does not rule out tax evasion as two reports can be consistent and false. Only without collusion between the reporters does third-party reporting fully live up to the promise.

Determining the feasibility and extent of collusion is critical to understanding the limits of third-party reporting as an enforcement tool. That is the purpose of this chapter. Under what conditions is evasion feasible in a third-party regime? How is the feasibility of evasion affected by the number of reporters, by the likelihood that collusion will be sustained, and by the number of related transactions subject to third-party reporting?

To answer these questions, I devise a model in which agents decide their underreporting for each taxable transaction on which they have been asked to report. Collusion is feasible when underreporting generates a surplus that can be shared between the reporters. I do not consider explicitly the division of the surplus, assuming simply that it is invariant to the degree of underreporting, which is sufficient for the reporters' optimal decisions to coincide. As such, the agents act to maximise their joint surplus, with the agent responsible for remitting the taxes sharing some sufficient portion of the proceeds.

The key ingredients of the model, therefore, are: 1) the reporters' underreporting costs which embody the essence of a third-party reporting regime (non-matching reports trigger an audit with certainty, resulting in a penalty that exceeds the evaded taxes); and 2) the sidepayment made by the remitter to the third-party reporter(s) so as to induce him (them) to underreport consistently (which must be sufficient to compensate for the cost of underreporting). The agents choose their levels of underreporting given the costs and benefits of doing so. I examine the feasibility of collusion and thus evasion (the extensive margin), determined by the total costs and benefits of underreporting for the reporters jointly. Where a collusive equilibrium is feasible, I examine the reporters' first-order conditions (the intensive margin) to determine the degree of underreporting and how it is affected by various features of the setting.

In the simplest case, underreporting always decreases under third-party reporting, because the remitter must account, at the margin, for the reporting cost of the other reporter when choosing his own level of underreporting. When there is more than one

third-party reporter, the level of underreporting is decreasing in the number of reporters. Third-party reporting can be particularly potent when applied to related transactions; under certain conditions, when a common report is required across transactions, third-party reporting can ensure full compliance. Uncertainty about the reporting costs of the other reporters reduces both the occurrence of underreporting and the level of underreporting when it occurs because agents must be compensated for the possibility that another reporter will report truthfully without their knowledge. When undetected underreporting requires consistent reports in each of a set of related transactions, it is less feasible the greater is the number of related transactions.

There has been only limited theoretical work on third-party reporting, though there has been a recent surge in empirical work. The canonical model of tax evasion of Allingham and Sandmo (1972) does not accommodate third-party reporting, a now-ubiquitous feature of tax systems. Yaniv (1992) was the first to consider formally third-party reporting and the possibility of collusion between reporters, proposing a simple model in which a firm offers to underreport a worker's wages in exchange for the worker receiving less than the free-market wage. Boadway, Marceau and Mongrain (2002) consider a non-cooperative model (in contrast to the cooperative framework I use) of collusive underreporting, focusing on the agents' incentives to cheat in a prisoners' dilemma, and reach the surprising conclusion that greater penalties could increase evasion by enforcing greater cooperation. Chang and Lai (2004) consider the role of social norms as a catalyst in the relationship between penalties and tax evasion. Most recently, Kleven, Kreiner and Saez (2016) focus on the interaction between the ubiquity of large firms and third-party reporting to explain the low observed levels of tax evasion. Using a model of a large firm in which all employees are assumed to collude jointly, they show that the larger the firm, the greater the probability a single employee will 'blow the whistle'.

Relative to the existing theoretical literature, my goal with this chapter is to achieve greater generality so as to accommodate a wider variety of cases. I offer a very simple setup of underreporting costs and sidepayments that can be accommodated in virtually any taxable transaction involving any form of agent. Though it is interesting, I do not consider the bargaining process between the reporters as some of the previous studies do, and instead focus on what the agents do once a bargain is achieved, as this is a more relevant consideration for assessing the efficacy of third-party reporting in enforcing compliance. By starting with a simple model, and gradually incorporating additional complexity, the implications of different features in different settings for evasion under third-party reporting are revealed. In a setup of intermediate complexity, I demonstrate the enforcement properties of the V.A.T. that have formed the basis of some recent empirical studies,[1] and only under the most complex setup do I nest the

[1]These include Pomeranz (2015) and Liu, Lockwood and Almunia (2017). Naritomi (2016) also offers

result of Kleven et al. (2016).

In section 3.2, I present the model, specifying the costs and benefits of underreporting, which drive the results. In section 3.3, I run through the simplest form of the model, in which there are two participants in a single transaction, both of whom may be asked to report and are aware of the other's cost of underreporting (and thus his report). In section 3.4, I extend the basic model to cases in which separate transactions may be linked by a single report, which includes the V.A.T. as an important example. In section 3.5, I introduce uncertainty into the model, both in terms of the other reporters' reporting costs within a transaction as well as other transactions that may be affected by a given report. In section 3.6, I conclude.

3.2 Model

Consider two participants in a taxable transaction, for example, the receipt of labour income in exchange for labour hours, of sales revenue in exchange for goods, or of dividends on the basis of a stockholding. Though a transaction has two participants (employer and employee, buyer and seller, firm and stockholder), a particular agent could participate in several transactions (an employer could have many employees, a buyer many sellers, a firm many stockholders, etc.). A transaction which appears to involve more than two participants can be represented as a set of related two-party transactions.

The transaction value is not observed by the tax authority, but it is observed by the participants, and potentially by others. In order to collect the taxes due, the tax authority attempts to acquire the transaction value (imposes reporting obligations) and selects the agent who is to remit the taxes (imposes a remittance obligation). The focus is on the reporting behaviour of the agents, given assignments of reporting and remittance obligations by the tax authority.

3.2.1 Setup

Agent $i \in \{1, \ldots, N\}$ is involved in transaction $t \in \{1, \ldots, T\}$, where N denotes the number of agents in the economy and T denotes the total number of transactions between all agents. For each transaction in which he participates, the agent could have a remittance obligation, indicated by $\mathbf{M}_{i,t} \in \{0, 1\}$.[2] For each transaction of which he observes the value, the agent could have a reporting obligation, indicated by $\mathbf{P}_{i,t} \in \{0, 1\}$. A remittance obligation must be imposed on one agent per transaction,

a relevant recent empirical study.

[2] All indicator variables are displayed in boldface.

and the remittance amount is implicitly a report by the remitter.[3,4] The value of the transaction is B_t, and the amount of underreporting by agent i is $H_{i,t}$.[5] His report then is $B_t - H_{i,t}$. The remitter must remit to the tax authority $\tau_t(B_t - H_{i,t})$, which is a function of his report. The evaded taxes are given by $E_t(B_t, H_{i,t}) = \tau_t(B_t) - \tau_t(B_t - H_{i,t})$.[6]

The perception that agent i has of the values of the set of tax enforcement instruments available to the tax authority (other than the assignment of remittance and reporting obligations) is described by the vector θ_i. A reporter, whether or not he is also the remitter, bears reporting cost $C_{i,t}(\mathbf{P}, H, \theta_i)$, which could be a function of any report for any transaction, as well as his perception of the tax enforcement settings. In order to underreport, a reporter must be compensated by the remitter, which takes the form of a side payment, $S_{i,t}(\mathbf{P}, H)$.

I assume that the utility the agent derives from what might be called his 'primary' choices (e.g., his labour supply or consumption) does not depend on his degree of underreporting. One can think of the agent as engaging in a two-stage decision process in which he first optimises over all of his choices other than reporting, taking his reporting decisions as given, and then decides his optimal level of underreporting. I assume that the outcome of the second step does not lead him to reassess his choice in the first step. This radically simplifies the analysis, and allows for my setup to be plugged into any generic optimisation procedure. It's important to note that, in reality, the agent's reporting and labour supply decisions, for example, could be interdependent (if underreporting costs take the form of a reporter's time), and the model would need to be modified to accommodate such a case. I also assumed that the reporting costs and sidepayments are assumed to have a dollar cost, and enter the agent's budget constraint.

Given the agent's exogenous income, $Y_i(H_i; \cdot)$, the agent's value function is given

[3]The remittance and reporting obligations are legal obligations, and do not require the tax authority to be aware of the transaction. The government might impose consistent obligations amongst transactions within a particular class (e.g., all employment income must be reported by both the employee and employer, but withheld taxes are remitted by the employer). A transaction which occurs in the shadow economy may be thought of as a transaction in which all reporters decide to fully underreport the value of the transaction. The model does not distinguish between a report of zero and the absence of a report.

[4]The set of: remittance obligations for transaction t is given by $\mathbf{M}_t$; reporting obligations for transaction t by $\mathbf{P}_t$; and reporting obligations for all transactions by $\mathbf{P}$.

[5]The set of underreporting amounts: for transaction t is given by H_t; across all transactions by agent i by H_i; and across all transactions by all agents by H.

[6]For example, under a linear tax with rate ξ_t and agent i the remitter, the tax remitted is $\xi_t \cdot (B_t - H_{i,t})$ and the tax evaded is $\xi_t \cdot B_t - \xi_t \cdot (B_t - H_{i,t}) = \xi_t H_{i,t}$.

by:

$$V_i(H_i;\cdot) = Y_i(H_i;\cdot) - \sum_{t=1}^{T}\Bigg\{\mathbf{M}_{i,t}\Big[\tau_t\left(B_t - H_{i,t}\right) + \sum_{j\neq i}\mathbf{P}_{j,t}S_{j,t}\left(\mathbf{P},H\right)\Big] - \mathbf{P}_{i,t}\Big[\left(1-\mathbf{M}_{i,t}\right)S_{i,t}\left(\mathbf{P},H\right) - C_{i,t}\left(\mathbf{P},H,\theta_i\right)\Big]\Bigg\},$$

which describes the impact on the agent of all transactions for which he is either the remitter or a reporter. This formula includes indicators for reporting and remittance obligations, the tax liability and the sidepayments he must make if a remitter, the sidepayment he receives if only a reporter, and the underreporting cost he bears if he is the remitter or only a reporter. Within the curly braces, the first line describes the impact if the agent is the remitter for transaction t, and the second line describes the impact if he is a reporter for transaction t.

