The London School of Economics and Political Science

Essays on Behavioral Responses to Social Insurance and Taxation

Arthur Seibold

Thesis submitted to the Department of Economics of the London School of Economics for the degree of Doctor of Philosophy.

London, July 2018

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent.

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of approximately 32,500 words (excl. graphs, tables and appendices).

Statement of Conjoint Work

I confirm that Chapter 3 was jointly coauthored with François Gerard and Joana Naritomi and I contributed 33% of this work.

Abstract

This thesis contains three essays on behavioral responses to social insurance and taxation. The first chapter documents and analyzes an important and puzzling stylized fact about retirement behavior: the large concentration of job exits at specific ages. In Germany, almost 30% of workers retire precisely in the month when they reach one of three "statutory" retirement ages, although there is often no incentive or even a disincentive to retire at these thresholds. To study what can explain the concentration of retirements around statutory ages, I use novel administrative data covering the universe of German retirees, and I take advantage of unique variation in retirement incentives as well as in the location of statutory ages across individuals created by the German pension system. Measuring retirement bunching responses to 644 different discontinuities in pension benefit profiles, I first document that financial incentives alone fail to explain retirement patterns in the data. Second, I show that there is a direct effect of "presenting" a threshold as a statutory age, which is substantially larger than that of financial incentives. Further evidence on mechanisms suggests the framing of statutory ages as reference points for retirement as an explanation. A number of alternative channels including firm responses are also discussed but they do not seem to drive the results.

The second chapter analyzes bunching responses around reference points and argues that bunching methods are naturally suited to quantify reference-dependent preferences. Using a standard labor supply model, the workhorse of the bunching literature, I first show that different types of reference dependence all have a key prediction in common: They imply sharp bunching of the outcome at the reference point. Observed bunching can be linked to underlying parameters, which motivates both structural and reduced-form estimation methods to implement an empirical bunching approach to reference dependence. Finally, I present two applications in the context of retirement decisions. First, I find significant bunching responses at a type of "pure" reference point, namely round retirement ages. Second, I complement the analysis from chapter 1 with structural estimation and find a quantitatively important role of reference dependence at statutory retirement ages. Counterfactual simulations highlight that shifting statutory ages via pension reforms can be an effective policy to increase actual retirement ages with a positive fiscal impact.

The third chapter turns to a topic from the realm of taxation. Modern systems of firm taxation typically feature a combination of payroll, valued-added, and corporate income taxes. However, they often exist alongside special presumptive tax regimes targeted at small and medium enterprises (SME), such as a single turnover tax. This chapter uses novel administrative data from São Paulo (Brazil), including data on inter-firm trade, to shed light on the effects of such dual tax systems on firm growth, market competition, and production decisions. First, we show that the firm size distribution is distorted by the eligibility threshold for the presumptive tax system. Second, ineligible (larger) firms are adversely affected by reductions in the tax and compliance burden for SME. Third, we study the relationship between tax systems and production choices. The presumptive tax mainly replaces a payroll tax and a value-added tax by a turnover tax in our context. Accordingly, we find that firms in the presumptive tax regime use relatively more labor

input and source more of their intermediate input from other firms in the same regime. This leads to partial segmentation of the trade network between firms in the two systems. We show that heterogeneity in firm production choices drives part of these correlations, but there is also a causal effect of tax regimes on input choices.

Acknowledgments

This thesis is the product of a long journey that would not have been possible without the help of many. First, I am deeply indebted to my supervisors Henrik Kleven and Camille Landais for their continued support and advice. Their approach to economics has shaped my understanding of the field. Henrik and Camille's conceptual clarity and their views on empirical research have guided my work, and they have taught me not to hesitate to tackle big questions. Moreover, they have offered invaluable advice on how to navigate the PhD and the job market.

In addition to my supervisors at LSE, I wish to thank Emmanuel Saez for hosting and advising me at UC Berkeley. The visit has given me new perspectives and and interactions with Emmanuel and others have had a large impact on my work. I also thank Wouter den Haan, Erik Eyster, François Gerard, Xavier Jaravel, Guy Michaels, Joana Naritomi, Steve Pischke, Daniel Reck and Johannes Spinnewijn for engaging with my research and for offering helpful advice. For insightful comments on my papers, I am grateful to Michel Azulai, Youssef Benzarti, David Card, Alex Gelber, Felix Koenig, Erzo F.P. Luttmer, Panos Mavrokonstantis, Jan Nimczik, Dominik Sachs, Sarah Smith, Stefan Staubli, Dimitri Taubinsky and Alisa Tazhitdinova.

Tatjana Mika, Michael Stegmann and their colleagues at the Research Data Centre of the German State Pension Fund have been great help in accessing the data and understanding the institutional setting for a large part of this thesis. I would like to thank administrative staff at LSE for their help, in particular Deborah Adams, Jane Dickson, Rhoda Frith and Mark Wilbor. Funding for this research has been generously provided by the Suntory and Toyota International Centres for Economics and Related Disciplines (STICERD).

I am grateful to Albert, Michel, Miguel, Panos, Pedro, Xuezhu, and all my colleagues and friends at LSE for making the PhD experience so much more enjoyable. Before I came to London, a number of people have played a big part in my personal and academic development. Francesca Fabbri and Dalia Marin have mentored me at LMU Munich and inspired me to start doing research. Maybe most importantly, my high-school teacher Wolfgang Benker has enabled me to discover economics and taught me to strive for excellence.

However, I owe everything to my parents and my grandparents whose upbringing set me up for life and who have always believed in me. They, and my brothers, have accompanied me every step of the way and without them none of this would have been possible. Finally, my deepest gratitude goes to Dana, whose love and companionship has filled my life with joy, and who has shared this long journey with me more than anyone else.

Contents

List of Tables

List of Figures

Chapter 1

Reference Dependence in Retirement Behavior: Evidence from German Pension Discontinuities

1.1 Introduction

For many countries, population aging poses looming questions over the fiscal sustainability of public pension systems. The average OECD country already spends 8% of GDP or 18% of total public expenditure on pensions (OECD (2015)). The old-age dependency ratio, measuring the number of individuals aged 65 and above relative to the working-age population, is projected to rise from currently 27% to 49% by 2050. In addressing these issues, a widely shared policy goal is to extend the working lives of the elderly population. Standard economic models prescribe the design of appropriate financial retirement incentives as a policy tool to influence labor supply at old age. However, existing studies find small effects of financial incentives on retirement behavior (e.g. Manoli and Weber 2016*a*).

Much of the public debate on pension reform revolves around a different policy: *statutory age thresholds*. Such thresholds are used by typical public pension systems to frame benefit rules. They may include an Early Retirement Age and a Normal Retirement Age, and they usually define retirement ages relative to which benefits are calculated. What is the role of statutory ages for retirement behavior? To provide motivational evidence, figure 1.1 shows that the distribution of job exits of German workers is strongly concentrated around statutory ages. There are sharp spikes in job exits at the main statutory ages 60, 63 and 65.[1] In total, 29% of job exits at age 55 and above occur precisely in the month when the worker reaches a statutory age.

The spikes in retirement at statutory ages are not only large, but also surprising from the point of view of standard labor supply models. To preview this, consider the stylized lifetime budget

[1] Note that different statutory ages apply to workers depending on their birth cohort and characteristics such as gender and contribution histories.

constraint in figure 1.2. Most workers face a reduction in the marginal return to work, i.e. an incentive to stop working, at ages 60 and 63, but a disincentive to retire at age 65. However, large bunching occurs at all three age thresholds. Similarly, bunching at statutory ages in spite of modest retirement incentives has been observed in other countries (e.g. Mastrobuoni 2009; Behaghel and Blau 2012; Cribb et al. 2016). Yet, statutory ages play no direct role in standard models of retirement, and what drives their prominent impact for retirement behavior has been a long-standing question in the literature (Lumsdaine et al. 1996).

Having documented this puzzling stylized fact about retirement behavior, this chapter asks what can explain the concentration of retirements around statutory ages. To address this question, I estimate bunching responses at 644 benefit discontinuities in the German public pension system, using administrative data on the universe of retirees in the country. The analysis finds little impact of financial incentives, but a large direct effect of statutory age thresholds on retirement behavior. Moreover, I show evidence suggesting that this "statutory age effect" is driven by the framing of statutory ages as reference points for retirement. Based on these results, chapter 2 of this thesis uses a model of retirement with reference-dependent decision utility for counterfactual simulations, and demonstrates that shifting statutory ages can be an effective policy tool to influence retirement behavior with a positive fiscal impact.

As the empirical setting, the German public pension system provides three key advantages. First, there is rich variation in financial retirement incentives and in the location of statutory ages. This creates more than 600 discontinuities in pension benefits corresponding to kinks and notches in lifetime budget constraints. The variation arises due to two sources: There are six pathways into retirement entailing different benefit profiles, and a series of pension reforms provide additional cohort-based variation at the monthly level. The second advantage is that some discontinuities are presented as "statutory age thresholds", while others are "pure financial incentives". Statutory ages are linked to notions such as a "normal" retirement date, and pension benefit adjustment is presented as a loss or gain relative to a reference level defined at statutory ages. This feature allows for joint estimation of responses to underlying financial incentives and the direct impact of presenting a threshold as a statutory age. In addition, discontinuities vary in the size of the financial incentive, their location, and the characteristics of affected workers, allowing to control for heterogeneity along these dimensions. In total, there are 644 discontinuities over the sample period, out of which 386 are linked to a statutory age. The large number of discontinuities provides a unique opportunity to study what determines the magnitude of retirement responses.

The third advantage of the empirical setting is that high-quality administrative data is available to exploit this fine-grained variation. The analysis is based on a novel data set provided by the German State Pension Fund, covering the universe of workers who retired between 1992 and 2014. The main sample contains around nine million individuals. The data includes a rich set of worker characteristics related to earnings careers and pension eligibility, based on which monthly job exits and individual lifetime budget constraints can be calculated.

I divide the analysis in this chapter into two parts. The first part of the chapter uses bunching

methods to estimate retirement responses at the 644 available benefit discontinuities. Two main results are established. First, financial incentives alone fail to explain retirement patterns. There are large responses at statutory ages even if there is a close to zero incentive or a disincentive to retire at the discontinuity. Large differences in bunching responses across types of discontinuities can also not be explained by different financial incentives. Second, presenting a threshold as a statutory age matters directly for retirement behavior. At all types of statutory ages and irrespectively of kink sizes, large additional bunching occurs compared to pure financial incentive discontinuities.

These results emerge from two complementary approaches. In the first approach, I focus on a few "cases" of specific discontinuities that lend themselves to natural comparison. For instance, the same group of workers is shown to respond much more to an Early Retirement Age kink than to a pure financial incentive notch, although the notch entails a larger financial incentive to retire. Moreover, workers respond more strongly to a Full Retirement Age kink than at a pure financial incentive kink of similar size occurring at the same retirement age in a similar pathway. In the second approach, I use the full set of discontinuities to generalize the results. The average observed elasticity of the retirement age w.r.t. the net-of-tax rate across all 386 statutory age kinks is 1.64. Across the 258 pure financial incentive discontinuities in the data, the average observed elasticity is only 0.15, less than a tenth of the estimate at statutory ages.

I also propose a reduced-form strategy that combines bunching estimates in a regression to jointly estimate the response to financial incentives and "statutory age effects" across discontinuities. The identification assumption is that responses to different types of discontinuities are driven by the same underlying parameters. The estimated average net-of-tax elasticity of around 0.1 is modest and statutory age effects are large and significant. Results are robust to controlling for a range of observables and fixed effects. Moreover, I show that there is some heterogeneity in responses, but the pattern of larger responses at statutory ages holds across all types of workers, including quantiles of income, education, firm size and retirement ages. Furthermore, results from the main estimation are robust to allowing for heterogeneous parameters along different dimensions.

The second part of this chapter explores mechanisms behind the reduced-form "statutory age effect". I begin by showing evidence that the framing of statutory ages as reference points for retirement provides a plausible mechanism. Exploiting a natural experiment where the frequency of information letters sent to workers is drastically increased, I document that more workers retire at a specific statutory age around which explanations in letters are framed. Additional information in the letters intended to inform workers about benefit calculation does not affect responses to financial incentives. Moreover, two patterns in the data speak against a mechanism purely driven by a lack of information or mistakes. First, retiring at statutory ages is positively associated with worker characteristics commonly used to proxy for financial literacy, including education, economic training and income. Second, the response to statutory ages is not diminished when a worker's stake in the retirement decision, e.g. measured by the size of their pension wealth relative to earnings, increases.

Firm responses are also discussed as a potential alternative mechanism. This may be a concern

since some statutory ages can play a role in the termination of labor supply contracts. However, I show that (i) self-employed workers and those in very small firms below the employment protection threshold also bunch at statutory ages, (ii) excluding those statutory ages where mandatory retirement is possible does not change the remaining results, and (iii) a number of proxies for firm incentives, including the fraction of workers in unlimited contracts and labor market tightness, are only weakly related to statutory ages retirements.