3.2.2 The costs and benefits of (under)reporting

The reporting cost function, $C_{i,t}(\mathbf{P},H,\theta_i)$, plays the leading role in the model, embodying the essence of a third-party reporting regime: that inconsistent reports trigger an audit with certainty, and the consequent penalty exceeds the taxes that otherwise would have been evaded. It is thus through the reporting cost function that the assignment of reporting obligations by the tax authority affects reporting behaviour.

If an agent is assigned a reporting obligation and reports truthfully (that is, $H_{i,t}=0$), then he bears some baseline reporting cost, $\phi_{i,t}$. If he underreports (that is, $H_{i,t}>0$), then the agent's perception of the tax enforcement settings depends on whether there is third-party reporting and, if so, whether there is any inconsistency among the reports. Let Σ_t represent the set of transactions for which the reports affect these perceptions for transaction t.[7] In the simplest case, in which a given transaction is independent of the reports for all others, this set is simply a singleton. This feature will become relevant when the reports for a given transaction could affect whether other transactions are audited as might be the case with a large firm if underreporting is suspected of being systematic. If $H_{i,t}>0$, then the agent's perception of the tax enforcement settings takes the following form:

$$\theta_i = \begin{cases} \tilde{\theta}_i & \text{if} \quad \mathbf{P}_{j,s}=0 \ \forall \ j\neq i \\ \bar{\theta}_i & \text{if} \quad H_{i,s}=\mathbf{P}_{j,s}H_{j,s} \ \forall \ j\neq i \\ \hat{\theta}_i & \text{if} \quad \exists \ j \ \text{ s.t. } \ \mathbf{P}_{j,s}H_{i,s}\neq\mathbf{P}_{j,s}H_{j,s} \end{cases}$$

[7] Σ_t is indexed by the transaction, rather than the agent; the relationship between transactions is a feature of the transactions themselves rather than their participating or reporting agents.

for all $s \in \Sigma_t$, and where: the second line requires that $\mathbf{P}_{j,s} = 1$ for some j and some s, and $H_{i,s} = H_{j,s}$ for all j for whom $\mathbf{P}_{j,t} = 1$, and all s; and the third line that both $\mathbf{P}_{j,t} = 1$ and $H_{i,t} \neq H_{j,t}$ for some j and some s. The first line refers to cases with no third-party reporting, the second to cases with third-party reporting and consistent reports, and the third to cases with third-party reporting and inconsistent reports.

The consistency of the reports affects a reporter's perception of the tax enforcement settings, which in turn affects his reporting cost. If there is third-party reporting and any of the reports for a transaction in the relevant set are inconsistent, then the remitter's reporting cost is assumed to exceed the taxes evaded; that is $C_{i,t}(H_{i,t}, \hat{\theta}_i) > E_t(B_t, H_{i,t})$ for all $H_{i,t} > 0$. Underreporting will only be worthwhile if all reports are consistent, since inconsistent reports indicate with certainty that evasion has occurred, triggering an audit with certainty, which generates a penalty that exceeds the evaded taxes.[8] I also assume that the remitter's reporting cost is higher when the reports are inconsistent than when they are consistent, and higher when they are consistent than when there is no underreporting; that is $C_{i,t}(H_{i,t}, \hat{\theta}_i) > C_{i,t}(H_{i,t}, \bar{\theta}_i) > \phi_{i,t}$ for all $H_{i,t} > 0$. The reporting cost function in each case is assumed to be twice continuously differentiable, increasing, and convex in $H_{i,t}$.

As a reporter faces a strictly positive cost of underreporting, he must be compensated by the remitter in order to underreport. I assume that this takes the form of a set of side payments by the remitter to the reporter(s). A reporter always may choose to report truthfully, so the side payment must leave him no worse off than if he did so, and the same can be said of the remitter. Between these two extremes—either agent being indifferent between participating and not—lies the set of side payments that neither surely will reject, which defines the feasible set of collusive reports.

As the reporting costs depend on whether the agents provide common reports, so too must the side payments. Let agents i be the remitting agents, and t the transaction of interest. Consider the case in which some agent's report differs from that of a remitter; that is $\mathbf{P}_{j,s}H_{i,s} \neq \mathbf{P}_{j,s}H_{j,s}$ for some j and some $s \in \Sigma_t$. When a report differs, since the reporting cost of the remitter exceeds the taxes evaded, there can be no gain from underreporting, so $S_{j,t}(\mathbf{P}, H) = 0$ for all j. Consider the case in which the reports are consistent; that is $H_{i,s} = \mathbf{P}_{j,s}H_{j,s}$ for all $j \neq i$ and all $s \in \Sigma_t$. Agent j in transaction t will participate if $\mathbf{P}_{j,t}S_{j,t}(H_{i,t}) \geqslant C_{j,t}(H_{i,t}, \bar{\theta}_j)$ for some $H_{i,t}$, while agent i in transaction t will participate if $\sum_{j \neq i} \mathbf{P}_{j,t}S_{j,t}(H_{i,t}) \leqslant E_t(B_t, H_{i,t}) - C_{i,t}(H_{i,t}, \bar{\theta}_i)$ for some $H_{i,t}$. That is, each reporter's side payment must exceed his reporting cost, and the total amount paid by the remitter cannot exceed the taxes evaded net of his own reporting costs.

These participation constraints together imply the following condition for underre-

[8] This might not be appropriate for all transactions in all settings. In some settings (e.g., in developing countries), an audit might not be carried out with certainty due to resource constraints, or where one is carried out, it might not uncover the evasion with certainty because of imperfect record keeping. The model would need to be extended to accommodate such cases.

porting to be feasible: $\sum_{j \neq i} \mathbf{P}_{j,t} C_{j,t}(H_{i,t}, \bar{\theta}_j) \leqslant E_t(B_t, H_{i,t}) - C_{i,t}(H_{i,t}, \bar{\theta}_i)$ for some $H_{i,t}$. If this condition is satisfied, then there exists a feasible set of side payments to ensure participation by all reporters.[9] Otherwise, there can be no gain from underreporting, so $S_{j,t}(\mathbf{P}, H) = 0$ for all j. The focus in the various cases is on establishing the feasibility of a set of side payments to support the existence of an equilibrium in which there are consistent reports.

3.3 Base case

Suppose that only the agents involved in a transaction may report on the value of that transaction, and each agent observes the reporting cost of the other. This is the simplest possible case of third-party reporting. Agent i chooses his underreporting for transaction t given the underreporting by agent j, as follows:

$$\max_{H_{i,t}} \; \mathbf{M}_{i,t}\Big[E_t(B_t, H_{i,t}) - \mathbf{P}_{j,t} S_{i,t}(H_{i,t}, H_{j,t})\Big] + \mathbf{P}_{i,t}\left[\mathbf{M}_{j,t} S_{j,t}(H_{i,t}, H_{j,t}) - C_{i,t}(H_{i,t}, H_{j,t}, \mathbf{P}_{j,t}, \theta_i)\right].$$

We are looking for equilibria with underreporting in which each agent chooses his report optimally given the report of the other agent. There are two types of potential equilibria with underreporting: those with consistent and those with inconsistent underreporting. One can immediately rule out inconsistent underreporting; that is, instances in which $H_{i,t} \neq H_{j,t}$, given the specification of the cost function: whenever the reports differ, the reporting costs are sufficiently high so as to render evasion infeasible. Therefore, the only equilibria that exist feature consistent underreporting; that is $H_{i,t} = H_{j,t}$. The remitter receives the surplus from underreporting, and shares some of it with the third-party reporter in the form of a sidepayment, which must be sufficient to cover the underreporting cost of the third-party reporter.

This governs the feasibility condition for underreporting. Whatever surplus is left over after covering the reporting costs of the agents is divided amongst them in some way. The sidepayment specifies the proportion of this surplus that each party receives, and I do not consider the implied bargaining process. When the sidepayment is specified in this way and the agents underreport consistently, the agents each can be seen to choose their underreporting so as to maximise the joint surplus from underreporting. So long as that share does not vary with the level of underreporting, then there will be some level of underreporting that is optimal for both parties, which constitutes the existence of an equilibrium.

Lemma 1. *Suppose that the division of the gains from evasion does not vary with the degree*

[9] At a minimum, $S_{j,t}(H_{i,t}) = C_{j,t}(H_{i,t}, \bar{\theta}_i)$ for all $j \neq i$ is feasible, and agreeable to all agents.

of underreporting. Then there exists a unique equilibrium in which, at the margin, the evaded taxes are equal to the sum of the reporters' costs of evasion.

Proof. Let $\alpha_{i,t}(H_t)$ represent the share received by agent i of the surplus from underreporting , with $\alpha_{i,t}(H_t) + \alpha_{j,t}(H_t) = 1$. Then the optimisation problem of agent i becomes:

$$\max_{H_t} \alpha_t(H_t) \left[E_t(B_t, H_t) - C_{i,t}\left(H_t, \bar{\theta}_i\right) - C_{j,t}\left(H_t, \bar{\theta}_j\right)) \right].$$

As $\alpha_t'(H_t) = 0$, this may further be simplified to:

$$\max_{H_t} E_t(B_t, H_t) - C_{i,t}\left(H_t, \bar{\theta}_i\right) - C_{j,t}\left(H_t, \bar{\theta}_j\right)),$$

with first-order condition:

$$\frac{dE_t}{dH_t}(B_t, H_t(\bar{\theta}) = \frac{dC_{i,t}}{dH_t}\left(H_t(\bar{\theta}), \bar{\theta}_i\right) + \frac{dC_{j,t}}{dH_t}\left(H_t(\bar{\theta}), \bar{\theta}_j\right),$$

provided that $E_t(B_t, H_t(\bar{\theta})) - C_{i,t}(H_t(\bar{\theta}), \bar{\theta}_i) - C_{j,t}(H_t(\bar{\theta}), \bar{\theta}_j) \geqslant 0$, or else $H_t = 0$. As this first-order condition is common to the agents, there exists an equilibrium in which $H_{i,t} = H_{j,t}$.[10] Uniqueness follows from the properties of $E_t(\cdot)$, $C_{i,t}(\cdot)$ and $C_{j,t}(\cdot)$ □

Under the assumption that the division of the gains from underreporting is invariant to the level of underreporting, it is possible to compare the outcomes with and without third-party reporting: it should be seen not as eliminating evasion, but rather as limiting its feasibility, as collusion and thus evasion requires the proceeds to cover the underreporting costs not only of the remitter but also of the third-party reporter. But it is also the case that, where evasion is feasible, third-party reporting lowers the degree of underreporting so long as the marginal cost of underreporting is increasing. At the margin, the agents underreport so as to equalise the benefit (for example, the marginal tax rate) and costs of underreporting borne by both agents. Where the marginal cost of underreporting is increasing, third-party reporting reduces the degree of underreporting.

Proposition 1. *Third-party reporting makes underreporting less feasible, and, under an increasing marginal cost of underreporting, reduces the degree of underreporting where it occurs.*

Proof. First, note the feasibility condition for underreporting:

$$E_t(B_t, H_t(\bar{\theta})) - C_{i,t}(H_t(\bar{\theta}), \bar{\theta}_i) - C_{j,t}(H_t(\bar{\theta}), \bar{\theta}_j) \geqslant 0,$$

[10] Note that, if $\alpha_t(\cdot)$ is continuously differentiable and varies with H_t, then there exists no H_t at which the first-order conditions of both agents are satisfied.

which is tightened by the remitter having also to cover the underreporting cost of the third-party reporter. Then compare the first-order condition of agent i when agent j does not report:

$$\frac{dE_t}{dH_{i,t}}(B_t, H_{i,t}(\tilde{\theta}_i)) = \frac{dC_{i,t}}{dH_{i,t}}\left(H_{i,t}(\tilde{\theta}_i), \tilde{\theta}_i\right),$$

to that when she does report:

$$\frac{dE_t}{dH_{i,t}}(B_t, H_{i,t}(\bar{\theta}_i)) = \frac{dC_{i,t}}{dH_{i,t}}(H_{i,t}(\bar{\theta}_i), \bar{\theta}_i) + \frac{dC_{j,t}}{dH_{i,t}}(H_{i,t}(\bar{\theta}_i), \bar{\theta}_i).$$

As the marginal cost of underreporting is increasing (as the reporting cost function is convex), and provided it is no lower when the second person reports than when she does not (which surely is the case), then the imposition of a reporting obligation on the non-remitting agent reduces underreporting. □

The case just considered allows reports only by the participants in a transaction. In reality, however, others might observe the transaction value, which the tax authority could attempt to obtain from them. For example, the tax authority could require a bank to provide information about a transaction between two agents, or it could offer a reward to whistleblowers aware of tax evasion. In such cases, underreporting, even if otherwise agreeable to both parties, could be undermined by the threat of a report by an outsider.

It is straightforward to extend the setup just considered to accommodate such cases. Agent i chooses an amount to underreport for transaction t, given the remittance and reporting obligations assigned by the tax authority, and the level of underreporting chosen by all other reporting agents j:

$$\begin{aligned}\max_{H_{i,t}} \; & \mathbf{M}_{i,t}\left[E_t(B_t, H_{i,t}) - \sum_{j \neq i} \mathbf{P}_{j,t} S_{i,t}(\mathbf{P}_t H_t)\right] \\ & + \mathbf{P}_{i,t}\left[(1 - \mathbf{M}_{i,t}) S_{i,t}(\mathbf{P}_t H_t) - C_{i,t}(H_t, \mathbf{P}_t, \theta_i)\right]\end{aligned}$$

for each transaction t in which agent i is involved, either as a participant or as a reporting non-participant. In deciding how much to underreport, agent i must now take into account not only the costs borne by himself and the participating reporter, but also the costs borne by all other reporters. As such, the level of underreporting is decreasing in the number of reporters.

Corollary 1. *If the marginal cost of underreporting is increasing, then the imposition of a reporting obligation on additional agents reduces the degree of underreporting where it occurs.*

Proof. Let agent i be the remitter, and j another agent. Consider the optimization

problem of agent i. As before, if any of the reports differ, then no agent will underreport, since the gains from underreporting will be insufficient to cover the costs. Consider instead the case in which all reports are consistent; that is, $H_{i,t} = \mathbf{P}_{j,t} H_{j,t}$ for all $j \neq i$. Then the optimization problem of agent i becomes:

$$\begin{aligned}
&\max_{H_{i,t}} \; E_t(B_t, H_{i,t}) - \sum_{j \neq i} \mathbf{P}_{j,t} S_{i,t}(\mathbf{P}_t H_t) - C_{i,t}(H_{i,t}, \bar{\theta}_i) \\
= &\max_{H_{i,t}} \; E_t(B_t, H_{i,t}) - C_{i,t}(H_{i,t}, \bar{\theta}_i) - \sum_{j \neq i} \mathbf{P}_{j,t} \Big\{ \alpha_{j,t}(H_{i,t}) C_{j,t}(H_{i,t}, \bar{\theta}_j) \\
&\qquad + (1 - \alpha_{j,t}(H_{i,t})) \Big[E_t(B_t, H_{i,t}) - \sum_{k \neq j} \mathbf{P}_{k,t} C_{k,t}(H_{i,t}, \bar{\theta}_k) \Big] \Big\}.
\end{aligned}$$

Analogous to the earlier case, this is equivalent to:

$$\max_{H_{i,t}} \; \Big(1 - \sum_{j \neq i} \mathbf{P}_{j,t} (1 - \alpha_{j,t}(H_{i,t})) \Big) \Big[E_t(B_t, H_{i,t}) - \sum_{k=1}^{N} \mathbf{P}_{k,t} C_{k,t}(H_{i,t}, \bar{\theta}_k) \Big],$$

where, again, the term in the square brackets represents the gains available from underreporting net of the agents' underreporting costs, and the term in the round brackets represents the division of these gains. The structure of the problem is thus unchanged.

Consider the effect on the value function of agent i of a marginal change in underreporting for transaction t:

$$\begin{aligned}
\frac{dV_i}{dH_{i,t}}(H_{i,t}; \cdot) = &\sum_{j \neq i} \frac{d\alpha_{j,t}}{dH_{i,t}}(H_{i,t}) \Big[E_t(B_t, H_{i,t}) - \sum_{k=1}^{N} \mathbf{P}_{k,t} C_{k,t}(H_{i,t}, \bar{\theta}_k) \Big] \\
&+ \Big(1 - \sum_{j \neq i} \mathbf{P}_{j,t} (1 - \alpha_{j,t}(H_{i,t})) \Big) \Big[\frac{dE_t}{dH_{i,t}}(B_t, H_{i,t}) - \sum_{k=1}^{N} \mathbf{P}_{k,t} \frac{dC_{k,t}}{dH_{i,t}}(H_{i,t}, \bar{\theta}_k) \Big].
\end{aligned}$$

By assuming, again, that the division of the gains is invariant to the level of underreporting, the solution simplifies considerably:

$$\begin{aligned}
\frac{dE_t}{dH_{i,t}}(B_t, H_{i,t}(\bar{\theta}_i)) &= \sum_{k=1}^{N} \mathbf{P}_{k,t} \frac{dC_{k,t}}{dH_{i,t}}(H_{i,t}(\bar{\theta}_k)), \bar{\theta}_k) \\
&= \frac{dC_{i,t}}{dH_{i,t}}(H_{i,t}(\bar{\theta}_i)), \bar{\theta}_i) + \frac{dC_{j,t}}{dH_{i,t}}(H_{i,t}(\bar{\theta}_j)), \bar{\theta}_j) \\
&\qquad + \sum_{\substack{k \neq i \\ k \neq j}} \mathbf{P}_{k,t} \frac{dC_{k,t}}{dH_{i,t}}(H_{i,t}(\bar{\theta}_k)), \bar{\theta}_k),
\end{aligned}$$

in which $H_{i,t}(\bar{\theta}_i)$ denotes the agent's optimal reporting function, and i, j, and k denote the remitter, participating reporter and non-participating reporters. This is identical to the earlier first-order condition, but for the addition of the marginal costs of the additional reporters. □

3.4 Multiple transactions with a common report

In the cases considered so far, transactions are independent from one another. I now consider a case in which an activity enters more than one taxable transaction. In such a case, third-party reporting will link together different transactions and the reporters for those transactions, which will lend greater potency to third-party reporting as an enforcement tool. It is typical for the agent to be a remitter for one of the transactions and a reporter for the other/s. The tax authority thus requires the agent's report to be consistent across transactions, perhaps mechanically via the agent making only a single report for the two transactions. In my framework, the agent's reporting decision for one transaction affects his report—and the available surplus—for another.

Wage income is a good example, in which a wage payment is both a deduction from income tax paid by the firm and taxable income to the worker. The tax authority requires the firm to report the worker's wages for tax withholding and to verify the wages that the worker reports when filing his taxes. But this report could also serve as the firm's reported wage deduction when calculating its income tax liability. Another example is the invoice-based Value-Added Tax (V.A.T.), under which a seller provides a buyer with an invoice for the taxes remitted for a transaction, with which the buyer can claim a credit when remitting taxes for a transaction subsequent in the value chain. A single report—embodied in the invoice—serves both as a third-party report for the earlier transaction and a report for claiming a credit for a latter transaction.

Such relationships between transactions can amplify the effectiveness of third-party reporting. In the simpler case considered earlier, the effectiveness of third-party reporting derives from its connecting agents together. In this case, connecting agents together means connecting transactions together. They key, of course, is to target linked transactions in which the common activity has opposing effects on the agents across transactions. This means enforcement is strengthened not only for a particular transaction but also for any linked transactions. As transactions are linked by third-party reporting, the agent must consider simultaneously the effect of his underreporting on the surplus generated in both transactions.

Consider an agent who must provide a common report for two transactions in which he is involved (I will refer to him as the 'remitter'). Suppose that the agent is a remitter for one of the transactions, and a reporter for the other. And, further, that the higher is the agent's report, the lower is his remittance (because, for example, the report

serves as a deduction when calculating his income-tax liability). Then the remitter chooses his report as follows:

$$\max_{H_i} \; E_u(B_u, H_i) - \mathbf{P}_{j,u} S_{j,u}(H_i, H_{j,u}) - C_{i,u}(H_i, H_{j,t}, \theta_i) \\ + S_{i,v}(H_i, H_{k,v}) - C_{i,v}(H_i, H_{k,v}, \theta_i),$$

which can be rewritten as:

$$\max_{H_i} \; \alpha_{i,u}(H_i) \left[E_u(B_u, H_i) - C_{i,u}(H_i, \bar{\theta}_i) - C_{j,u}(H_i, \bar{\theta}_j) \right] \\ + \alpha_{i,v}(H_i) \left[E_v(B_v, H_i) - C_{i,v}(H_i, \bar{\theta}_i) - C_{k,v}(H_i, \bar{\theta}_k) \right],$$

where the first line refers to the transaction for which the agent is a reporter and the second a remitter.