This chapter relates to three strands of literature. First, it contributes to the recent empirical literature on retirement behavior. Brown (2013) and Manoli and Weber (2016*a*) investigate the responsiveness of retirement to pure financial incentives and find very small elasticities. Those papers come closest to this study in terms of the bunching methods used. On the other hand, here are several studies of statutory age reforms that find large effects on retirement behavior, including in the U.S. (Mastrobuoni 2009), the U.K. (Cribb et al. 2016), Austria (Manoli and Weber 2016*b*) and Switzerland (Staubli and Zweimüller 2013; Lalive and Staubli 2015). This chapter proposes an explanation for these diverging results, where responses to statutory ages are exacerbated by reference point effects. Moreover, some studies provide support for framing effects or reference dependence around statutory ages. In particular, the argument by Behaghel and Blau (2012) that workers are loss averse relative to the Full Retirement Age is closely related to this chapter. Brown et al. (2013) and Merkle et al. (2017) find experimental evidence of framing effects around statutory ages. Shoven et al. (2017) present survey evidence indicating that retiring at statutory ages may be perceived as a social norm.

Second, this chapter contributes to the bunching literature reviewed by Kleven (2016). Initially, the bunching method was used to estimate a price elasticity at a budget set discontinuity (Saez 2010; Chetty et al. 2011), but recent studies have moved towards using additional bunching moments to estimate additional parameters. For instance, Kleven and Waseem (2013) estimate elasticities and the share of individuals subject to frictions at notches. Gelber et al. (2017) develop a difference-in-bunching approach to estimate an elasticity and an adjustment cost parameter. The present setting with more than 600 discontinuities allows for a rich analysis of the drivers of bunching. This type of setting can help address challenges of bunching estimation where results can be somewhat local to specific discontinuities. Building on the literature, this chapter develops methods for a setting where many discontinuities are available and, in contrast to previous work, bunching methods are used to estimate parameters that exacerbate bunching in this chapter.

The remainder of this chapter is organized as follows. Section 1.2 outlines the empirical context and the data, section 1.3 describes the empirical methodology, section 1.4 presents reduced-form evidence, section 1.5 explores mechanisms behind the statutory age effect, and section 1.6 concludes.

1.2 Context and Data

1.2.1 The German Public Pension System

Germany has a pay-as-you-go pension system that covers the vast majority of workers in the country (86% of the labor force in 2014). Enrolment is mandatory for private-sector employees, but most self-employed workers and civil servants are exempt. Contributions are levied as a payroll tax on gross earnings.[2] Benefits are defined according to a pension formula based on a worker's lifetime contribution history.[3] Hence, pensions are roughly proportional to lifetime income and the system is characterized by relatively little redistribution. The average replacement rate is 50% (OECD 2015). Public pensions are the main source of income for most recipients.[4] Moreover, there is a relatively strict earnings test for pension recipients where earnings above €450 per month lead to reductions in benefit payments. Only 2.5% of workers in the main sample have any income from employment while receiving a pension, making retirement an absorbing state for most.

The key advantage of the institutional setting is the unique number of pension discontinuities it provides. To begin with, the system features three types of statutory age thresholds where pension eligibility changes. First, the *Early Retirement Age (ERA)* is the earliest age from which any pension can be claimed. Second, the *Full Retirement Age (FRA)* is the earliest age from which workers can claim their *full pension*. Third, the *Normal Retirement Age (NRA)* is the age from which workers can get more than their full pension.[5]

There is a large amount of variation in statutory ages across workers along two dimensions. First, there are six pathways into retirement that differ in their ERA and FRA, and workers need to meet specific requirements to be eligible. Pathways are summarized in table 1.1. The basic pathway is the regular pension that requires only 5 years of contributions. However, early retirement is not possible in the regular pathway and the full pension can only be claimed at the NRA. In order to enter a more generous pathway, workers must have contributed for longer and/or satisfy other requirements such as disability. Specifically, at 15 and 35 years of contributions, workers become eligible for pathways with ERAs between 60 and 63, and FRAs between 63 and 65. Other pathway requirements include gender, disability and periods of unemployment.

The second dimension of variation arises due to a series of cohort-based pension reforms enacted since the early 1990s. Figure 1.3 shows the evolution of ERAs and FRAs for birth cohorts 1932 to 1949. Statutory ages differ considerably among the early cohorts, and they were changed in

[2] Workers in so-called mini jobs with earnings less than €450 are exempt from contributions. Besides, contributions have to be paid for some non-work periods such as receiving certain types of unemployment benefits.

[3] Appendix 1.A.2 provides additional details on benefit calculation and other aspects of the institutional setting. See also Börsch-Supan and Schnabel (1999) and Börsch-Supan and Wilke (2004) for a more comprehensive overview.

[4] See Heien et al. (2005). In 2003, 11% of retirees reported to receive any income from employer pension schemes and only 1% had a private pension, and the average income from those sources is small relative to public pensions. Among retirees with any employer pension income, the employer pension amounts to 34% of their public pension on average. The corresponding figure for private pensions is 23%. The numbers seem to increase somewhat for younger cohorts, but remain small throughout the sample period.

[5] The distinction between the FRA and NRA is somewhat peculiar to the German system. Essentially, the FRA was created to allow some workers to claim a "full pension" before the NRA if they satisfy certain requirements. However, all workers can claim more than their full pension only after the NRA.

different pathways at different times. For instance, the women's FRA was increased from 60 to 65 for cohorts 1940 to 1944. This was done gradually: the FRA increases by one month for each month of birth in the reform cohort window. Similar gradual changes to the ERA and FRA were also implemented in all other pathways.

In addition, the setting provides two sources of pension discontinuities not linked to statutory ages, representing "pure financial incentives". First, the contribution requirements of pathways mentioned above create "notches" where workers discontinuously become eligible for more generous benefits. Second, there is an invalidity pathway where pensions can be claimed at any age. This pathway has a low contribution requirement of only 5 years, but a relatively strict disability requirement. No statutory ages are defined for this pathway, but pension eligibility changes at some thresholds.[6] Section 1.2.3 explains in more detail how statutory ages and other benefit rules translate into budget constraint discontinuities.

1.2.2 The Role of Statutory Age Thresholds

Link to Pension Benefits. Benefit eligibility directly depends on statutory ages in all pathways (except invalidity pensions). A full pension level is defined at the FRA, and there are permanent reductions in benefits for workers claiming before the FRA as well as permanent increases in benefits for claiming after the NRA. The adjustment function follows a kinked schedule, with a penalty of 0.3% for each month of retirement before the FRA, no adjustment between the FRA and the NRA, and a reward of 0.5% for each month of retirement after the NRA.

Framing of Benefits and Retirement. Moreover, statutory age thresholds play an important role for how benefits and retirement are presented to workers. Appendix figure 1.A1 provides an example of of framing from a leaflet designed to inform workers about a pension reform that increases the NRA to 67. Three features stand out. First, statutory ages are somewhat directly suggested as retirement dates. For instance, the title "retirement at 67" refers to the post-reform NRA at age 67. Workers are also told that "if they want to retire early" they can do so at the ERA, and if they wish a full pension, they should retire at the FRA. Second, workers are warned of losses if they retire before the FRA ("the penalty will remain for your entire retirement"). Third, different pathways are explained to workers via differences in statutory ages, and the pension reform is explained as a change to statutory ages.

The example illustrates how pensions and retirement are framed in terms of statutory ages in three ways. First, linking statutory ages to notions such as "full" and "normal" retirement may be suggestive to workers and contribute to expectations regarding retirement dates. Second, pension adjustment for early retirement is framed as a loss relative to a "full pension" level linked to the FRA, while adjustment for late retirement is framed as a gain. In other words, statutory ages are used as *institutional reference points* in terms of both retirement dates and benefit levels. Third,

[6] Moreover, contribution points are credited to invalidity pensioners as if they had continued working until age 60, making benefits less dependent on their contribution history.

while different pathways effectively entail different benefit levels for any given retirement age, the distinction between pathways is framed via different statutory ages rather than directly in terms of benefit levels. This logic originates from German Social Law, where each pathway is defined in terms of its statutory ages, and pension adjustment based on statutory ages is defined in a separate section. Relatedly, major pension reforms are equally framed as changes to statutory ages rather than the changes to benefit levels that they effectively entail. Similar framing of retirement dates and benefit levels has been shown to affect reported retirement plans in lab settings (e.g. Brown et al. 2013, Merkle et al. 2017).

Labor Supply Contracts. Finally, statutory ages can play a role in the termination of some labor supply contracts. In particular, mandatory retirement clauses linked to workers' NRA can be specified in collective industry agreements or in individual contracts. This is sometimes cited as a way for firms to avoid high costs of firing of older workers. Importantly, there is no possibility for mandatory retirement clauses linked to the ERA or FRA.[7] The potential role of firms is further addressed in section 1.5.2.

1.2.3 Lifetime Budget Constraint Discontinuities

In order to see how the pension system affects incentives for the timing of retirement, the net present value of a worker i's lifetime income can be written as a function of her retirement (job exit) age R_i:

$$NPV_i(R_i) = \sum_{t=0}^{R_i-1} \delta^t w_{it}(1-\tilde{\tau}_{it}) + \sum_{t=\max(R_i,ERA)}^{T_i} \delta^t B_i(R_i) \tag{1.1}$$

The worker earns a gross wage w from starting age 0 to the period before retirement, which is subject to income tax and social insurance contributions summarized in $\tilde{\tau}$. Pension benefits B depend on R both via contributions paid until retirement and pathway-specific pension adjustment. Benefits can be claimed from the age at job exit if the worker has already reached her ERA (and from the ERA otherwise) and are paid until time of death T. Finally, all payments are discounted at factor $\delta = \frac{1}{1+r}$, where r is the interest rate.

To satisfy the lifetime budget constraint, $C_i = NPV_i$, i.e. lifetime consumption possibilities C are given by discounted lifetime income streams. The slope of the budget constraint, i.e. the marginal gain in lifetime consumption from delaying retirement by one period, defines the *implicit net wage* $w^{net} = \frac{dC}{dR}$. Expressing the consumption gain as a fraction of gross earnings, the *implicit net-of-tax rate* can be calculated as $1-\tau = \frac{w^{net}}{w}$. In general, delaying retirement affects consumption in three ways. First, the worker gains an additional period of wage earnings. Second, she sees a permanent change in her benefit eligibility $\frac{dB}{dR}$. In the German case $\frac{dB}{dR}$ is always strictly positive, since later retirement implies both more favorable pension adjustment and a larger sum

[7] Individual agreements between the worker and the firm to terminate a contract at the ERA/FRA can be added to the contract no earlier than three years before the desired time of job exit, but similar agreements can also be made in reference to other dates.

of contribution points. Third, if she retires at the ERA or later, i.e. she is already eligible to claim benefits, there is an opportunity cost of work in terms of foregoing one period of benefits.

Discontinuous changes in pension eligibility introduce discontinuities into the lifetime budget constraint. The empirical setting provides more than 600 distinct discontinuities, which can be grouped into the following three types:[8]

Kinks Linked to Statutory Ages. Figure 1.2 shows a stylized version of the lifetime budget constraint. There are *convex kinks*, i.e. reductions in the marginal net-of-tax rate, at the ERA and the FRA. Moreover, there is a *non-convex kink*, i.e. an increase in the marginal return to work, at the NRA.[9] The kinks at the FRA and NRA arise as a direct consequence of the discontinuous pension adjustment described in the previous section, where annual adjustment falls from 3.6% to 0 at the FRA and jumps from 0 to 6% at the NRA. The convex kink at the ERA arises due to a combination of pension adjustment and an additional opportunity cost of working, since workers start foregoing benefits once they reach the ERA.[10] The location of statutory ages varies by pathway and birth cohort and there are a total of 386 budget constraint kinks linked to statutory ages.

Contribution Notches. The contribution thresholds required for different pathways create further budget set discontinuities in the form of *notches*, i.e. jumps in average net-of-tax rate. In figure 1.2, for instance, the worker reaches 35 years of contributions when working until age 58, where he becomes eligible for the long-term insured pathway and now faces both a lower ERA and a lower FRA. Thus, he can receive a pension earlier (i.e. for more years) and his pension is higher at any given age due to more favorable adjustment, which implies a discontinuous jump in pension wealth. Similarly, workers face notches when they become eligible for other pathways at 5, 15 and 35 years of contributions.[11] The precise location of these notches is worker-specific since it depends on the individual career starting age. Combining variation across pathways, cohorts and age groups yields a total of 180 such notches.

Kinks in the Invalidity Pathway. Pensions are also discontinuously adjusted in the invalidity pathway. Specifically, benefits are increased by 3.6% p.a. for retiring between 60 and 63, with no further adjustment when claiming before 60 and after 63. These kinks in the benefit schedule imply budget constraint kinks similar to those around statutory ages. However, the key difference to other pathways is that there are no statutory ages in the invalidity pathway.[12] Including a

[8] See appendix 1.A.4.2 for the a complete list of all discontinuities used for bunching.

[9] An exception is the regular pathway where the ERA coincides with the NRA. In this case, there is a convex kink at the ERA/NRA.

[10] The ERA kink could be smoothed out by actuarially fair adjustment of pensions. However, the actual adjustment of 3.6% annually is not sufficient (see Börsch-Supan and Wilke 2004).

[11] The notches at 5 years of contributions are not used in this chapter since the data on workers with less than 5 years of contributions is incomplete.

[12] This is presumably intended to mirror adjustment in the other pathways in order to avoid incentives for switching to invalidity pensions. Notice that the invalidity adjustment function is equivalent to adjustment based on an ERA of 60 and an FRA of 63, and thus coincides precisely with the benefit schedule in the *disability pathway* which may be seen as the closest substitute.

gradual introduction period, there are 78 kinks invalidity pension kinks.