Lemma 2. *When an agent provides a common report across two transactions, and the reports have opposing effects on tax liability, gains from underreporting in one transaction must at least offset losses in the other if evasion is to be feasible.*

Proof. The first-order condition is as follows:

$$\alpha_{i,u} \left[\frac{dE_u}{dH_i}(B_u, H_i) - \frac{dC_{i,u}}{dH_i}(H_i, \bar{\theta}_i) - \frac{dC_{j,u}}{dH_i}(H_i, \bar{\theta}_j) \right] \\ + \alpha_{i,v} \left[\frac{dE_v}{dH_i}(B_v, H_i) - \frac{dC_{i,v}}{dH_i}(H_i, \bar{\theta}_i) - \frac{dC_{k,v}}{dH_i}(H_i, \bar{\theta}_k) \right] = 0,$$

which may be rewritten as:

$$\alpha_{i,u} \left[\frac{dE_u}{dH_i}(B_u, H_i) \right] + \alpha_{i,v} \left[\frac{dE_v}{dH_i}(B_v, H_i) \right] \\ = \alpha_{i,u} \left[\frac{dC_{i,u}}{dH_i}(H_i, \bar{\theta}_i) + \frac{dC_{j,u}}{dH_i}(H_i, \bar{\theta}_j) \right] + \alpha_{i,v} \left[\frac{dC_{i,v}}{dH_i}(H_i, \bar{\theta}_i) + \frac{dC_{k,v}}{dH_i}(H_i, \bar{\theta}_k) \right],$$

where the term on the left-hand side describes the marginal benefits of underreporting, and the term on the right-hand side the marginal costs. As $dE_v/dH_i(B_v, H_i) < 0$, and the terms on the right-hand side are greater than zero, a positive level of underreporting requires that the first term exceeds the second term. Formally, the feasibility condition is:

$$\alpha_{i,u} \left[\frac{dE_u}{dH_i}(B_u, H_i) \right] > \alpha_{i,v} \left[\frac{dE_v}{dH_i}(B_v, H_i) \right].$$

□

There are two cases in which the necessary conditions for underreporting will be

met when transactions are linked in this way. The first is, given the remitter's share of the proceeds, that the tax rate on the transaction for which he is a remitter is less than that on the transaction for which he is a reporter. The second is, given the two tax rates, that the remitter's share of the surplus is greater for the transaction in which he is a reporter than that for the transaction in which he is a remitter.

In either case, there will be a benefit to the remitter of underreporting, as the loss he incurs due to a reduced deduction will be more than offset by the surplus generated by underreporting in the other transaction. Of course, this is a necessary, rather than sufficient, condition for underreporting—the surplus generated must also be sufficient to cover the reporting costs of both agents across both transactions. This highlights cases in which we definitively can rule out underreporting, which offers a guide for imposing third-party reporting obligations so as to ensure compliance.

Proposition 2. *Suppose an agent must provide a common report for two transactions, where greater underreporting reduces the tax liability for one transaction and increases it for the other. Further, suppose that neither the tax rate nor the agent's share of the proceeds is greater in the latter than the former. Then third-party reporting eliminates tax evasion.*

Proof. This follows immediately from lemma 2. □

While I earlier established that the imposition of third-party reporting obligations can reduce tax evasion, I have now identified settings—corresponding to real-world scenarios—in which they do, in fact, ensure perfect compliance. This illustrates the desirability of imposing third-party reporting obligations in situations in which agents' actions are common to multiple taxable transactions, and of structuring agents' tax liabilities in this way given the ability to impose third-party reporting obligations.

This explains the well-known enforcement property of the V.A.T., as described by Pomeranz (2015). Underreporting at one link in the chain is offset completely by a lower credit at the subsequent link, deeming evasion under a V.A.T. infeasible. And the only point in the chain at which evasion is feasible is the final one, where the consumer is unconstrained as his purchase does not serve as a credit for a later transaction. While this has been attributed to the "paper trail" created by the V.A.T., the present analysis points to the powerful incentives engendered in a third-party reporting regime.

3.4.1 Application

Consider the reporting decisions of the members of a supply chain under an invoice-based V.A.T..[11] Suppose that there are three firms (a producer, a wholesaler and a re-

[11]These invoices are not a feature of an alternative, accounts-based V.A.T., in which reporting obligations are only imposed on the remitter. All countries with a V.A.T. other than Japan use the invoice-based method, and thus it serves as our application.

tailer) and a consumer involved in three transactions (producer–wholesaler, wholesaler–retailer and retailer–consumer).

Under a V.A.T., each member of the supply chain remits taxes to the government on the basis of the value of production net of non-labour input costs. One agent's input costs are another agent's value of production. In order to deduct from taxes his input costs, an agent must declare them to the tax authority via an invoice provided by the seller of those inputs. A V.A.T. effectively is a chain of remittance and reporting obligations.

Agent 1 sells z_1 to agent 2 at price q_1; agent 2 uses production function $f_2(\cdot)$ to transform z_1 into z_2, which it sells to agent 3 at price q_2; agent 3 uses production function $f_3(\cdot)$ to transform z_2 into x, which it sells to agent 4 at price p. For simplicity, there are no labour inputs, nor imports and exports. The V.A.T. rate is τ. Take the quantities of the various inputs and outputs to be the optima of the agents' supply and demand decisions, and the prices those in equilibrium. Figure 3.1 indicates the transactions in which each agent is involved.

i	t		
	1	2	3
1	✓		
2	✓	✓	
3		✓	✓
4			✓

Figure 3.1: Involvement of each agent i in each transaction t

The producer must report the value of its production and remit V.A.T. accordingly. There are no input costs, so the producer claims no credits when calculating its V.A.T. obligation. The producer sells its output to a wholesaler, which then must report the value of the transaction when calculating its value added and thus V.A.T. obligation.

Therefore $\mathbf{M}_{1,1} = 1$, $\mathbf{M}_{2,1} = 0$, $\mathbf{P}_{1,1} = 1$ and $\mathbf{P}_{2,1} = 1$. The producer's reporting problem is then:[12]

$$\max_{H_{1,1}} \pi_1 = q_1 z_1 - \tau(q_1 z_1 - H_{1,1}) - S_{1,1}(H_{1,1}, H_{2,1}) - C_{1,1}(H_{1,1}, H_{2,1}, \mathbf{P}_{2,1} = 1, \theta_i),$$

[12] I suppress much of the functional notation for clarity, but recall that the agents are assumed to have optimized with respect to their primary objective, and markets to have cleared in equilibrium.

which is just a special case of the value function introduced earlier with:

$$\begin{aligned} T &= 3 \\ i &= 1 \\ j &= 2 \\ V_i(H_{i,t}, \cdot) &= \pi_1 \\ Y_i(H_{i,t}, \cdot) &= q_1 z_1 \\ B_t &= q_1 z_1, \end{aligned}$$

and the aforementioned indicator values.

The wholesaler transforms the input it buys from the producer into output, which it sells to the retailer. The wholesaler remits V.A.T. on the basis of the value of the transaction less a credit for the value of the prior transaction, which it reports to the tax authority. The tax authority thus receives two reports of the value of the first transaction. The retailer will, in turn, need to report the value of the second transaction in order to claim a credit when remitting the V.A.T. due for the final transaction.

Therefore $\mathbf{M}_{1,1} = 0$, $\mathbf{M}_{2,1} = 0$, $\mathbf{M}_{2,2} = 1$, $\mathbf{M}_{3,2} = 0$, $\mathbf{P}_{2,1} = 1$, $\mathbf{P}_{2,2} = 1$ and $\mathbf{P}_{3,2} = 1$. The wholesaler's reporting problem is then:

$$\begin{aligned} \max_{H_{2,1}, H_{2,2}} \quad & \pi_2 = q_2 f_2(z_1) - q_1 z_1 \\ & + S_{1,1}(H_{1,1}, H_{2,1}) - C_{2,1}(H_{1,1}, H_{2,1}, \mathbf{P}_{1,1} = 1, \theta_i) \\ & - \tau(q_2 f_2(z_1) - (q_1 z_1 - H_{2,1}) - H_{2,2}) - S_{2,2}(H_{2,2}, H_{3,2}) \\ & - C_{2,2}(H_{2,2}, H_{3,2}, \mathbf{P}_{3,2} = 1, \theta_i), \end{aligned}$$

where the second line refers to the first transaction, in which the wholesaler is a reporter but not the remitter, and the third and fourth lines refer to the second transaction, in which the wholesaler is the remitter (and thus also a reporter).

Relabelling the reporting cost function for the second transaction, separating that credit and rearranging yields:

$$\begin{aligned} \max_{H_{2,1}, H_{2,2}} \quad & \pi_2 = q_2 f_2(z_1) - q_1 z_1 \\ & + S_{1,1}(H_{1,1}, H_{2,1}) - \tilde{C}_{2,1}(H_{1,1}, H_{2,1}, \mathbf{P}_{1,1} = 1, \theta_i) - \tau H_{2,1} \\ & - \tau(q_2 f_2(z_1) - q_1 z_1 - H_{2,2}) - S_{2,2}(H_{2,2}, H_{3,2}) \\ & - C_{2,2}(H_{2,2}, H_{3,2}, \mathbf{P}_{3,2} = 1, \theta_i), \end{aligned}$$

then this is just a special case of the value function introduced earlier with:

$$\begin{aligned}
T &= 3 \\
i &= 2 \\
j &= 1 \text{ if } t = 1, \text{and} \\
j &= 3 \text{ if } t = 2 \\
V_i(H_{i,t}, \cdot) &= \pi_2 \\
Y_i(H_{i,t}, \cdot) &= q_2 f_2(z_1) - q_1 z_1 \\
B_t &= q_2 f_2(z_1) - q_1 z_1 \\
C_{i,1}(H_{i,1}, H_{j,1}, \mathbf{P}_{j,1}, \theta_i) &= \tilde{C}_{2,1}(H_{1,1}, H_{2,1}, \mathbf{P}_{1,1} = 1, \theta_i) + \tau H_{2,1},
\end{aligned}$$

and the aforementioned indicator values. The final line is crucial—the wholesaler bears an additional cost of underreporting for the first transaction, since doing so limits the credit it can claim when calculating the V.A.T. it must remit for the second transaction. Indeed, this is the linchpin of the V.A.T. system as a tax enforcement mechanism.