1.2.4 Data

The analysis is based on a novel set of administrative data covering the universe of retirees who claimed a public pension between 1992 and 2014. The main data set is assembled from 23 single-year cross sections provided by the German State Pension Fund.[13] The sample is limited to workers in the six main pathways who claimed a pension for the first time between ages 55 and 67, have earned at least 5 contribution points from at least 5 years of contributions and do not continue work after retirement. Moreover, individuals part of whose earnings careers have been abroad and members of a special scheme for miners are excluded. Finally, East Germans retiring in 1995 and earlier are excluded since their pension was calculated under a particular set of post-reunification rules. In order to have sufficient parts of each cohort's retirement age distribution available, the analysis focuses on workers born between 1933 and 1948. After applying those restrictions, the *individual sample* contains around 8.9 million observations.

The data includes all variables necessary for the pension fund to determine a worker's pension eligibility as well as a number of socioeconomic characteristics. Monthly benefit claims and last contributions can be directly observed. The month of job exit can be inferred from the time of the last contribution for most of the sample. For those workers where the last contribution does not coincide with employment, the time of job exit is imputed using additional information on the insurance status in the last three years before retirement. Lifetime earnings and average annual earnings are backed out using information on contribution periods and contribution points,[14] and a pension benefit simulator is built to calculate individual benefit eligibility across possible retirement ages. Lifetime budget constraints are simulated as a version of equation (1.1) with a 3% discount rate and heterogeneous life expectancies by gender and year of birth. In order to account for the fact that observed take-up of pathways may reflect workers' choices, pathways are assigned in terms of eligibility as far as possible. This may be particularly important for cohorts where reforms could induce some "switching" between pathways, which may change group composition over time.

In addition, survey data from the German Socioeconomic Panel (SOEP) is used for part of the analysis.[15] SOEP is an unbalanced panel of around 1.4 million individual-year observations spanning the period 1984 to 2013. It contains a wide range of socioeconomic variables including labor market outcomes. Variables of interest are collapsed at the three-digit occupation level and merged with the main data where occupation can be observed from 2000 onwards. This sample is referred to as the *occupation-matched sample*.

As explained in section 1.2.1, pension discontinuities differ across pathways and cohorts. In

[13]Data citation: *Versichertenrentenzugang 1992-2014, source: FDZ-RV*. See appendix 1.A.3 for details of key variables and other definitions.

[14]Contribution points are generally proportional to gross earnings. The only caveat is top-coding of earnings above the contributions cap.

[15]Data citation: *Socio-Economic Panel (SOEP), data for years 1984-2013, version 30i, SOEP, 2015*. Appendix 1.A.3.3 provides details on survey variables and matching with the main data.

practice, workers can be grouped by pathway and year of birth to capture this variation. Workers born during reform periods where policy varies at the monthly level are grouped by pathway and month of birth instead. The sample split yields 375 groups each of whom faces a distinct set of statutory ages and lifetime budget constraint discontinuities. When analyzing contribution notches, groups by pathway and year of birth are further divided into those retiring at ages 55 to 60 and 60 to 65 in order to capture variation of notch sizes with retirement age. In total, bunching is estimated at 644 discontinuities, among which there are 386 statutory ages and 258 pure financial incentive discontinuities. For the analysis across discontinuities, bunching observations are collected in the *bunching sample*, where each observation represents a discontinuity faced by a particular group of workers.

Table 1.2 shows summary statistics for the individual sample in column (1), for the occupation-matched sample in column (2) and for the bunching sample in column (3). In spite of the varying sample restrictions, key observables are relatively balanced across the different samples. Table 1.3 summarizes the budget constraint discontinuities in the bunching sample. Across all statutory age kinks, the average "kink size" is 0.08, i.e. the net-of-tax rate is reduced by 8 percent at the threshold. This is driven by a combination of convex kinks at ERAs and FRAs with average size between 0.3 and 0.4, and NRAs which feature non-convex kinks of average size -0.35. At pure financial incentive discontinuities, the average change in the net-of-tax rate is around 0.5, and the contribution notches entail an average approximate kink size of 0.9. There is also some within-group variation in the effective size of discontinuities due to different individual earnings histories, but the within-group standard deviations tend to be small.

1.3 Empirical Methodology

1.3.1 Basic Bunching Method

The first step of the empirical analysis is to measure retirement responses at each discontinuity. The bunching method developed by Saez (2010) and Chetty et al. (2011), which can be applied to the retirement age distribution,[16] provides a way of detecting such responses. A bunching strategy is naturally suited to the present context, since excess retirements measure both responses to kinks in the budget constraint and any other impact of certain thresholds on retirement. An additional advantage of the method is that it allows for the identification of responses within groups where all individuals face the same incentives.

The bunching mass B at an age threshold $\hat{R}$ can be measured as the observed local spike in the density of retirement ages above a counterfactual density $h_0(\hat{R})$. The standard approach to estimate $h_0(\hat{R})$ is to fit a flexible polynomial to the observed density excluding the threshold. The excess mass $b = B/h_0(\hat{R})$ is computed as the bunching mass relative to the counterfactual. While B measures the absolute number of excess retirements at $\hat{R}$, b expresses bunching in multiples of the counterfactual and can thus be compared across thresholds.

[16] See e.g. Brown (2013) and Manoli and Weber (2016*a*) for previous work on retirement bunching.

Assuming that the density would have been smooth in the absence of the threshold,[17] bunching can be interpreted in terms of a local retirement response. A standard approach focused on pure price changes then computes an elasticity by relating the excess mass to the kink size defined as the local percentage change in the implicit net-of-tax rate $\frac{\Delta\tau}{1-\tau}$. The elasticity of the retirement age with respect to the net-of-tax rate can be calculated as

$$\hat{\varepsilon} = \frac{b/\hat{R}}{\Delta\tau/(1-\tau)} \tag{1.2}$$

The formula is based on the insight that the excess mass is directly related to the labor supply response of the marginal bunching individual (Saez 2010), i.e. $b \approx \Delta R$. Elasticities computed according to (1.2) are referred to as *observed elasticities* for the remainder of the chapter.

1.3.2 Estimation Using Multiple Bunching Observations

The observed elasticity $\hat{\varepsilon}$ corresponds to a structural labor supply elasticity in a frictionless model without any responses to non-price factors. In such a model, bunching is only a function of the elasticity and a vector of observable variables x related to the threshold, including the counterfactual density and the kink size. Following the notation of Kleven (2016), $B = B(\varepsilon, x)$, and ε can be estimated using a single bunching observation as above. However, the recent literature has cast doubt on the structural interpretation of observed elasticities, and moved towards estimating additional parameters to explain differences in bunching across kinks. Writing bunching at threshold i as $B_i = B(\varepsilon, \omega, x_i)$, where ω is a vector of k additional parameters, identification necessitates observing $n \geq k+1$ bunching moments. If $n = k+1$, the implied system of n equations has an exact solution given the set of observed bunching moments. If $n > k+1$, parameters can be estimated across "bunching observations" B_i.

Existing studies focus mostly on optimization frictions (e.g. Chetty et al. 2011, Kleven and Waseem 2013, Gelber et al. 2017), where ω contains parameters such as a fraction of workers unable to adjust or a fixed cost of adjustment. This chapter, in contrast, is interested in estimating the effect of statutory ages on bunching, which is later interpreted in terms of reference dependence. Denoting D_i^s an indicator for the presence of a statutory age at bunching threshold i,

$$B_i = B(\varepsilon, \omega(D_i^s), x_i) \tag{1.3}$$

Hence, the presence of a statutory age affects bunching via ω. Parameters can be identified when bunching is observed at sufficiently many thresholds that vary in D_i^s and x_i under the following assumption:

> ASSUMPTION A. $E(\varepsilon_i | D_i^s) = \varepsilon$. *That is, structural elasticities do not vary systematically between statutory age thresholds and pure financial incentive discontinuities.*

[17] The empirical implementation allows for round number effects at the threshold. See appendix 1.A.4.1 for details of bunching estimation in practice.

Intuitively, the assumption rules out that stronger responses to financial incentives are falsely interpreted as statutory age effects. Note that the assumption is concerned with underlying structural elasticities, which differ from observed elasticities estimated according to (1.2) in the presence of statutory age effects. Indeed, equations (1.2) and (1.3) imply $\hat{\varepsilon} = f(\varepsilon, \omega(D_i^s))$, such that differences in observed elasticities are a corollary of the equations. An observed elasticity overestimates the true elasticity if some of the bunching occurs due to non-financial factors.[18] It is also important to note that the bunching approach generally allows for heterogeneity in underlying elasticities (and other parameters). In this case, bunching identifies an average retirement response, and local average parameter values at the threshold (see Kleven (2016)).

Within-group Estimation. For part of the analysis, parameters can be estimated within groups indexed by g:

$$B_{ig} = B(\varepsilon_g, \omega_g(D_{ig}^s), x_{ig}) \tag{1.4}$$

This requires observing bunching both at statutory ages and pure financial incentive discontinuities for the same group of workers g. Restricting the analysis to groups of workers facing both types of discontinuities allows for identification under a weaker assumption.

ASSUMPTION B. $E(\varepsilon_{ig}|D_{ig}^s) = \varepsilon_g$. *That is, a given group of workers g exhibits the same structural elasticity at statutory age thresholds and pure financial incentive discontinuities.*

Hence, elasticities can vary across groups in unrestricted ways, but a given group of workers are required to respond to all financial incentives in the same manner.

Optimization Frictions. Evidence from previous work indicates that optimization frictions seem to play a relatively minor role for the timing of retirement (e.g. Manoli and Weber 2016*b*). More generally, extensive margin responses are less subject to frictions than intensive margin responses (Chetty 2012). These findings are also mirrored by the sharp retirement responses documented in this chapter. However, it is not necessary to assume that there are no frictions for the purpose of the above analysis. Denoting a vector of friction parameters by ϕ, if $B_i = B(\varepsilon, \omega(D_i^s), \phi, x_i)$, the additional assumption necessary to identify a statutory age effect is that frictions do not vary systematically with D_i^s. In other words, if frictions attenuate responses to different thresholds in the same way, the relative magnitude of the effects of interest can still be identified.[19]

[18] This contrasts to a situation with optimization frictions, where the observed elasticity underestimates the true elasticity.

[19] For instance, this would be given if there was a constant share of non-optimizers, leading to a proportional attenuation of bunching as in Kleven and Waseem (2013).

1.4 Reduced-Form Evidence

1.4.1 Basic Bunching Analysis

1.4.1.1 Bunching at Specific Discontinuities: Some Cases

I begin by presenting some cases of bunching at specific discontinuities in order to illustrate the variation in the data. In particular, this section focuses on cases that lend themselves to two natural comparisons between statutory ages and pure financial incentive discontinuities.

Statutory Age vs. Contribution Notch Within Group. First, panels A1 and A2 of figure 1.4 show that the same group of workers responds more strongly to a discontinuity linked to a statutory age than to pure financial incentives. Panel A1 plots the job exit age distribution of women born in 1945 and 1946 around their ERA of 60. The average kink size is 0.08, implying an 8% reduction in the implicit net-of-tax rate at the threshold.[20] There is large excess mass of 12.3 and the observed retirement age elasticity calculated according to equation (1.2) is 4.45. Panel A2 shows the distribution of years of contributions of women in the same birth cohorts around the threshold of 15 years required for the women's pathway. At 14 years and 11 months of contributions, women face a notch of size 1.007, i.e. they gain an average of 0.7% of lifetime wealth from working an additional month. Following Kleven and Waseem (2013), the notch can be approximated as a kink for the marginal buncher. Here, the notch corresponds approximately to a kink of size 0.38. Indeed, there is sharp bunching at 15 years and some missing mass to the left of the notch. However, the excess mass of 1.32 is significantly less than that at the ERA in panel A1 where workers face smaller kink. The observed elasticity of 0.12 is much smaller than that of the same group at the ERA.

Kinks in Disability vs. Invalidity Pathways. For the second comparison, panels B1 and B2 show bunching at two very similar kinks, with and without the presence of a statutory age. Panel B1 shows bunching around the FRA at 63 for cohorts 1945 and 1946 in the disability pathway. The kink size is 0.51 and the excess mass is estimated at 10.5, which implies an observed elasticity of 0.67. Panel B2 shows the distribution of job exit ages for workers born between 1938 and 1946 in the invalidity pathway. They face a kink of size 0.43 at age 63. Consequently, workers in panels B1 and B2 face very similar kinks at the same age, but the threshold is not framed as the FRA in the invalidity pathway. Indeed, responses are very different. In contrast to the large excess mass at the FRA, bunching is hardly visible and the excess mass is only 0.08 at the invalidity kink. Consequently, the observed elasticity of 0.006 is far below the estimate at the FRA.

1.4.1.2 Bunching Across all Discontinuities

Table 1.4 summarizes responses across all 644 discontinuities in the data. In columns (1), the average excess mass of 21.8 across the 386 budget set kinks linked to statutory ages is very large.

[20] See appendix figure 1.A1 for lifetime budget constraints of the groups shown in figure 1.4.

Columns (2) to (4) show that it is driven by large responses at all three types of statutory ages. Attributing all bunching to the discontinuity in the implicit net-of-tax rate implies an average observed elasticity of 1.64.[21] This observed elasticity is two orders of magnitude above previous estimates of around 0.01 to 0.04 by Brown (2013) and Manoli and Weber (2016*a*) from pure financial incentives. Moreover, a first indication that bunching seems to occur somewhat irrespectively of the financial incentive is given by the large excess mass at the non-convex NRA kinks.

Next, columns (5) to (7) report bunching responses at the 258 pure financial incentive discontinuities. The average excess mass is 2.99. The average observed elasticity is 0.01 at pure financial incentive kinks, and 0.22 at contribution notches.[22] Averaging across all pure financial incentive discontinuities yields an elasticity of 0.15, compared to 1.64 at statutory age kinks in column (1). This implies that, conditional on kink size, bunching at statutory ages is more than ten times larger.

To further investigate to what extent differences in bunching are driven by differences in financial incentives, figure 1.5 shows binned scatterplots of the excess mass at a discontinuity against kink size. Two main insights emerge from the figure. First, financial incentives cannot explain the bunching patterns well. In panel A, there is large bunching at statutory ages independently of the underlying incentive.[23] The estimated slope is positive but insignificant and there is large bunching across all kink sizes, including close to zero and even negative ones. The second insight is that statutory ages matter directly for bunching. There are much larger responses at statutory ages in panel A than at pure financial incentives in panel B for any given kink size. Note that this does not necessary imply that financial incentives do not matter. Panel B shows a significant positive relationship between bunching at pure financial incentive discontinuities and the underlying kink size. The estimated slope corresponds to a difference-in-bunching elasticity of 0.18. However, even at the largest pure financial incentive discontinuities there is less bunching than at statutory ages.

1.4.2 Reduced-Form Estimation

The analysis so far suggests that there is large amount of additional bunching at discontinuities linked to statutory ages. In order to gauge the quantitative importance of this "statutory age effect", I use the following regression specification:

$$\frac{b_i}{\hat{R}_i} = \varepsilon \frac{\Delta\tau_i}{1-\tau_i} + \sum_s \beta^s D_i^s + Z_i'\gamma + \nu_i \tag{1.5}$$

[21] Note that non-convex NRA kinks are not included in the elasticity estimation since bunching in response to those would imply a negative elasticity.

[22] This difference could be driven by several factors. First, kinks apply to the invalidity pathway where workers may display a lower true elasticity than in other pathways. Second, observed elasticities measured at notches represent an upper bound: Kleven and Waseem (2013) point out that the approximation of the notch as a kink for the marginal buncher in order to compute a reduced-form elasticity undestimates the size of the discontinuity since everyone between the marginal buncher and the notch faces a larger change in the marginal tax rate. Third, additional months of contributions could come from some non-work periods such that workers may have additional margins of adjustment to bunch at contribution notches.

[23] Appendix figure 1.A2 shows plots separately by statutory age types, suggesting that the flat slope is driven by a combination of positive slopes at ERAs and FRAs and a negative slope at NRAs.

where an observation indexed by i corresponds to a discontinuity in the bunching sample. D_i^s is an indicator for a statutory age of type $s \in (ERA, FRA, NRA)$ attached to discontinuity i, and the coefficients β^s measure the additional bunching due to the respective statutory age type. Finally, Z_i is a vector of control variables, and ν_i is an error term.

Equation (1.5) may be a natural way to detect a reduced-form statutory age effect, but it can be also be interpreted as a simple, linear version of the bunching equation (1.3), where the parameter vector ω consists of a set of linear regression coefficients on the dummies D_i^s. The empirical setting provides many more bunching observations than parameters in the equation, which has two advantages. First, additional regressors can be included, allowing to control for a number of group-level characteristics and fixed effects in a flexible way. Second, rather than finding a solution to an exactly identified system of bunching equations, the equation can be estimated via OLS, combining the information provided by all available bunching moments. Intuitively, the specification fits a regression to the discontinuity-level data, where the slope is interpreted as the elasticity and additional intercepts are interpreted as statutory age effects. Hence, statutory age effects are identified from the difference in bunching *between* statutory age kinks and pure financial incentive discontinuities while the elasticity is identified from bunching at kinks of different sizes *within* each type of discontinuity. Standard errors are obtained via bootstrap by re-sampling bunching observations.[24]

The key identification assumption can be phrased in terms of this regression. ν_i needs to be uncorrelated with the regressors, which requires assumption A. To see this, consider a case where true elasticities vary across discontinuities with $D_i = 0$ and $D_i = 1$. Then $E(\nu_i|D_i) \neq 0$, since ν_i contains some residual bunching not captured by the average elasticity ε, and this would introduce bias into the estimation of β. In practice, the inclusion of control variables and fixed effects somewhat weakens the required assumption, such that elasticities should be independent of D_i conditional on these controls. Note that some support for assumption A is lent by the results from figure 1.5, where the implied difference-in-bunching elasticities at statutory ages and other discontinuities are relatively similar.

Table 1.5 reports results from regressions based on equation (1.5). To begin with, column (1) shows results from a basic specification without controls. This yields an elasticity of 0.11 and large and significant statutory age effects. Next, column (2) adds interactions between different statutory age types in order to account for the fact that more than one type is present at some discontinuities. Columns (3) adds a set of worker controls, as well as pathway and year-of-birth fixed effects accounting for the dimensions along which groups are defined. Column (4) adds the maximum set of group fixed effects, controlling for pathway times year-of-birth fixed effects. Finally, column (5) controls for occupation-level characteristics including firm size and unionization rates. In spite of the varying set of controls and fixed effects, the estimated statutory age effects remain at similar magnitudes. These coefficients do not have a direct interpretation, but magnitudes can

[24] This corresponds to a block bootstrap procedure on the individual-level data, where blocks are defined by groups of workers facing the same discontinuity.

be compared. With a coefficient of 0.8, the NRA has the largest effect on bunching, while the FRA effect is around 0.3 and the ERA effect is 0.2. Elasticity estimates remain within a narrow range between 0.05 and 0.1.

1.4.3 Heterogeneity

An important advantage of the empirical setting is that the large number of discontinuities allows for an exploration into the determinants of bunching beyond the scope of existing bunching studies. This section examines heterogeneity along a number of dimensions. The main finding is that the "statutory age effect" found in the previous section does not seem to be confounded by observable factors. Moreover, results from an additional estimation strategy are presented, where parameters are allowed to vary across groups.

Worker and Firm Characteristics. Besides varying incentives, differences in bunching between statutory ages and other discontinuities could be driven by differences in elasticities across groups. Figure 1.6 shows average observed bunching elasticities at statutory ages and pure financial incentive discontinuities by a range of observables. Panels A to C focus on worker characteristics, namely lifetime wealth (panel A), years of schooling (panel B), and health status proxied by sick leave periods (panel C). Workers are grouped by quintiles of each variable. Those with higher lifetime wealth and higher education seem to respond less strongly at statutory ages, but more strongly at other discontinuities. In particular, workers in the highest schooling quintile seem to respond more to pure financial incentives. Groups in worse health, on the other hand, are less responsive to both statutory ages and other discontinuities. Panels D to F sort bunching observations by some occupation-level characteristics, in particular firm size (panel D), unionization rate (panel E) and tenure in the firm (panel F). Recall that these characteristics are obtained by matching the individual data with SOEP data at the 3-digit occupation level. Again, the gap between bunching at statutory ages and other discontinuities differs somewhat across groups. However, observed bunching elasticities are higher at statutory ages by at least a factor of two in each quintile of each variable. Finally, panels G and H sort bunching observations by birth cohort and the retirement age at the discontinuity. Responses are substantially larger at statutory ages for each birth cohort and across the range of available retirement ages.

Appendix figures 1.A3 and 1.A4 provide two pieces of additional evidence. First, figure 1.A3 shows that the main results hold when considering heterogeneity closer to the group definitions in the bunching sample, namely by retirement pathway, year of birth and the age at the discontinuity. Both excess mass and observed elasticities are always larger at statutory ages than at pure financial incentive discontinuities. Second, figure 1.A4 shows that the patterns above hold when considering raw excess mass rather than observed bunching elasticities.

Explanatory Power of Observed Characteristics. In order to quantify the explanatory power of firm incentives and other observable variables, table 1.6 reports results from a Oaxaca-Blinder decomposition. Bunching observations are grouped into discontinuities linked to statutory

ages and pure financial incentive discontinuities. The decomposition attributes differences in excess mass across groups to a component explained by differences in observables and an unexplained component. Since results vary with the choice of reference group, the table reports results using statutory ages as the reference group in column (1), pure financial incentive discontinuities as the reference group in column (2) and a weighted average of the two in column (3). Results confirm that most of the additional bunching at statutory ages cannot be explained by observable factors. Financial incentives account for a maximum of 23% of observed differences, while worker and firm variables including those discussed above explain up to 12% and 15%, respectively. Between 64% and 104% of the additional bunching at statutory ages cannot be explained by differences in observable characteristics.

Estimation with Heterogeneous Parameters. In the reduced-form estimation presented in section 1.4.2, a concern for identification arises if parameters are heterogeneous across workers facing different types of discontinuities. Adding fixed effects somewhat alleviates the concern by allowing for group-specific bunching intercepts. However, a more direct way to address this is to allow for heterogeneous parameters in the following specification:

$$\frac{b_{ig}}{\hat{R}_{ig}} = \varepsilon_g \frac{\Delta\tau_{ig}}{1-\tau_{ig}} + \sum_s \beta_g^s D_{ig}^s + \nu_{ig} \tag{1.6}$$

where g indexes groups. Since the main issue with the previous specification is heterogeneity correlated with kink sizes and statutory ages, a natural solution is to allow for heterogeneous parameters at the level where these variables are determined, namely pathway and year of birth. This strategy corresponds to a linear version of the within-group bunching equation (1.4), which requires the weaker identification assumption B. The assumption states that the same group of workers exhibits the same elasticity at different types of discontinuities, while true elasticities can vary arbitrarily across groups.[25]

Table 1.7 reports results from estimating equation (1.6) with varying group definitions. Note that table 1.7 reports weighted averages of coefficients, while pathway- and cohort-specific estimates are shown in appendix table 1.A1. First, column (1) replicates the basic specification with homogenous parameters. Column (2) estimates a specification with pathway-specific coefficients, and column (3) repeats the exercise with groups defined by birth cohorts. In both specifications, the elasticity estimate increases somewhat compared to column (1), but statutory age effects remain significant and increase slightly in magnitude. Column (4) reports estimates with groups defined by pathway and birth cohort. In the spirit of the comparison presented in figure 1.4, this specification estimates elasticities and statutory age effects within narrowly defined groups such as women born in 1945. Again, the estimates in column (4) are similar to the previous columns. Appendix table

[25] Also note that if underlying parameters are heterogeneous, even if its idenfitication assumption is satisfied, equation (1.5) identifies weighted averages where the parameter of each group is weighted by the conditional variance of the corresponding regressor within the group (see Angrist 1998). These weighted averages are not necessarily equal to "true" population averages. However, the coefficients identified in equation (1.6) can be used to calculate population averages by weighting estimates by group size.

1.A1 suggests some parameter heterogeneity across pathways, and little heterogeneity across birth cohorts, which is in line with the evidence from appendix figure 1.A3. However, the fact that average parameters in table 1.5 change little when allowing for more heterogeneity indicates that there is little bias in the basic specification with homogeneous coefficients. Overall, the results suggest large and significant statutory age effects, and elasticity estimates are between 0.09 and 0.29.

1.5 Mechanisms

This section discusses potential mechanisms behind the "statutory age effect" found by the preceding analysis. I argue that the framing of statutory ages as reference points is the most plausible mechanism for a number of reasons. First, section 1.5.1 shows evidence suggestive of framing effects and argues that mistakes or a lack of information are unlikely to explain the observed patterns. Second, the mechanism is consistent with the institutional setting presented in section 1.2.2, where statutory ages are *institutional* reference points in the framing of benefits and retirement. This is likely to facilitate the formation of goals or expectations, which is typically interpreted as reference-dependent preferences.[26] Third, there is complementary evidence from surveys and experiments on framing effects. For instance, Shoven et al. (2017) suggest that retiring at the NRA may be perceived as a norm, and Merkle et al. (2017) find experimental support for framing effects and behavior consistent with reference dependence around statutory ages.

Fourth, alternative mechanisms including firm responses are addressed in details in section 1.5.2 and 1.5.3, and they do not seem to explain much of the effect. Finally, in the model presented in chapter 2 of this thesis, reference dependence is a fairly general way the effect of interest: a discontinuity in utility may capture a number of potential sources, including framing effects, but also norms, "deep" internal preferences, or audience effects.[27]

1.5.1 Workers, Framing and Information

Worker characteristics. First, it may be interesting to examine which types of workers are most likely to retire at statutory ages. Table 1.8 shows regressions of dummies for bunching at different types of discontinuities on a number of worker characteristics.[28] If statutory age retirements were driven by a lack of information or mistakes, one might expect that workers with lower financial literacy are more likely to choose these ages. In column (1), however, workers retiring at statutory ages have *higher* education and are *more* likely to be economically trained. They also

[26] In contrast to other decision environments, there is no status quo or previous outcomes in terms of the retirement date. In the absence of such backward-looking experience, and faced with the difficulty of predicting one's retirement date, it may be natural to take public "advice" or perceived norms as a reference point for retirement.