While the retailer must report the value of the second transaction in order to claim a credit for it when calculating the V.A.T. to be remitted for the third transaction, the consumer does not report the value of the third transaction. Therefore $\mathbf{M}_{2,2} = 1$, $\mathbf{M}_{3,2} = 0$, $\mathbf{M}_{3,3} = 1$, $\mathbf{M}_{4,3} = 0$, $\mathbf{P}_{2,2} = 1$, $\mathbf{P}_{3,2} = 1$, $\mathbf{P}_{3,3} = 1$ and $\mathbf{P}_{4,3} = 0$. The retailer's reporting problem is then:

$$\begin{aligned}
\max_{H_{3,2}, H_{3,3}} \quad \pi_3 &= q_3 f_3(z_2) - q_2 z_2 \\
&\quad + S_{2,2}(H_{2,2}, H_{3,2}) - C_{3,2}(H_{2,2}, H_{3,2}, \mathbf{P}_{2,2} = 1, \theta_i) \\
&\quad - \tau(q_3 f_3(z_2) - (q_2 z_2 - H_{3,2}) - H_{3,3}) - C_{3,3}(H_{3,3}, \theta_i),
\end{aligned}$$

where the second line refers to the second transaction, in which the retailer is a reporter but not the remitter, and the third line refers to the third transaction, in which the retailer is the remitter (and thus also a reporter).

After performing the same rearrangement as in the case of the wholesaler:

$$\begin{aligned}
\max_{H_{3,2}, H_{3,3}} \quad \pi_3 &= q_3 f_3(z_2) - q_2 z_2 \\
&\quad + S_{2,2}(H_{2,2}, H_{3,2}) - \tilde{C}_{3,2}(H_{2,2}, H_{3,2}, \mathbf{P}_{2,2} = 1, \theta_i) - \tau H_{3,2} \\
&\quad - \tau(q_3 f_3(z_2) - q_2 z_2 - H_{3,3}) - C_{3,3}(H_{3,3}, \theta_i),
\end{aligned}$$

then this is just a special case of the value function introduced earlier with:

$$\begin{aligned}
T &= 3 \\
i &= 3 \\
j &= 2 \text{ if } t = 2, \text{and} \\
j &= 4 \text{ if } t = 3 \\
V_i(H_{i,t}, \cdot) &= \pi_3 \\
Y_i(H_{i,t}, \cdot) &= q_3 f_3(z_2) - q_2 z_2 \\
B_t &= q_3 f_3(z_2) - q_2 z_2 \\
C_{i,2}(H_{i,2}, H_{j,2}, \mathbf{P}_{j,2}, \theta_i) &= \tilde{C}_{3,2}(H_{2,2}, H_{3,2}, \mathbf{P}_{2,2} = 1, \theta_i) + \tau H_{3,2},
\end{aligned}$$

where the reporting cost function for the second transaction may be interpreted equivalently to that in the case of the wholesaler.

Recall that, in order for the agents to underreport, they must underreport equally, and it must be the case that $\tau_t H_{i,t}(\bar{\theta}_i) - C_{i,t}(H_{i,t}(\bar{\theta}_i), \bar{\theta}_i) - C_{j,t}(H_{i,t}(\bar{\theta}_i), \bar{\theta}_i) \geqslant 0$. But one can see, if we let $t = 1$, $i = 1$ and $j = 2$, that:

$$\begin{aligned}
\tau H_{1,1}(\bar{\theta}_i) - C_{1,1}(H_{1,1}(\bar{\theta}_i), \bar{\theta}_i) - C_{2,1}(H_{1,1}(\bar{\theta}_i), \bar{\theta}_i) &\geqslant 0 \\
\tau H_{1,1}(\bar{\theta}_i) - C_{1,1}(H_{1,1}(\bar{\theta}_i), \bar{\theta}_i) - \tilde{C}_{2,1}(H_{1,1}(\bar{\theta}_i), \bar{\theta}_i) - \tau H_{1,1}(\bar{\theta}_i) &\geqslant 0 \\
-C_{1,1}(H_{1,1}(\bar{\theta}_i), \bar{\theta}_i) - \tilde{C}_{2,1}(H_{1,1}(\bar{\theta}_i), \bar{\theta}_i) &\ngeqslant 0,
\end{aligned}$$

since the agents face a strictly positive reporting cost when underreporting. The argument follows equivalently for the second transaction. This is the crux of the V.A.T. as a tax enforcement mechanism—there can be no net gain from underreporting, since any evaded taxes will be fully offset by reduced credits. A V.A.T. is, in essence, a system of reporting obligations assigned in a way so as to eliminate underreporting.

The third transaction, however, demonstrates the hole in the V.A.T. enforcement system: since the consumer is not required to report the value of the transaction in order to obtain a credit for a subsequent transaction (as there is no subsequent transaction), the retailer is free to underreport. The third transaction corresponds to the simplest general case considered earlier with no third-party reporting in which the agent simply weighs the costs and benefits of underreporting at the margin when deciding how much to underreport. Imposing a reporting obligation on the consumer would reduce underreporting, though one should consider the reporting costs incurred under such an arrangement.

Naritomi (2016) considers the case of Brazil, where consumers are given an incentive to report the value of transactions subject to V.A.T.. The Brazilian tax authority runs a lottery offering consumers the chance to win a prize in return for reporting the value

of their retail purchases. The model presented here predicts that such an arrangement will reduce, but not necessarily eliminate, underreporting, since conditions could exist for the retailer and consumer to jointly underreport. For example, the consumer could agree to receive a falsified receipt in exchange for a discount on the purchase price. Whether this will occur depends on the reporting costs of the agents. While evasion at earlier stages can be ruled out, the V.A.T. tax enforcement chain will inevitably be broken at the final stage, since the consumer, as the final link in the chain, will always have less 'skin in the game' than those involved at a preceding stage.

3.5 Uncertainty about other reporters

I have so far assumed implicitly that the reporters observe the reporting costs of one another. This generates the somewhat unrealistic result that no equilibrium exists in which one agent underreports while another tells the truth—all of the agents either underreport consistently or report truthfully. If this were true, a third-party reporting regime would never uncover evasion.

But this is artificial. Suppose, for instance, that a reporter is offered a reward by the tax authority in exchange for a truthful report. The reward could be represented by an increased underreporting cost when underreporting consistently, since doing so would entail his foregoing the reward. The other reporters must make their underreporting decisions without observing the true reporting cost of the reporter offered the reward. Instead, they must consider this cost probabilistically, which will affect their reporting decision. In order to accommodate cases such as this, the model must accommodate equilibria in which reporters provide inconsistent reports to the tax authority.

3.5.1 Unobservable reporting costs within a transaction

Consider agent j, who participates in transaction t, for which agent i is the remitter, and let $\mathbf{P}_{k,t} = 1$ for some $k \neq j$, and continue to assume that $\Sigma_t = \{t\}$. Recall that if $E_t(B_t, H_{i,t}) \geqslant \sum_{k=1}^{N} \mathbf{P}_{k,t} C_{k,t}(H_{i,t}, \bar{\theta}_k)$, then there exists an $S_{k,t}(H_{i,t})$ for all k such that $H_{i,t} = \mathbf{P}_{k,t} H_{k,t}$ for all $k \neq i$. Each agent will consistently underreport, since he knows for sure that his side payment will exceed his reporting cost; he knows that this is feasible, since he knows that the proceeds of the underreporting are sufficient to cover the reporting costs of all reporters. If the proceeds of the underreporting are insufficient to cover all reporting costs, then none of the agents will underreport, since at least one report will be inconsistent.

When an agent does not observe the reporting costs of the other reporters, so does not know whether the other reports will be consistent, he must choose his level of underreporting under uncertainty. Assume that agent j observes the reporting cost he

will face if he reports truthfully, as well as that if he underreports and there is either consistent or inconsistent reporting by the other reporters; that is, he observes $\phi_{j,t}$, $C_{j,t}(H_{i,t}, \bar{\theta}_j)$, and $C_{j,t}(H_{i,t}, \hat{\theta}_j)$. He also observes $\mathbf{P}_{k,t}$ for all k, and $E_t(B_t, H_{i,t})$. Assume that the agent knows, if he is to underreport, the report that collectively is optimal for all reporters; that is $H_{i,t}(\bar{\theta}_i)$.[13] The agent does not observe $C_{k,t}(\mathbf{P}, H, \theta_k)$ for any $k \neq j$, so does not know whether $E_t(B_t, H_{i,t}) \geqslant \sum_{k=1}^{N} \mathbf{P}_{k,t} C_{k,t}(H_{i,t}, \bar{\theta}_k)$.

The agent must commit to a level of underreporting without knowing for sure the cost function or payoff he will face *ex post*, as these depend on the consistency of the reports. Only once all agents report will each observe the outcome of his reporting decision. For agent j, the subjective probability that the reports for transaction t are consistent is $p_{j,t} \equiv \mathbb{P}_j[H_{i,t} = \mathbf{P}_{k,t} H_{k,t} \ \forall \ k \neq i]$ and that there is an inconsistent report for transaction t is $1 - p_{j,t} \equiv \mathbb{P}_j[\exists \ k \text{ s.t. } \mathbf{P}_{k,t} H_{i,t} \neq \mathbf{P}_{k,t} H_{k,t}]$.