[27] Although I argue in favor of framing effects, it remains outside the scope of this chapter to offer fully conclusive proof of the exact source of reference dependence. Instead, the model is devised in a "pragmatic approach" (Chetty 2015) in the sense that it may not necessary to fully specify the sources of behavior to draw conclusions about the effects of a policy such as a statutory age, and to provide useful predictions regarding counterfactual policies.

[28] While the heterogeneity analysis in section 1.4.3 tests whether differences in responses across discontinuities can be explained by differences in observables, the specification presented here asks which individual workers are more likely to be among the bunchers.

have higher lifetime earnings, higher last earnings before retirement and higher pension wealth. Interpreting these characteristics as a proxy for financial literacy, this suggests that a pure information mechanism does not seem to be at work. Columns (2) and (3) show that workers retiring at the ERA/FRA are less educated but more likely economically trained, while retiring at the NRA is associated with higher education but negatively related to economic training. All types of statutory ages seem to attract workers with higher earnings. Hence, even though there is usually a disincentive to retire, more educated and higher-income workers are more likely to choose the NRA. Economic training somewhat reduces the probability of bunching at the NRA, but the effect is limited in magnitude.

Columns (4) repeats the exercise with pure financial incentive discontinuities. These are positively associated with education and economic training, but negatively with lifetime earnings. Finally, column (5) focuses on other round ages. Figures 1.1 and 1.4 indicate clear bunching at round ages not linked to statutory ages. Round-number bunching has been observed in a number of contexts and has been attributed to reference point effects (see Kleven 2016). Hence, they may offer an interesting point of comparison. Round-number bunchers are less educated, more likely to economically trained and with higher earnings, which makes them similar to workers bunching at the ERA/FRA.

Information Provision and Framing. As explained in section 1.2.2, the way information is provided to workers and how it is framed could play an important role for retirements at statutory ages. To test this, I exploit a reform implemented in Germany in the early 2000s, where the state pension fund drastically increased the frequency of information letters to workers. Before June 2002, a detailed letter was sent on workers' 55th birthdays. Under the new regime phased in between June 2002 and December 2003, a basic letter is sent to every worker every year, and detailed letters are sent every 3 years from age 55. The stated goal of the reform was to inform workers better about benefits and retirement. Letters are personalized and they provide detailed information on the worker's contributions so far, how pensions are calculated, and some guidance on making intertemporal decisions.[29] Projected benefit amounts for the individual at different retirement ages are also shown. However, letters also seem to reinforce the framing of retirement in terms of statutory ages. In particular, workers are shown the date when they will reach the NRA at the beginning of the letter, and two out of three benefit scenarios use the NRA as a reference point.

Figure 1.7 shows monthly fractions bunching at different types of discontinuities around the reform. In panel A, there is no change in the response to pure financial incentives, but there is an *increase* in the probability of bunching at statutory ages. Panel B show that this is driven by more workers bunching at the NRA, whereas the probability of bunching at the ERA/FRA is constant throughout the reform. Since letters specifically emphasize the NRA, this pattern may imply an additional framing effect. Appendix table 1.A2 shows results from a version of the main estimation

[29] Appendix figure 1.A5 shows an example letter. See also appendix 1.A.2.4 for further details on the content of letters and information provision.

with effects of annual information letters. The insignificant coefficients on the interaction between kink size and annual letters confirm that additional information does not affect the response to financial incentives. However, letters increase the response to statutory ages in columns (1) and (2). Columns (3) and (4) suggest that this is indeed driven by larger responses at the NRA and possibly the FRA. Overall, two conclusions arise. First, increasing the frequency of information provision does not increase responses to financial incentives. Second, framing retirement around statutory ages seems to induce more workers to retire at these ages.

Mistakes and Inattention. A potential alternative to framing effects and reference dependence may be mistakes, for instance because workers misperceive the incentives linked to statutory ages. However, this would require that workers perceive a strong incentive to retire at all statutory ages, including NRAs that entail a disincentive to retire. In addition, the size of perceived incentives would have to extremely large: a back-of-the-envelope calculation suggests that in order to generate the observed excess mass at statutory ages, workers would have to perceive an average kink size of at least 95%.[30] Such a large kink does not exist anywhere in the pension system, and actual kink sizes at statutory ages vary between -40% and 50%.

Moreover, when optimization requires costly attention, one may expect that the frequency of mistakes depends on the stake workers have in the decision. Retirement is generally a high-stake decision with large consequences in terms of lifetime consumption possibilities, but stakes vary across workers. For instance, some workers may have contributed to the public system only for part of their earnings career, or they may not be the main household earner. Table 1.8 shows that workers with higher stakes seem to be, if anything, more likely to bunch at statutory ages. The ratio of pension wealth to annual earnings, which proxies for the relative importance of public pensions for the worker, increases the probability of bunching at statutory ages. Married females, who tend to rely less on their own pension for old-age consumption, are less likely to bunch at statutory ages. Both effects also hold when focusing only on the NRA where mistakes may be particularly likely. Finally, another measure of stakes could be the size of the local incentive. For instance, if a kink is very small, choosing a "wrong" retirement age may not be very costly. This would imply that observed elasticities are increasing in kink sizes because it is more worthwhile to optimize at large kinks. Appendix figure 1.A6 shows binned scatterplots of observed elasticities vs. kink sizes. There is no evidence of larger responses at large kinks: across all types of discontinuities observed elasticities are flat or even decreasing with kink size. Appendix table 1.A3 reports results from corresponding regressions, confirming that observed elasticities do not increase with the kink size.

[30] This calculation assumes that the elasticity w.r.t. perceived kinks is the same as the measured elasticity w.r.t. actual kinks. Even if the elasticity w.r.t. perceived kinks was substantially larger, implausibly large perceived kink sizes would be needed to justify observed responses. Importantly, a large elasticity w.r.t. perceived kinks alone could not rationalize observed patterns, since bunching occurs at negative kink sizes as well.

1.5.2 The Role of Firms

Self-Employed and Very Small Firms. To begin with, I show that subgroups where firm incentives play a small role or no role at all also bunch at statutory ages. Two such groups are available in the present context. First, although limited, there is a number of self-employed individuals enrolled in the public pension system.[31] Second, very small firms with less than 10 employees are exempt from employment protection, which implies that there should be less firing frictions and hence less need for employers to use statutory ages to lay off older workers. Figure 1.8 shows job exit age distributions among the full occupation-matched sample (panel A), self-employed workers enrolled in the public pension system (panel B), the 20 occupations most frequently in very small firms, including medical receptionists, hairdressers, pharmacists, florists, and dental technicians (panel C), and the 20 occupations least frequently in small firms, including bankers, executives, machine operators, miners and train drivers (panel D). There are sharp spikes among the self-employed and the fraction of 28% who bunches at statutory ages is only marginally smaller than in panel A. Moreover, although the vast majority of contracts falls below the threshold for employment protection, there are also sharp spikes among workers most frequently in small firms, 31% of whom bunch at statutory ages. There is however somewhat more bunching among those least frequently in small firms. Hence, groups of workers where firm incentives play a small role or no role at all still exhibit large bunching, which suggests that firm responses are not the main driver of statutory age retirements.

Mandatory Retirement. The most direct way for firms to induce workers to retire at statutory ages is through mandatory retirement clauses linked to the NRA. A natural way to check whether the previous results are driven by this is to exclude all statutory ages from the analysis where this firm-induced mandatory retirement is possible. Appendix table 1.A4 shows results from regressions analogous to table 1.7, but excluding all discontinuities linked to a NRA. Results are virtually unchanged, with elasticity estimates ranging from 0.14 to 0.26, and highly significant statutory age effects at the ERA and FRA similar in magnitude to those in table 1.7. Thus, even in the unlikely case that all NRA job exits are driven by mandatory retirement, there remains a large portion of unexplained statutory ages retirements.

Proxying for Firm Incentives. To shed light on the role of firms more generally, statutory age retirements can be related to a number of variables proxying for firm incentives. First, firing frictions are more severe for larger firms since employment protection becomes stricter in line with certain size thresholds. Second, firing costs may change when workers are more unionized. Third, firing costs are higher for workers with longer tenure since employment protection increases as a function of tenure thresholds. Fourth, job exits at statutory ages are not necessary at all as a

[31] Self-employed individuals can be enrolled in the public pension system for two reasons. First, there may be part of a small set of occupations where enrollment is mandatory. This includes mainly craftspersons, workmen, self-employed teachers and educators, nurses and artists. Second, self-employed workers can voluntarily enrol in the public scheme.

tool for firms when workers have contracts that end automatically after a term limit. Finally, in a tighter labor market it may be more valuable to firms to keep older workers beyond statutory age thresholds.

Figure 1.9 show binned scatterplots of the fraction of workers bunching at statutory ages against the proxies discussed above. Panels A to D include firm size, unionization, tenure and the frequency of unlimited contracts at the occupation level. Labor market tightness in panel E is constructed from annual vacancy and unemployment data at the state level. The fraction bunching at statutory ages is large in all bins of the explanatory variables and the estimated slopes are relatively flat. While there seems to be no effect of unionization, the fraction bunching is indeed increasing in firm size, average tenure and the fraction of workers in unlimited contracts. Somewhat surprisingly, there seem to be more statutory age job exits in tighter labor markets.

Table 1.9 shows results from corresponding individual-level regressions of a dummy for statutory age job exits on all these variables, including worker controls as well as pathway and year-of-birth fixed effects. The positive relationship of statutory age retirements with firm size remains, while the effect of unionization, tenure and unlimited contract turns negative. Columns (2) and (3) report results for the NRA and the ERA/FRA separately. The NRA offers the clearest channel for firm responses, but results do not differ substantially from column (1). Overall, the probability that an individual worker retires at statutory ages seems to be only weakly related to firm incentives. Larger firms may induce some additional statutory age retirements, but the magnitude of this effect is limited.[32] This is consistent with the results from section 1.4.3, where firm-related variables explain only a small share of differences in bunching across types of discontinuities.

1.5.3 Other Checks

Benefit claiming at statutory ages. While job exits could be externally induced in some cases, the timing of benefit claiming is determined by workers themselves. In particular, if they were laid off but did not want to retire yet, one might expect that some affected workers delay benefit claiming beyond statutory ages in order to search for another job. Appendix figure 1.A7, panel A shows a histogram of the gap between job exit and benefit claiming among workers who exit at a statutory age (white bars), and workers who exit at other ages but would be eligible to claim immediately (red bars). Around half of workers exiting at other ages claim immediately, while a substantial shares seem to wait up to around 12 months to claim. For statutory age job exits, however, the fraction claiming immediately is a striking 95%. In addition, panel B shows the job exit age and claiming age distribution of workers not exiting at statutory ages who are eligible to claim immediately. There are sharp spikes in claiming at the main statutory age thresholds, in particular at age 65, which suggests that many workers seem to wait until a statutory age to claim, even if their job exit occurs at other ages.

[32] Of course these correlations do not prove a role for firms themselves. For instance, the effect of firm size may stem from stronger social norms or peer effects in larger firms, possibly enhancing reference point effects of statutory ages.

Salience of incentives. It could also be that statutory ages make underlying financial incentives more salient rather than serving as reference points. In other words, there could be more bunching at statutory ages than at other discontinuities because workers are more aware of the underlying budget set kink. A priori evidence against this hypothesis is provided by the results from figure 1.5, where large bunching at statutory ages occurs at all kink sizes, including negative ones. If incentives were made more salient by statutory ages, one would expect a different pattern, where bunching is more steeply increasing in kink size than at pure financial incentives. Appendix table 1.A5 provides a further test, repeating the analysis of table 1.7 with additional interactions of kink size with statutory age dummies. Interaction effects are insignificant or even negative, implying that workers do not respond more to financial incentives at statutory ages.

Finally, a number of other channels have been discussed in the literature, but are unlikely to play a role in the present context.

- **Default options.** In the German context, statutory age thresholds do not constitute a default option for retirement. Benefit claiming always requires an active choice in the form of an application by workers themselves.[33]
- **Liquidity constraints.** Liquidity constraints may be a potential explanation for job exits at the ERA. Since pension benefits are only paid from the ERA onwards, workers who exit their jobs before the ERA may have to use savings or borrow against their future pension in order to smooth lifetime consumption throughout the gap between job exit and ERA. This may not be possible to the desired extent with credit constraints. However, recent evidence by Goda et al. (2018) suggests that liquidity constraints are not the main driver of ERA retirements in the U.S. In addition, table 1.8 shows no indication of liquidity constraints at the ERA. Workers retiring at the ERA/FRA have both higher lifetime incomes and higher last incomes before retirement.
- **Health insurance.** Finally, health insurance availability has been suggested in the U.S. context as a potential driver of retirements at the NRA. However, in Germany there is universally mandated public health insurance that covers workers as well as pensioners. Hence, the availability of health insurance does not depend on age and is unlikely to be a driver of retirements at any particular age.

1.6 Conclusion

Recent years have seen a surge of interest in retirement decisions and their responsiveness to pension system features. While there have been studies on some reforms and idiosyncratic features, the overall evidence is inconclusive. This chapter aims at filling this gap by providing a comprehensive

[33] This contrasts with the setting of Lalive et al. (2017) where workers are retired by default at the FRA in Switzerland.