Lemma 3. *When agents do not observe the reporting costs of one another, but instead perceive some probability of inconsistent underreporting, the feasibility condition takes the following form:*

$$E_t(B_t, H_{i,t}) \geqslant \sum_{j=1}^{N} \mathbf{P}_{j,t} \frac{p_{i,t}}{p_{j,t}} \underset{j}{\mathbb{E}} \left[C_{j,t}(H_{i,t}, \theta_j) \right],$$

Proof. The agent's expected reporting cost is:

$$\underset{j}{\mathbb{E}}[C_{j,t}(H_{i,t}, \theta_j)] = p_{j,t} C_{j,t}(H_{i,t}, \bar{\theta}_j) + (1 - p_{j,t}) C_{j,t}(H, \hat{\theta}_j),$$

as long as $H_{j,t} > 0$.[14] If agent j is not the remitter, then he will compare this with his

[13] In order to observe the optimal level of underreporting, the agent probably would need also to observe the reporting costs of the other reporters, which has been ruled out. This suggests that there could be a coordination problem in practice, since the agents do not observe the information necessary to select the jointly optimal level of underreporting. The focus is on establishing the existence, rather than inevitability or even likelihood, of an equilibrium with consistent reports. Such questions are left for future work.

[14] Otherwise $\mathbb{E}_j[C_{j,t}(H_{i,t}, \theta_j)] = \phi_{j,t}$, as before.

expected side payment, which is:[15]

$$\begin{aligned}\mathop{\mathbb{E}}_{j}[S_{j,t}(H_{i,t})] = p_{j,t}\Bigg\{ & \alpha_{j,t}(H_{i,t})C_{j,t}(H_{i,t},\bar{\theta}_j) \\ & + \left(1-\alpha_{j,t}(H_{i,t})\right)\left[E_t(B_t,H_{i,t}) - \sum_{k\neq j}\mathbf{P}_{k,t}C_{k,t}(H_{i,t},\bar{\theta}_k)\right]\Bigg\} \\ & + (1-p_{j,t})\cdot 0 \\ \equiv\ & p_{j,t}S_{j,t}(H_{i,t}),\end{aligned}$$

for ease of exposition. If agent j is the remitter, then he will instead compare it with the expected side payments he must make, which are:

$$\begin{aligned}\mathop{\mathbb{E}}_{i}\left[\sum_{j\neq i}\mathbf{P}_{i,t}S_{j,t}(H_{i,t})\right] &= p_{i,t}\sum_{j\neq i}\mathbf{P}_{i,t}S_{j,t}(H_{i,t}) + (1-p_{i,t})\cdot 0 \\ &= p_{i,t}\sum_{j\neq i}\mathbf{P}_{i,t}S_{j,t}(H_{i,t}).\end{aligned}$$

A reporter will underreport if and only if $\mathbb{E}_j[S_{j,t}(H_{i,t})] \geqslant \mathbb{E}_j[C_{j,t}(H_{i,t},\theta_j)]$
$\Rightarrow \mathbb{E}_j[\sum_{j\neq i} S_{j,t}(H_{i,t})] \geqslant \mathbb{E}_j[\sum_{j\neq i} C_{j,t}(H_{i,t},\theta_j)]$, and the remitter will underreport if and only if $E_t(B_t,H_{i,t}) \geqslant \mathbb{E}_i[\sum_{j\neq i}\mathbf{P}_{j,t}S_{j,t}(H_{i,t})] + \mathbb{E}_i[C_{i,t}(H_{i,t},\theta_i]$. □

Note the ratio of probabilities on the right hand side of the inequality in lemma 3: the numerator deflates the feasibility threshold, as there is now some probability that the remitter will not have to make the sidepayments; the denominator inflates the feasibility threshold, as there is now some probability that the reporters will not receive a side payment. If all reporters expect consistent reporting to occur with equal probability, then the two effects cancel out.

Proposition 3. *If the remitter does not consider consistent reporting any less likely than do the reporters, then the threshold for consistent underreporting is always higher under uncertainty.*

Proof. The expectation in the feasibility condition in lemma 3 assigns some positive probability to the reporting cost under inconsistent reports, $C_{j,t}(H_{j,t},\hat{\theta}_j)$, and by assumption $C_{j,t}(H_{j,t},\hat{\theta}_j) > C_{j,t}(H_{j,t},\bar{\theta}_j)$ for all j and t. □

[15] To clarify, $S_{j,t}(H_{i,t})$ is the amount offered by the remitter to agent j, *ex ante*, in exchange for his consistent report. When making his choice, the agent knows he will receive this side payment for sure if all reports (including his own) are consistent, and will otherwise receive nothing. Conditional on providing a consistent report, agent j expects to receive this side payment with probability $p_{j,t}$. I do not consider how the remitter knows what side payment to offer without observing the agents' costs; again, the concern solely is in establishing the feasibility of such a side payment, and thus the existence of an equilibrium with consistent reports.

One can say that undetected evasion is less likely under uncertainty in the sense that some cases of undetected evasion, which would be feasible under certainty, are infeasible under uncertainty. This is because each reporter demands a larger side payment to compensate for the possibility that his underreporting will be detected due to inconsistent reports. However, one can say nothing about inconsistent underreporting. Inconsistent reporting never occurs under certainty. Cases which might result in truthful reporting under certainty, but inconsistent underreporting under uncertainty, cannot be ruled out.

If the feasibility condition is sufficient to ensure consistent reports, then $p_{j,t}$ represents, from the perspective of agent j, the probability that the condition is satisfied. The feasibility condition is necessary for the existence of a set of side payments which can support consistent underreporting. No claims may be made about whether such an equilibrium is inevitable or even likely, or how it might be reached. Such concerns may be reflected in $p_{j,t}$, since it represents the probability that there are consistent reports, rather than that the feasibility condition is satisfied, which is implied by the former.

While uncertainty enhances the efficacy of third-party reporting in making evasion less feasible, it also reduces the degree of underreporting where it remains feasible. This is because the risk that another agent will report truthfully raises the marginal cost of underreporting, which leads the agents to underreport less.

Proposition 4. *If the remitter does not consider consistent reporting any less likely than do the reporters, and the marginal reporting cost is higher when the reports are inconsistent than when they are consistent, then the level of underreporting, when it occurs, is lower under uncertainty.*

Proof. Consider the optimal level of underreporting chosen under uncertainty. Analogous to the earlier cases, the optimisation problem of agent i is:

$$\max_{H_{i,t}} \left(1 - \sum_{j \neq i} \mathbf{P}_{j,t}(1 - \alpha_{j,t}(H_{i,t}))\right) \left[E_t(B_t, H_{i,t}) - \sum_{j=1}^{N} \mathbf{P}_{j,t} \frac{p_{i,t}}{p_{j,t}} \mathbb{E}_j \left[C_{j,t}(H_{i,t}, \theta_j)\right]\right].$$

Assuming, as before, that the division of the gains from underreporting is invariant to the level of underreporting, the solution is characterised by:

$$\begin{aligned}
\frac{dE_t}{dH_{i,t}}(B_t, H_{i,t}(\theta_i)) &= \sum_{j=1}^{N} \mathbf{P}_{j,t} \frac{p_{i,t}}{p_{j,t}} \frac{d\,\mathbb{E}_j \left[C_{j,t}(H_{i,t}(\theta_j)), \theta_j)\right]}{dH_{i,t}} \\
&= \sum_{j=1}^{N} \mathbf{P}_{j,t} \frac{p_{i,t}}{p_{j,t}} \left(p_{j,t} \frac{dC_{j,t}}{dH_{i,t}}(H_{i,t}(\theta_j)), \bar{\theta}_j) \right. \\
&\qquad \left. + (1 - p_{j,t}) \frac{dC_{j,t}}{dH_{i,t}}(H_{i,t}(\theta_j)), \hat{\theta}_j) \right).
\end{aligned}$$

Given the convexity of the reporting cost function, the proposition follows immediately. □

3.5.2 Unobservable reporting costs between transactions

Now consider the case in which reporting costs for a given transaction may be affected by the reports for another transaction. I have in mind situations in which a number of transactions are related such that the reports within each of the transactions must be consistent if underreporting in any of the transactions is to be feasible. It is typical for some agents to be engaged in numerous transactions; for example, a firm pays wages and salaries to all of its workers. Such transactions tend to be centrally administered, and each employee could reasonably be aware of the transactions involving the other employees. One could imagine that the detection of evasion by the tax authority for one such transaction might trigger audits and subsequent detection of tax evasion for the others.

Consider agent j, who participates in transaction t, which is one of a set of related transactions, Σ_t. Recalling the cost function introduced earlier, an agent's reporting cost depends on all reports for all transactions in this set; if there are inconsistent reports for any transaction, then every reporter for every transaction in the set will face a discrete increase in his reporting cost if he underreports. Continue to assume that an agent does not observe the reporting costs of the other reporters for transaction t, but now consider the implications for his not observing the reporting costs of reporters for any of the set of related transactions. The goal is to consider how the number of transactions in such a circumstance affects his reporting behavior.

The agent continues to compare his expected reporting cost, $\mathbb{E}_j\left[C_{j,t}(H_{i,t}, \theta_j)\right]$, with his expected side payment, $\mathbb{E}_j\left[S_{j,t}(H_{i,t})\right]$, in the case of a non-remitter, or his expectation of the side payments he must make, $\mathbb{E}_i\left[\sum_{j\neq i} \mathbf{P}_{i,t} S_{j,t}(H_{i,t})\right]$, in the case of the remitter. Recall that these expectations were based on the subjective probability of consistent reports, which continues to be the case. Now, however, this probability must account for the uncertainty of reporting across all of the related transactions, rather than simply the transaction for which the agent reports. Recall that $p_{j,s}$ refers to the probability assigned by agent j to the consistency of reports for transaction s; that is $p_{j,s} \equiv \mathbb{P}_j[H_{k,s} = \mathbf{P}_{l,s} H_{l,s} \ \forall \ l \neq k]$. Without loss of generality, consider the case in which $p_{j,u} = p_{j,v}$ for all $u, v \in \Sigma_t$.[16]

Proposition 5. *When the number of related transactions for which reports are provided to the tax authority becomes arbitrarily large, underreporting is infeasible.*

[16]This assumption may be interpreted as each agent being uninformed about differences in the probabilities of consistent reports between the related transactions; that is, consistent reports are equally as likely in a particular transaction as in any other.