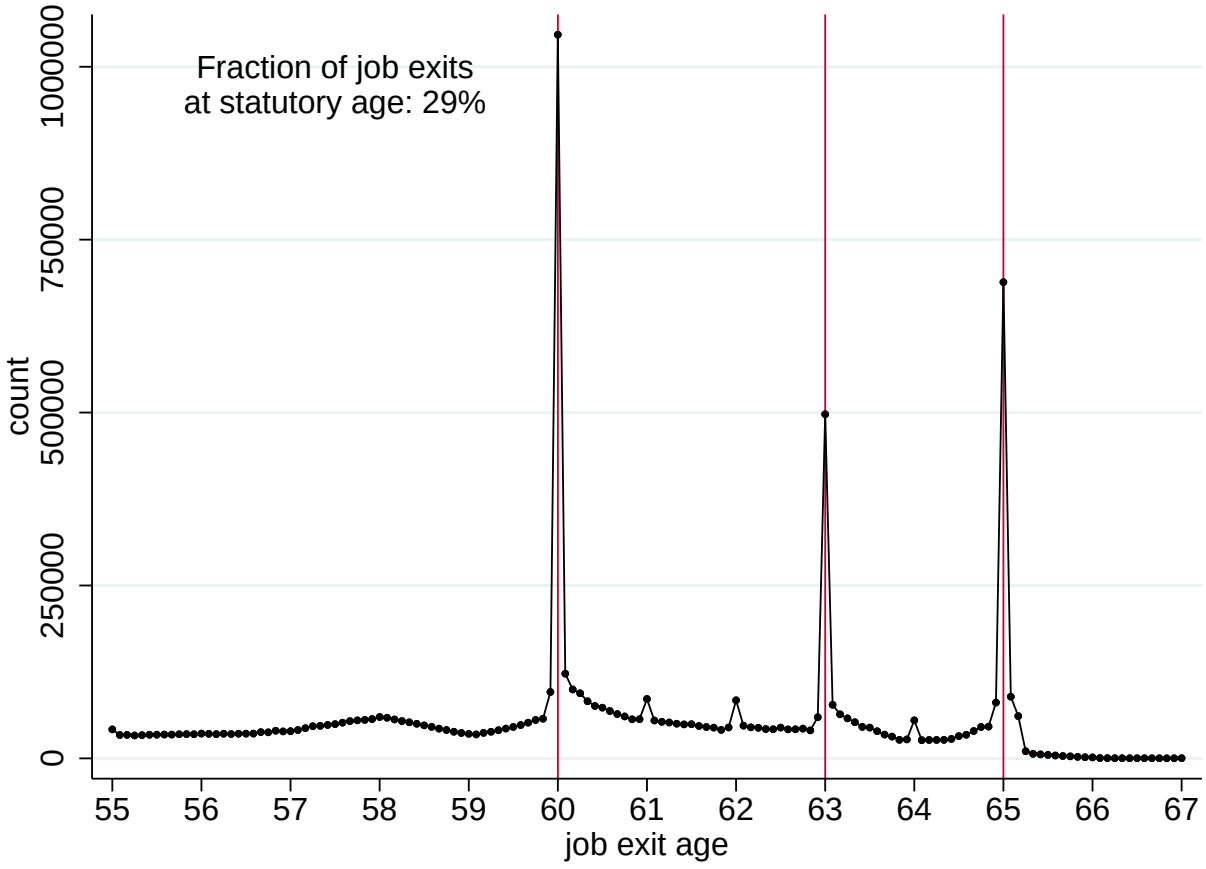
view of the effect of two key features of pension systems, namely statutory age thresholds and financial incentives. The results highlight the important role of statutory ages: around 30% of job exits occur at a statutory age, and the responses are largely driven by reference point effects. Nevertheless, workers also respond to financial incentives, as is particularly visible in bunching at notches. Job exit age elasticities with respect to the net-of-tax rate are larger than those found in previous studies, with estimates around 0.1 to 0.3.

There are implications for the design of pensions and reform options. Having established their direct impact on behavior, statutory age thresholds themselves can be viewed as a policy instrument independent of financial incentives. Chapter 2 shows in simulations that increasing average retirement ages can be an effective way to increase actual retirement ages with a positive fiscal effect. However, questions remain to what extent policy can exploit reference dependence with respect to established thresholds by arbitrarily shifting those, since workers' notion of reference points may originate to some degree from a "real" incentive. Moreover, this raises important new issues of distribution since certain types of workers seem to be more prone to responding to these reference points.

The result may also have more general implications for the interpretation of bunching patterns. Incentive schemes in other contexts are also framed by statutory thresholds that may serve as reference points. For instance, taxes and social insurance contributions are often defined in terms of thresholds, and many other policies divide the choice space into discrete categories rather than emphasizing continuous choices. This may reduce complexity and thus help individuals make decisions, but the results of this chapter highlight that strongly framed thresholds can become reference points somewhat independently of the original incentive.

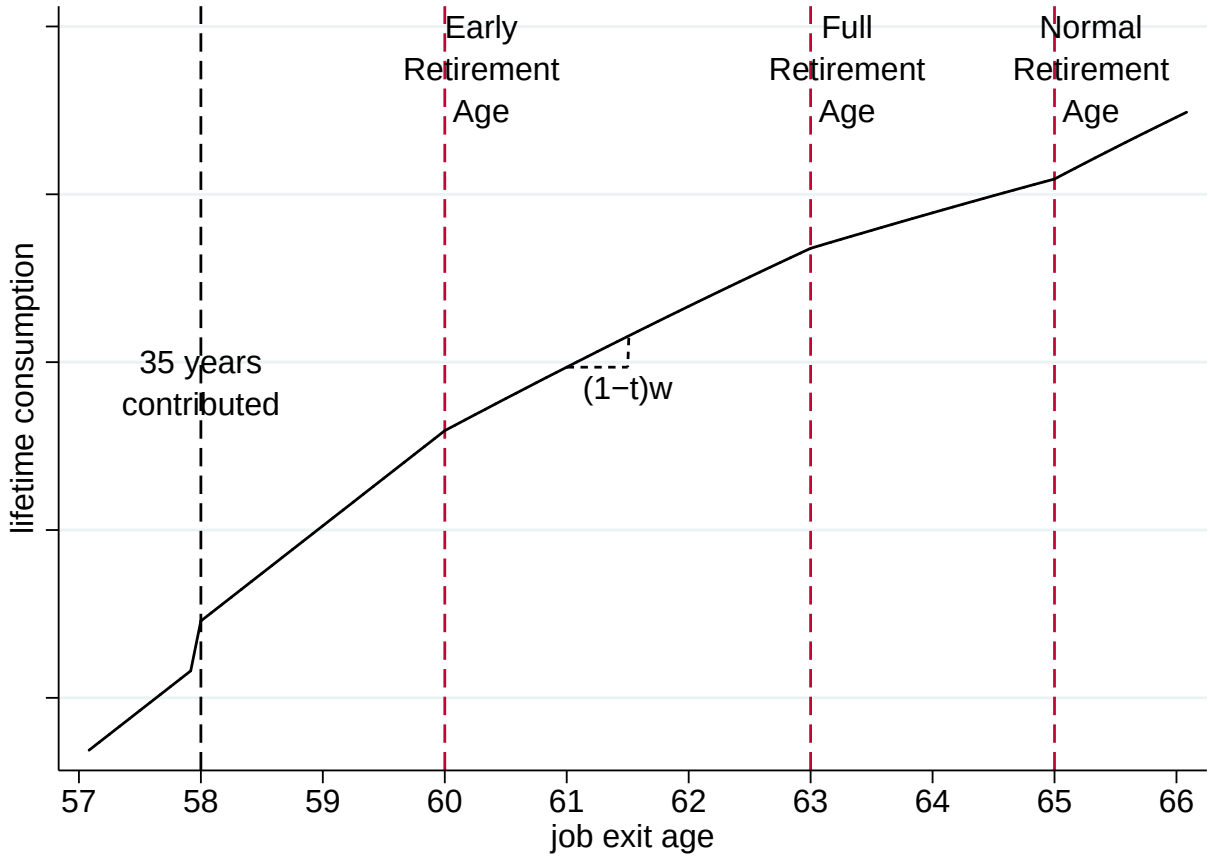
It may be worthwhile for future research to study the sources of reference-dependent retirement behavior more closely. For instance, this could be done via some controlled variation in the framing of benefits and retirement and the information environment more generally. Similarly, it may be interesting to test for individuals' perceptions and intentions more explicitly, for example via surveys. Finally, the degree to which workers internalize reference points set by the government may depend on cultural aspects, and it may be interesting to see how results compare across cultural contexts.

Figure 1.1: Job Exit Age Distribution (Full Sample)



Note: This figure shows the pooled distribution of job exit ages for all workers born between 1933 and 1948. The connected dots show the count of job exits within monthly bins. Vertical red lines indicate the location of main statutory ages throughout the sample period.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

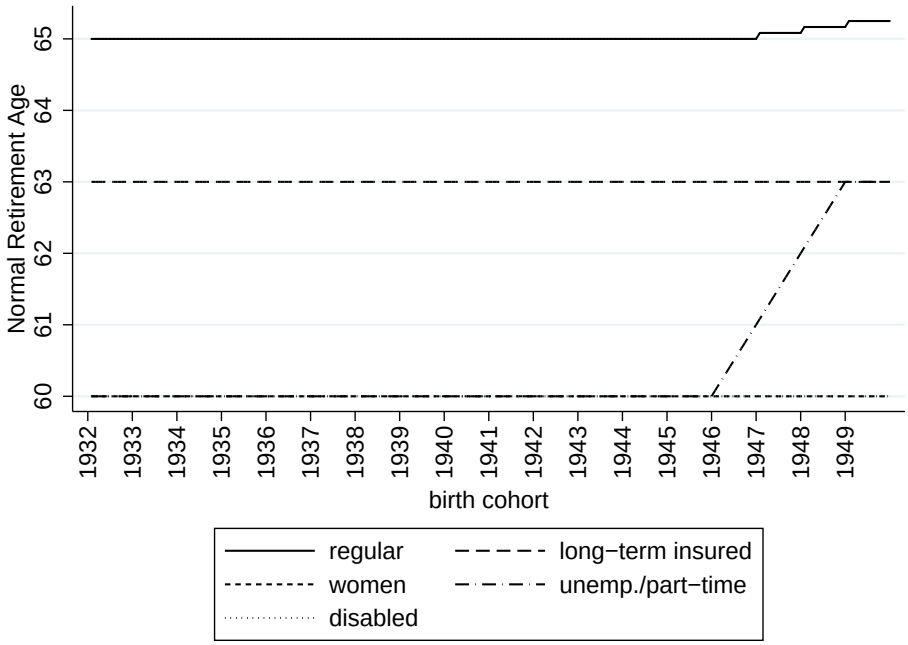
Figure 1.2: Stylized Lifetime Budget Constraint



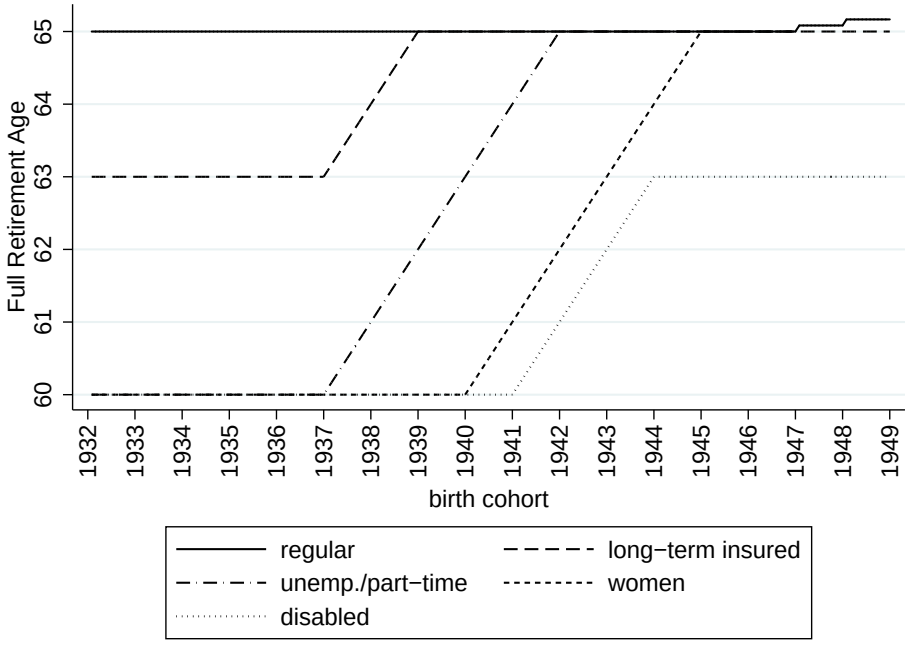
Note: The figure shows a stylized lifetime budget constraint for a worker who faces an Early Retirement Age of 60, a Full Retirement Age of 63 and an Normal Retirement Age of 65, who becomes eligible for a pathway requiring 35 years of contributions at age 58. The slope of the BC is the implicit net wage defined as $w_i^{net} = (1 - \tau_i)w_i$ as shown in section 1.2.3. The stylized shape of the constraint corresponds to incentives faced by the average worker: On average, workers face a 32% reduction in the implicit net wage ("kink size") at age 60, a 42% reduction at age 63%, and a 21% increase in the implicit net wage at age 65.