Proof. The expected reporting cost of agent j is given by:

$$\mathop{\mathbb{E}}_{j}\left[C_{j,t}(H_{i,t},\theta_j)\right] = p_{j,t}^{\#\Sigma_t} C_{j,t}(H_{i,t},\bar{\theta}_j) + \left(1 - p_{j,t}^{\#\Sigma_t}\right) C_{j,t}(H,\hat{\theta}_j),$$

where $\#\Sigma_t \in \mathbb{N}$ denotes the cardinality of Σ_t; that is, the number of related transactions which can affect the reporting cost of a reporter for transaction t.[17]

As the number of related transactions becomes arbitrarily large, it becomes all but certain that at least one audit is triggered, making underreporting uneconomical:

$$\lim_{\#\Sigma_t \to \infty} \mathop{\mathbb{E}}_{j}\left[C_{j,t}(H_{i,t},\theta_j)\right] = C_{j,t}(H_{i,t},\hat{\theta}_j) > E_t(B_t,H_{i,t}),$$

for all $H_{i,t} > 0$, since $p_{j,t} \in [0,1]$. □

Truthful reporting in just one of the related transactions is sufficient to render infeasible underreporting in all transactions; as the number of related transactions increases, so too does the chance that a truthful report will occur in any one of those transactions. The analogy of tax evasion in large organzations follows naturally. Since many transactions occur within large organisations, there are many opportunities for truthful reporting by a single individual to render infeasible underreporting by all others. Whenever the tax authority receives information that tax evasion has occurred, this alerts it to the likelihood that evasion is widespread in the firm.

Corollary 2. *For any given transaction, once the number of related transactions reaches a certain level, consistent underreporting is infeasible.*

Proof. For each t, there exists a $\#\Sigma_t$, beyond which:

$$\begin{aligned} E_t(B_t,H_{i,t}) &< \sum_{j=1}^{N} \mathbf{P}_{j,t} \frac{p_{i,t}}{p_{j,t}} \mathop{\mathbb{E}}_{j}\left[C_{j,t}(H_{i,t},\theta_j)\right] \\ &= \sum_{j=1}^{N} \mathbf{P}_{j,t} \frac{p_{i,t}}{p_{j,t}} \left\{ p_{j,t}^{\#\Sigma_t} C_{j,t}(H_{i,t},\bar{\theta}_j) + \left(1 - p_{j,t}^{\#\Sigma_t}\right) C_{j,t}(H_{i,t},\hat{\theta}_j) \right\}, \end{aligned}$$

which follows from the continuity of the function in the curly braces, and the fact that $C_{j,t}(H_{i,t},\bar{\theta}_j) < \mathbb{E}_j\left[C_{j,t}(H_{i,t},\theta_j)\right] < C_{j,t}(H_{i,t},\hat{\theta}_j)$. □

The $\#\Sigma_t$ which satisfies this condition can be interpreted as the maximum size of a firm in which the evasion of taxes for a particular transaction can be sustained. If agents were to account for the number of related transactions when forming their

[17] Suppose that $\Sigma_t = \{1,\ldots,S\}$, so $\#\Sigma_t = S$. Then $\mathbb{P}_j[H_{k,s} = \mathbf{P}_{l,s}H_{l,s} \;\forall\; l \neq k \;\&\; \forall\; s \in \Sigma_t] = \mathbb{P}_j[H_{k,1} = \mathbf{P}_{l,1}H_{l,1} \;\forall\; l \neq k] \times \ldots \times \mathbb{P}_j[H_{k,S} = \mathbf{P}_{l,S}H_{l,S} \;\forall\; l \neq k] = \mathbb{P}_j[H_{k,s} = \mathbf{P}_{l,s}H_{l,s} \;\forall\; l \neq k]^S = p_{j,s}^S$.

subjective probabilities of consistent reports, then one could reasonably expect no agent to underreport when the number of transactions exceeds this level; in such a case, even inconsistent underreporting is infeasible.

Uncertainty about reporting costs allows one to describe the efficacy of a third-party reporting regime to all but eliminate evasion in large firms. In a large firm, information about the payment of employees' wages might be known by many individuals within the firm. If the firm engages in a collusive arrangement with its employees in order to evade taxes, there is some chance that somebody with knowledge of the true wages will provide that information to the tax authority. In engaging in the evasion, the employee and employer will need to consider this possibility. This is a reformulation of that presented by Kleven et al. (2016). Evasion of wage and salary income within large firms is said in practice to be virtually nonexistent due to the presence of a large number of potential whistleblowers within the firm. My model nests this case.

3.6 Conclusion

I consider how a variety of factors interact with the enforcement properties of third-party reporting. Third-party reporting lowers the feasibility and degree of underreporting, as does the number of parties with knowledge of a transaction who are asked to report. Related transactions, where an agent is required to provide a common report across the transactions in a way that has opposing effects, can be a particularly promising target for third-party reporting. In some cases, this will ensure full compliance, as in the case of the V.A.T.. The feasibility and degree of underreporting under third-party reporting both decrease when the agents are uncertain about the underreporting costs faced by the other agents. And when the number of related transactions is sufficiently large, as in the case of a large firm, evasion is infeasible under third-party reporting.

In this chapter, I offer a canonical model of tax evasion under third-party reporting. The model explains the efficacy and limitations of third-party reporting regimes as tax enforcement tools. Under third-party reporting, tax evasion requires cooperation between the reporting agents; the costlier is such cooperation, the less likely is tax evasion to occur. This is the essential quality of third-party reporting. Understanding the factors that drive up the cost of collusion is essential to devising a third-party reporting regime that fulfills the conventional wisdom of full compliance. A few of the factors I consider ensure full compliance, and others ensure compliance increases, possibly substantially. These offer a guide to the design of an optimal third-party reporting system.

Appendix A

A.1 Nonparametric density ratio estimation

Determining the range of taxable income in which the densities in the treatment and nontreatment periods diverge is equivalent to determining that in which their ratio diverges from one. With that in mind, note that Bayes' rule implies that the ratio of the two densities may be estimated as:

$$\frac{\hat{h}(z \mid T = 1)}{\hat{h}(z \mid T = 0)} = \frac{\hat{\mathbb{P}}[T = 1 \mid z]}{1 - \hat{\mathbb{P}}[T = 1 \mid z]} \cdot \frac{1 - \hat{\mathbb{P}}[T = 1]}{\hat{\mathbb{P}}[T = 1]}. \tag{A.1.1}$$

I estimate these conditional probabilities via local-likelihood logit regression, which is a standard nonparametric local regression but with a logistic rather than simple linear specification for the local regressions, and estimation via maximum likelihood (Frölich, 2006). Local regression has the attractive property of performing well at endpoints, such as at the threshold in the present context, hence its popular use in regression-discontinuity designs. For each value of $z = \tilde{z}$, the local-likelihood logit estimator is given by $\hat{\mathbb{P}}[T = 1 \mid \tilde{z}] = 1/(1 + \exp(-g(\tilde{z}; \hat{\boldsymbol{\beta}}_{\tilde{z}})))$, where:

$$\hat{\boldsymbol{\beta}}_{\tilde{z}} = \underset{\boldsymbol{\beta}_{\tilde{z}}}{\arg\max} \sum_{i=1}^{n} \left(T_i \ln \left(\frac{1}{1 + e^{-g(z_i; \hat{\boldsymbol{\beta}}_{\tilde{z}})}} \right) + (1 - T_i) \ln \left(\frac{1}{1 + e^{g(z_i; \hat{\boldsymbol{\beta}}_{\tilde{z}})}} \right) \right) \times K(z_i - \tilde{z}),$$

where T_i is the treatment-period indicator for observation i, and:

$$g(z_i; \boldsymbol{\beta}) = \beta_0 + \beta_1 \cdot \mathbf{1}[z_i > z^*] + \beta_2 \cdot z_i + \beta_3 \cdot (z_i \times \mathbf{1}[z_i > z^*]),$$

which, given the discontinuity at the threshold, includes taxable income, a dummy

variable indicating taxpayers located above the threshold, and their interaction, and:

$$K(z_i - \tilde{z}) = \frac{3}{4} \cdot \frac{1 - \left(\frac{z_i - \tilde{z}}{h}\right)^2}{h},$$

which is the Epanechnikov kernel with a bandwidth of h. I set $h = 500$, which visually appears to provide an acceptable balance between over- and under-smoothing. This maximisation problem can be solved via maximum likelihood. The predicted values from this regression are consistent estimates of the conditional probabilities in equation A.1.1. The unconditional treatment probabilities in equation A.1.1 can be estimated via simple averages of the treatment indicator in the relevant ranges of taxable income considered. One then can use these predicted values and means to compute the density-ratio estimate in equation A.1.1.

A.2 Extensive-margin response

The tax applies only to those who do not have private health insurance. There are two ways taxpayers can avoid the tax: by forgoing private health insurance, and reducing taxable income to below the threshold (an intensive-margin response), or by forgoing reducing taxable income, and taking up private health insurance (an extensive-margin response). If private health insurance were randomly assigned, then any extensive-margin response could be ignored for the purposes of estimating the effect of the notch on deductions and gross income. As the take up of private health insurance is a choice, it's possible that the propensity to do so is related to deductions and gross income, which would introduce a selection bias.

I observe the complement of the set of taxpayers considered for the main analysis, which includes the extensive-margin responders (the exempt group). One complication, however, is that there is mild, but visually evident bunching below the threshold among this group in the treatment period (which disappears in the nontreatment period). This must be because not all either are legally exempt from the tax, or believe that they are. One possible reason is that I exclude all those with a spouse, but not all spouses can be classified as a dependent for the purposes of the policy. Some taxpayers with a spouse might therefore face an incentive to move below the threshold to avoid the tax. This makes it difficult to distinguish between changes due to bunching, and changes due to the extensive-margin response.

I implement the following strategy to address this problem. I assume that bunchers in the exempt group have on average the same responses as the bunchers in the non-exempt group. I measure the proportion of bunchers in the manipulation region (as defined for the non-exempt group) in the exempt group (using the methods described

earlier), and then use this proportion to impute the expected effect of bunching among all taxpayers (bunchers and nonbunchers) in the exempt group. I then subtract this from the observed differences-in-differences for the exempt group to determine the effect of the extensive-margin responders on deductions, gross income, and taxable income among the exempt group.

The results are displayed in table 1.6. The estimated probability of bunching is 5.17%. In the first column are estimates of the expected effect of bunching in the exempt group if the bunchers in the exempt and non-exempt groups had the same response (the probability of bunching multiplied by the estimates among the bunchers in the final column of table 1.3). The second column is the observed effect for the exempt group, based on the differences-in-differences. By subtracting the implied bunching effect from the observed effect, I obtain the implied extensive-margin effect in the final column. This captures the estimated effect of the extensive-margin responders on average deductions, gross income, and taxable income. The estimated effects are both substantively modest and statistically insignificant.