Figure 1.3: Evolution of Statutory Ages

Panel A: Early Retirement Ages (ERA)



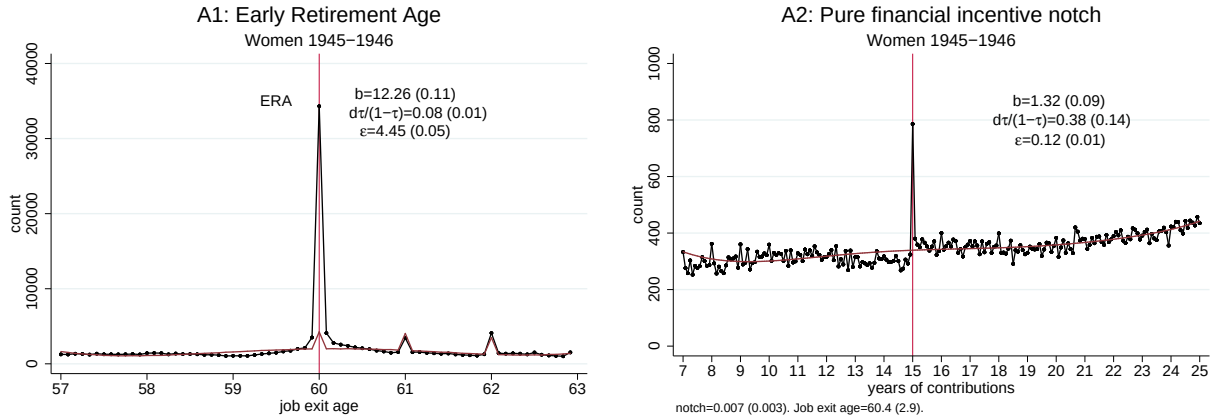
Panel B: Full Retirement Ages (FRA)



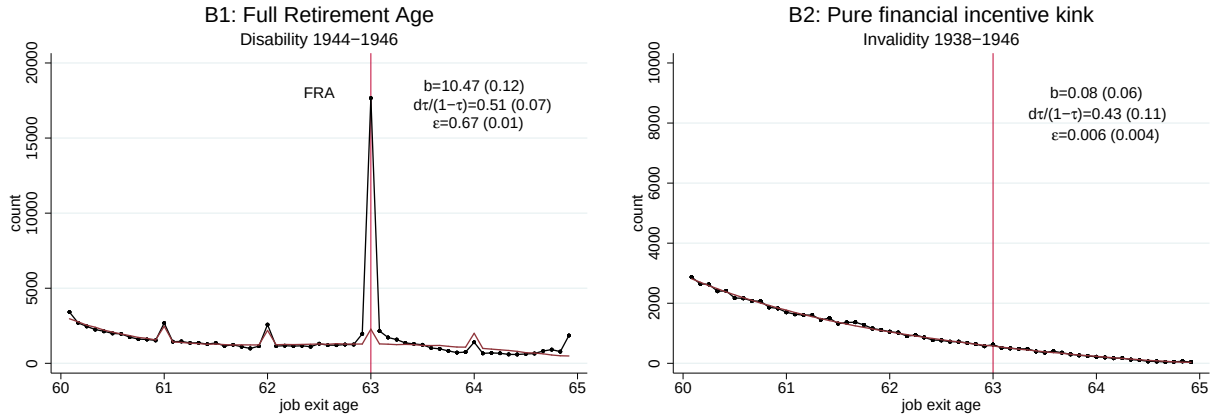
Note: The figures show the evolution of Early Retirement Ages (ERA) and Full Retirement Ages (FRA) of different pathways across monthly birth cohorts. In Panel A, the regular ERA is increased from 65 to 65/3 between 1947 and 1949 and the unemployed/part-time ERA is gradually increased from 60 to 63 between 1946 and 1948. In Panel B, the long-term insured FRA is increased from 63 to 65 between 1937 and 1938 and from 65 to 65/3 for cohort 1949, the women's FRA from 60 to 65 between 1940 and 1944, the unemployed/part-time FRA from 60 to 65 between 1937 and 1941, the disability FRA from 60 to 63 between 1941 and 1943, and the regular FRA 65 to 65/3 between 1947 and 1949. See table 1.1 for an overview of pathways.

Figure 1.4: Bunching at Specific Discontinuities

Panel A: Statutory age vs. pure financial incentive notch



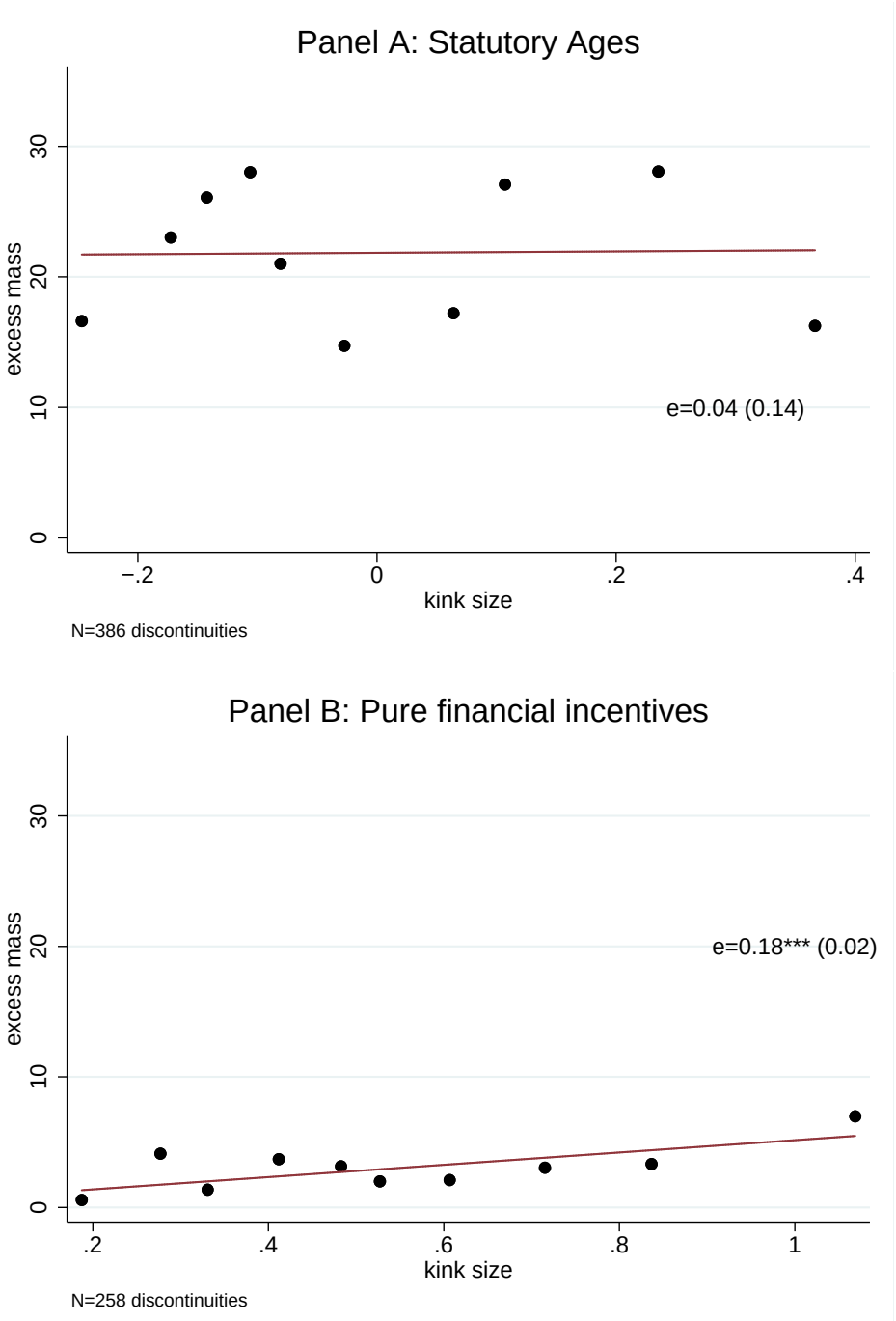
Panel B: Statutory age vs. pure financial incentive kink



Note: This figure shows bunching at some cases of specific discontinuities. Panel titles indicate the type of discontinuity and panel subtitles indicate pathways and birth cohorts used. In panels A1, B1 and B2, the connected black dots show counts of job exit ages in monthly bins for the group indicated by the respective panel title. In panel A2, the black dots show counts of years of contributions instead. In all panels, the red line shows the counterfactual distribution estimated as a 7th-order polynomial, including round-age dummies in panels A1 and B1. Vertical red lines indicate the location of the discontinuity. b is the excess mass, $d\tau/(1-\tau)$ is the change in the implicit net-of-tax rate at the discontinuity (kink size), and ε is the implied elasticity of the retirement age w.r.t. the implicit net-of-tax rate. See appendix figure 1.A1 for the lifetime budget constraints of the four groups.

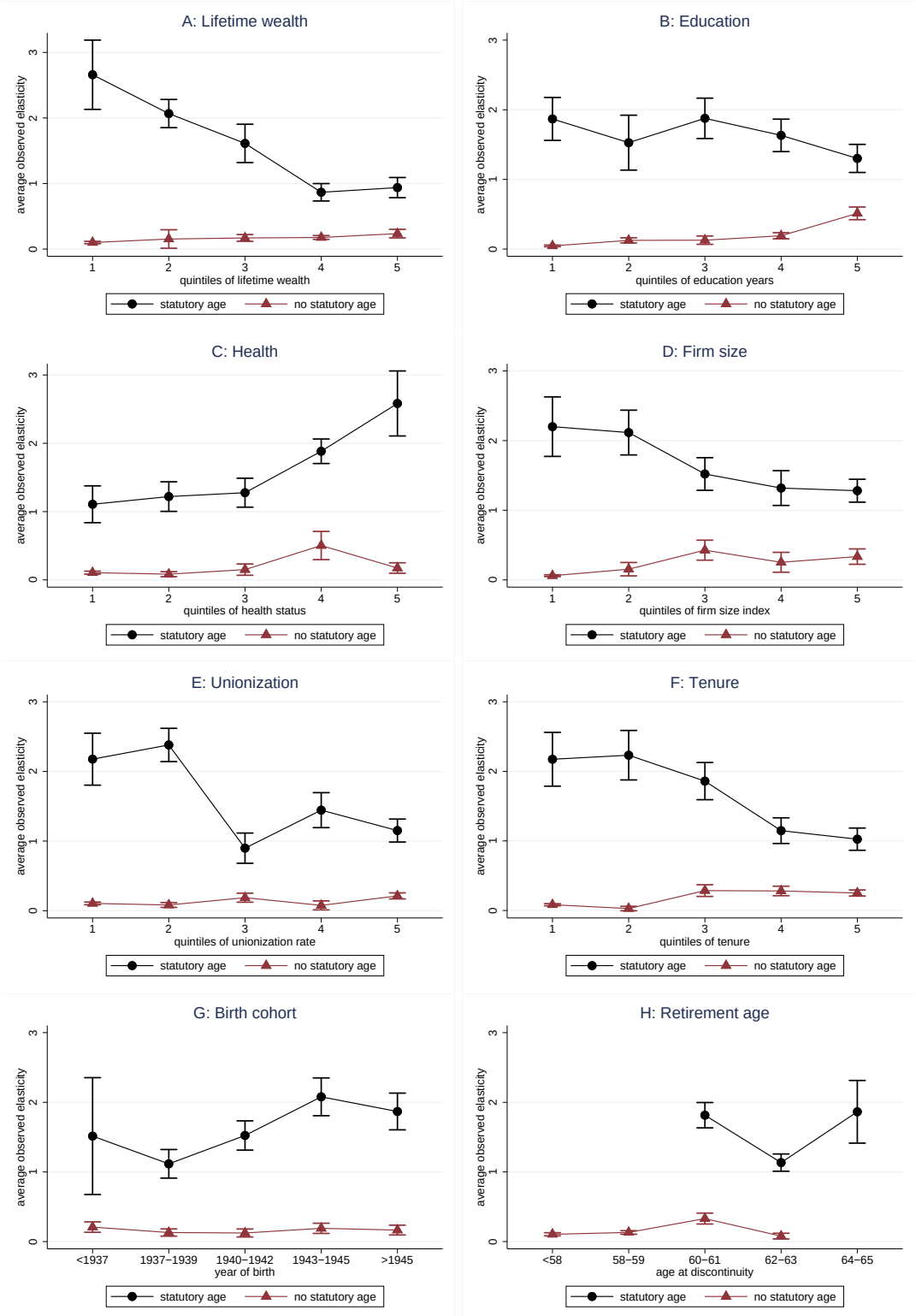
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Figure 1.5: Bunching and Financial Incentives



Note: The figure shows binned scatterplots of the excess mass at pure financial incentive discontinuities (panel A) and statutory ages (panel B) against kink size. In panel B, the type of statutory ages (Early, Full or Normal Retirement Age) controlled for. Each panel also includes the coefficient from a regression of normalized excess mass $b/\hat{R}$ on kink size, which can be interpreted as a difference-in-bunching elasticity, with bootstrapped standard error in parantheses. Appendix figure 1.A2 shows additional graphs by type of statutory age.
Data sources: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