Appendix B

B.1 Identification proof

The following is a proof of the main identification result in chapter 2. It shows that the observed conditional expectation, $\mathbb{E}[Y_i \mid Z_i = z, S_i = 1]$, may be reweighted using propensity-score weights to obtain the non-observed conditional expectation, $\mathbb{E}[Y_i \mid Z_i(0) = z, S_i = 1]$.

Proof.

$$
\begin{aligned}
&\mathbb{E}[Y_i \mid Z_i(0) = z, S_i = 1] \\
&\qquad \equiv \int y f_Y(y \mid Z_i(0) = z, S_i = 1)\, dy \\
&\qquad = \int\int y f_{Y,\mathbf{X}}(y, \mathbf{x} \mid Z_i(0) = z, S_i = 1)\, d\mathbf{x} dy \\
&\qquad = \int\int y f_Y(y \mid Z_i(0) = z, \mathbf{X}_i = \mathbf{x}, S_i = 1)\, f_{\mathbf{X}}(\mathbf{x} \mid Z_i(0) = z, S_i = 1)\, d\mathbf{x} dy \\
&\qquad = \int\int y f_Y(y \mid Z_i(0) = z, \mathbf{X}_i = \mathbf{x}, S_i = 1)\, f_{\mathbf{X}}(\mathbf{x} \mid Z_i(0) = z, S_i = 0)\, d\mathbf{x} dy \\
&\qquad = \int\int y f_Y(y \mid Z_i(0) = z, \mathbf{X}_i = \mathbf{x}, S_i = 1)\, f_{\mathbf{X}}(\mathbf{x} \mid Z_i(1) = z, S_i = 1)\, \omega d\mathbf{x} dy \\
&\qquad = \int\int y f_Y(y \mid Z_i(1) = z, \mathbf{X}_i = \mathbf{x}, S_i = 1)\, f_{\mathbf{X}}(\mathbf{x} \mid Z_i(1) = z, S_i = 1)\, \omega d\mathbf{x} dy \\
&\qquad = \int\int y f_{Y,\mathbf{X}}(y, \mathbf{x} \mid Z_i = z, S_i = 1)\, \omega d\mathbf{x} dy \\
&\qquad \equiv \tilde{\mathbb{E}}[Y_i \mid Z_i = z, S_i = 1],
\end{aligned}
$$

where the third equality holds under Assumption 3 (random assignment), the fourth under the definition of the propensity-score weights and Assumption 4 (common support), and the fifth under Assumption 2 (manipulation on observables). □

Bibliography

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

Australian Bureau of Statistics (2009), Average weekly earnings, Technical Report 6302.0, Australian Bureau of Statistics. Available here: http://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/376AFC773288BA43CA25761000197A9F/$File/63020_may%202009.pdf.

Australian Bureau of Statistics (2013), Australian health survey: Health service usage and health related actions, 2011-12, Technical Report 4364.0.55.002, Australian Bureau of Statistics. Available here: http://www.abs.gov.au/ausstats/abs@.nsf/lookup/E334D0A98272E4DCCA257B39000F2DCF?opendocument.

Australian Taxation Office (2009), Taxation statistics 2008-09, Technical report, Australian Taxation Office. Available here: https://www.ato.gov.au/assets/0/104/300/362/2cb0b6c4-67b5-47a5-b4c1-b1554a05e7d1.pdf.

Bachas, P. and Soto, M. (2017), 'Not(ch) your average tax system: Corporate taxation under weak enforcement. Unpublished working paper, available here: https://dl.dropboxusercontent.com/content_link/siuYy8luN6NcAbjydJrg0HLEy5E12TGeXKrI1fBap92PB6tOrffhyM0qNSd2K9E9/file.

Barreca, A. I., Lindo, J. M. and Waddell, G. R. (2015), 'Heaping-induced bias in regression-discontinuity designs', *Economic Inquiry* **54**(1), 268–293.

Best, M., Brockmeyer, A., Kleven, H., Spinnewijn, J. and Waseem, M. (2015), 'Production versus revenue efficiency with limited tax capacity: Theory and evidence from Pakistan', *Journal of Political Economy* **123**(6), 1311–1355.

Boadway, R., Marceau, N. and Mongrain, S. (2002), 'Joint tax evasion', *Canadian Journal of Economics* **35**(3), 417–435.

Bovenberg, A. L. and de Mooij, R. A. (1994), 'Environmental levies and distortionary taxation', *American Economic Review* **84**(4), 1085–1089.

Calonico, S., Cattaneo, M. and Titiunik, R. (2014), 'Robust nonparametric confidence intervals for regression-discontinuity designs', *Econometrica* **82**(6), 2295–2326.

Chang, J. and Lai, C. (2004), 'Collaborative tax evasion and social norms: Why deterrence does not work', *Oxford Economic Papers* **56**(2), 344–368.

Chapman, D., Cook, T., Zurovac, J., Coopersmith, J., Finucane, M., Vollmer, L. and Morris, R. (2018), 'The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons', *Journal of Policy Analysis and Management* **37**(2), 403–429.

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., Friedman, J. N., Olsen, T. and Pistaferri, L. (2011), 'Adjustment costs, firm responses, and micro vs. macro labor supply responses: Evidence from Danish tax records', *The Quarterly Journal of Economics* **126**(2), 749–804.

DiNardo, J., Fortin, N. M. and Lemieux, T. (1996), 'Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach', *Econometrica* **64**(5), 1001–1044.

Doerrenberg, P., Peichl, A. and Siegloch, S. (2015), 'The elasticity of taxable income in the presence of deduction possibilities', *Journal of Public Economics* **151**, 41–55.

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', *The Review of Economics and Statistics* **81**(4), 674–680.

Frölich, M. (2006), 'Non-parametric regression for binary dependent variables', *The Econometrics Journal* **9**(3), 511–540.

Grembi, V., Nannicini, T. and Troiano, U. (2016), 'Do fiscal rules matter?', *American Economic Journal: Applied Economics* **8**(3), 1–30.

Gruber, J. and Saez, E. (2002), 'The elasticity of taxable income: Evidence and implications', *Journal of Public Economics* **84**, 1–32.

Hahn, J., Todd, P. and Van der Klaauw, W. (2001), 'Identification and estimation of treatment effects with a regression-discontinuity design', *Econometrica* **69**(1), 201–209.

Holland, P. W. (1986), 'Statistics and causal inference', *Journal of the American Statistical Association* **81**(396), 945–960.

Jordan, C. (2017), 'Commissioner's address to the National Press Club'. Transcript available here: https://www.ato.gov.au/Media-centre/Speeches/Commissioner/Commissioner-s-address-to-the-National-Press-Club/.

Keele, L., Titiunik, R. and Zubizarreta, J. R. (2015), 'Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout', *Journal of the Royal Statistical Society A* **178**(1), 223–239.

Kleven, H. (2016), 'Bunching', *Annual Review of Economics* **8**, 435–464.

Kleven, H., Kreiner, K. and Saez, E. (2016), 'Why can modern governments tax so much? An agency model of firms as fiscal intermediaries', *Economica* **83**, 219–246.

Kleven, H. and Waseem, M. (2013), 'Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan', *The Quarterly Journal of Economics* **128**(2), 669–723.

Lee, D. S. (2008), 'Randomized experiments from non-random selection in U.S. House elections', *Journal of Econometrics* **142**(2), 675–697.

Lee, D. S. and Lemieux, T. (2010), 'Regression discontinuity designs in economics', *Journal of Economic Literature* **48**(2).

Linden, A. and Adams, J. (2012), 'Combining the regression discontinuity design and propensity score-based weighting to improve causal inference in program evaluation', *Journal of Evaluation in Clinical Practice* **18**, 317–325.

Liu, L., Lockwood, B. and Almunia, M. (2017), VAT notches, voluntary registration, and bunching: Theory and UK evidence. Available at: https://warwick.ac.uk/fac/soc/economics/staff/blockwood/vat-notches-voluntary_31_7_17.pdf.

McCaffrey, D. F., Ridgeway, G. and Morral, A. R. (2004), 'Propensity score estimation with boosted regression for evaluating causal effects in observational studies', *Psychological Methods* **9**(4), 403–425.

McCrary, J. (2008), 'Manipulation of the running variable in the regression discontinuity design: A density test', *Journal of Econometrics* **142**(2), 698–714.

Naritomi, J. (2016), Consumers as tax auditors. Available at: https://www.dropbox.com/s/1e0bctgjji4s01c/naritomi_enforcement_May2016.pdf?dl=0.

Paetzold, J. (2017), 'How do wage earners respond to a large kink? Evidence on earnings and deduction behavior from Austria'. Unpublished working paper, available here: https://www.uni-salzburg.at/fileadmin/multimedia/SOWI/documents/VWL/Working_Papers/Workingpaper_2017_01.pdf.

Pomeranz, D. (2015), 'No taxation without information: Deterrence and self-enforcement in the value added tax', *American Economic Review* **105**(8), 2539–2569.

Ramsey, F. (1927), 'A contribution to the theory of taxation', *The Economic Journal* **37**(145), 47–61.

Rosenbaum, P. R. and Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), 41–55.

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies', *Journal of Educational Psychology* **66**(5), 688–701.

Saez, E. (2010), 'Do taxpayers bunch at kink points?', *American Economic Journal: Economic Policy* **2**(3), 180–212.

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.

Schächtele, S. (2016), Deduction responses to the income tax: Bunching evidence from Germany. Unpublished working paper, available here: https://www.econstor.eu/bitstream/10419/145748/1/VfS_2016_pid_6770.pdf.

Slemrod, J. (1998), 'Methodological issues in measuring and interpreting taxable income elasticities.', *National Tax Journal* **51**(4), 773–788.

Slemrod, J. and Kopczuk, W. (2002), 'The optimal elasticity of taxable income', *Journal of Public Economics* **84**(1), 91–112.

Stavrunova, O. and Yerokhin, O. (2014), 'Tax incentives and the demand for private health insurance', *Journal of Health Economics* **34**, 121–130.

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.

Yaniv, G. (1992), 'Collaborated employee-employer tax evasion', *Public Finance* **47**(2).