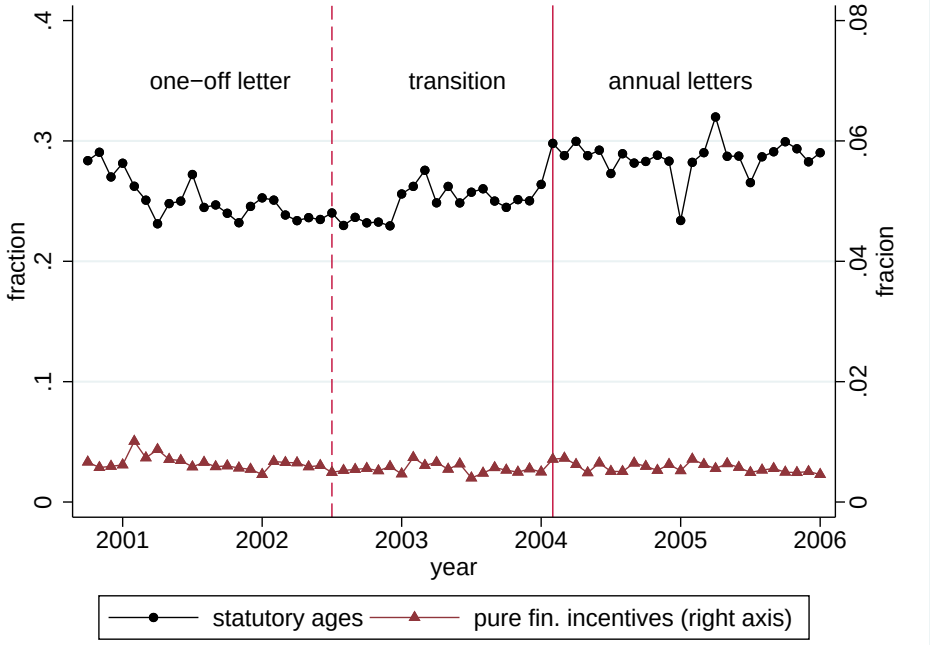
Figure 1.6: Heterogeneity



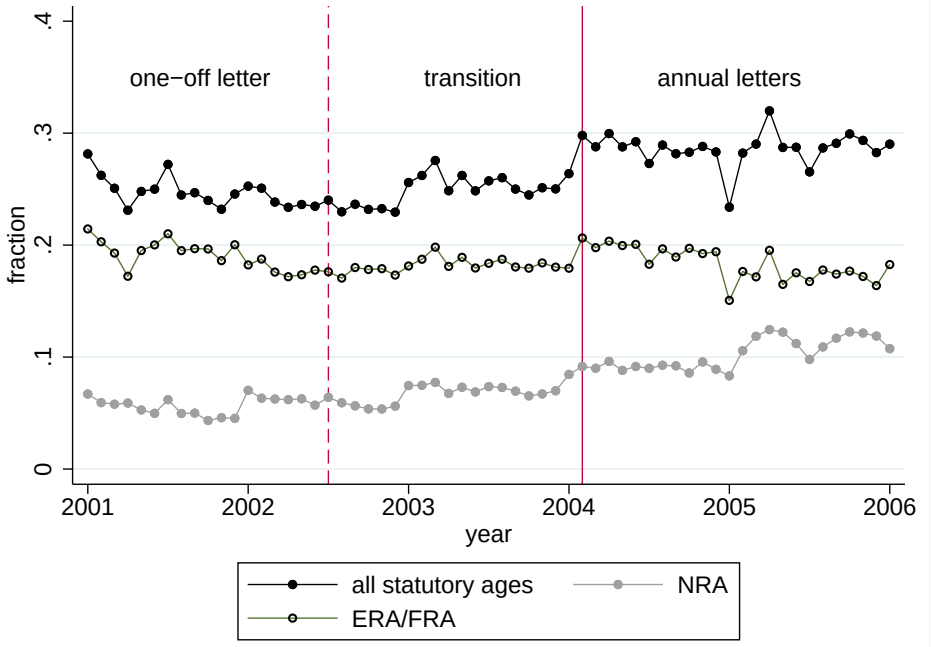
Note: The figure shows average observed bunching elasticities by quintiles of worker and firm-related characteristics, namely estimated lifetime wealth, schooling periods, health status (5=healthiest), a firm size index computed from discrete size categories, unionization rate, tenure, birth cohort and the retirement age at the discontinuity. Black dots indicate bunching at statutory ages, whereas red triangles are for bunching at pure financial incentive discontinuities. The horizontal bars around the point estimates mark confidence intervals based on bootstrapped standard errors. Observed elasticities are only calculated at convex kinks. Corresponding graphs with excess mass are in appendix figure 1.A4. *Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold*

Figure 1.7: The Effect of Information Letters

Panel A: Statutory ages vs. pure financial incentves

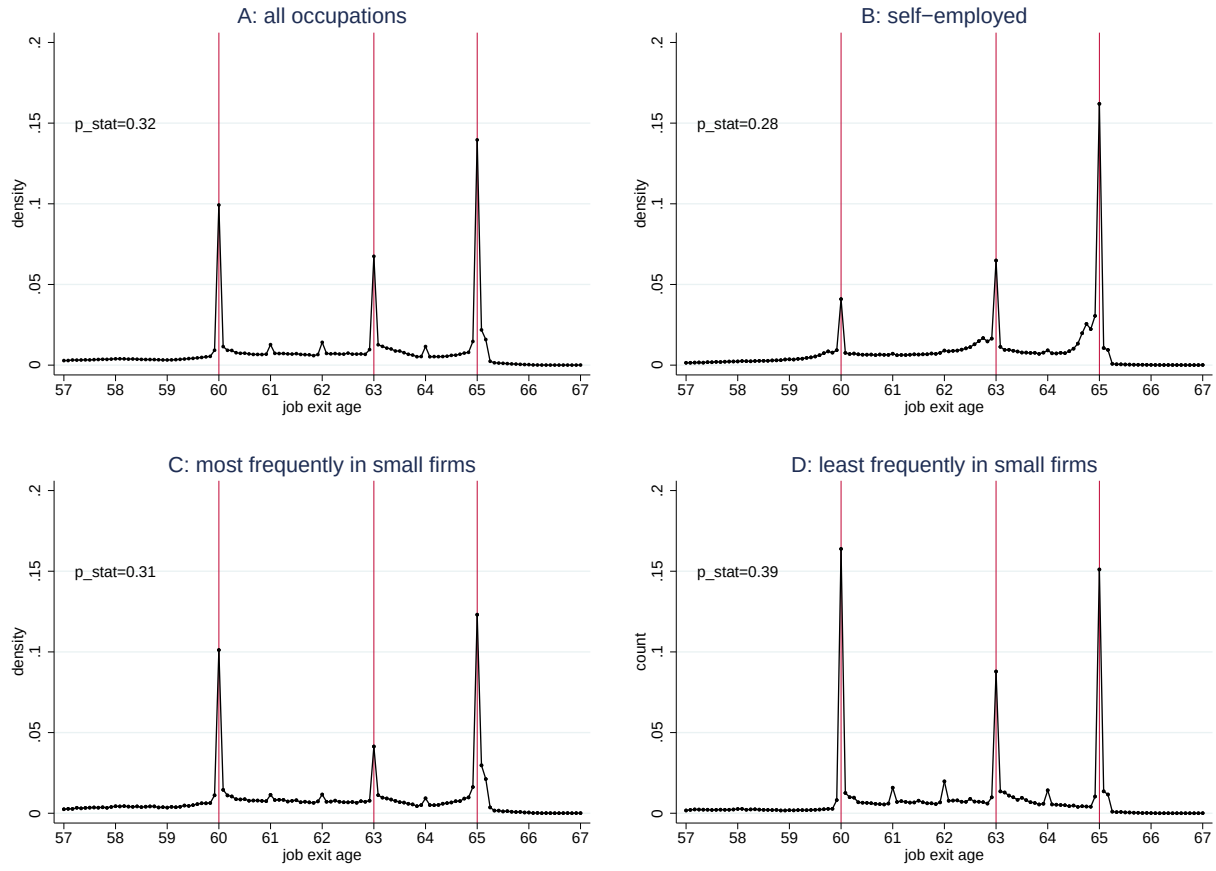


Panel B: Statutory ages by type



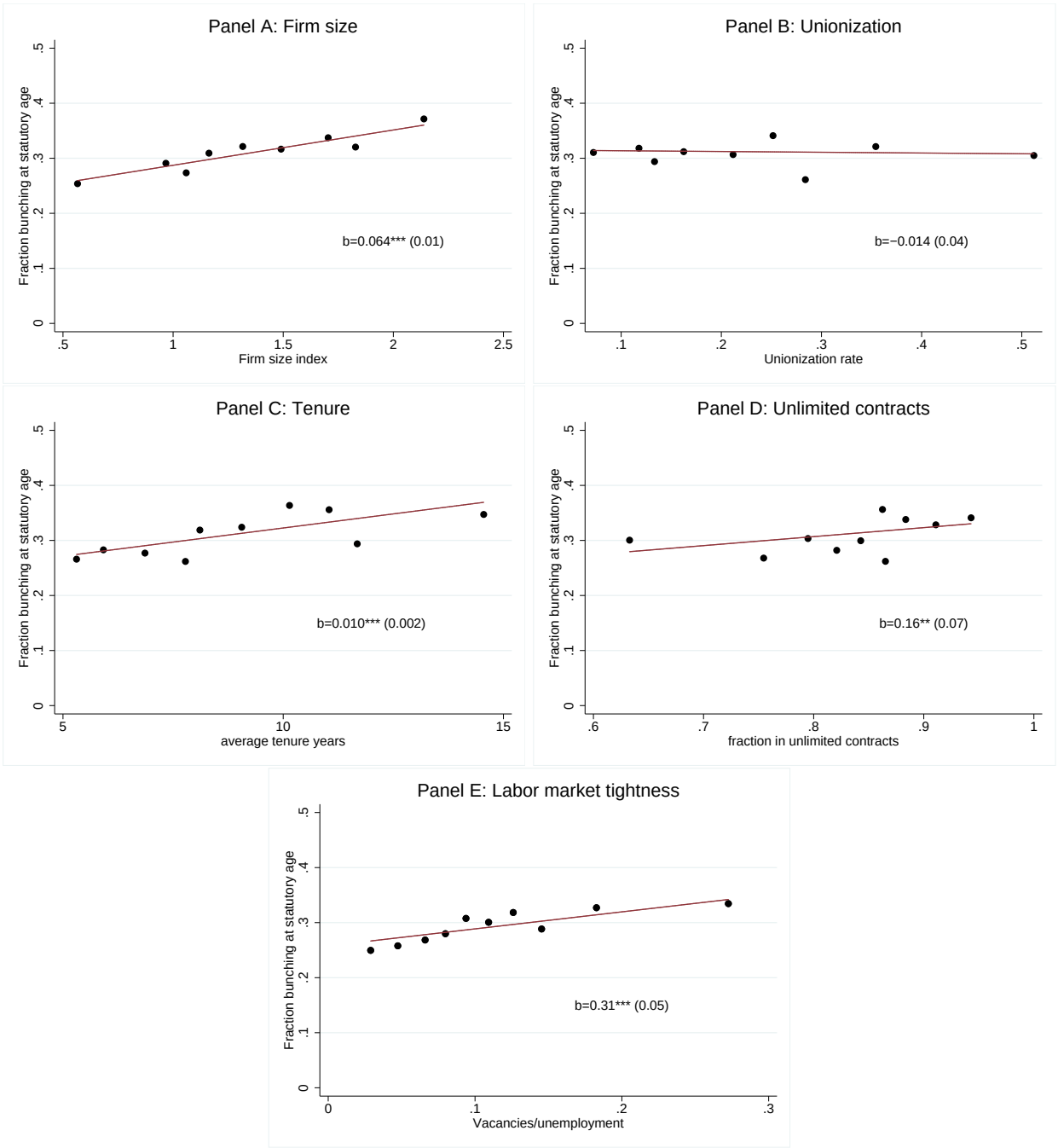
Note: Panel A of the graph shows the fraction of job exits at ages 55 and above occurring at statutory ages and other round ages through a period where the pension fund increased the amount of information provided to workers. Before the reform, a detailed letter was sent to each worker only once. Under the new regime, a basic letter is sent to workers every year, and a detailed letter is sent every 3 years. The dotted vertical line indicates the beginning of the phase-in in June 2002 and the solid vertical line in December 2003 indicates the time when the reform was fully phased in. The black connected dots show the fraction bunching at any statutory age in every month, and the red connected dots show the fraction bunching at pure financial incentive discontinuities (values on the right axis). Series residualized for calendar month effects. Panel B shows the same series for statutory ages by type, where the hollow connected dots show the fraction bunching at the Early/Full Retirement Age (ERA/FRA), and the grey connected dots show the fraction at the Normal Retirement Age (NRA).
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Figure 1.8: Self-Employed and Small Firms



Note: This figure shows the pooled distribution of job exit ages for all workers in the occupation-matched sample (panel A), self-employed workers (panel B), the 20 occupations most frequently in small firms with less than 20 employees (panel C), and the 20 occupations least frequently in small firms (panel D). The connected dots show the count of job exits within monthly bins. Vertical red lines indicate the location of main statutory ages throughout the sample period. p_stat indicates the fraction of workers bunching at statutory ages among the group in each panel.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Figure 1.9: Bunching and Firm Incentives



Note: This figure plots the fraction of workers in an occupation exiting their jobs at a statutory age against a number of variables related to firm incentives. Black dots show average values by decile of the respective explanatory variable. Firm size index is average value out of four size categories, 0=below 20 employees, 1=20 to 200, 2=200 to 2000, 3=above 2000. Unionization rate is fraction with union membership. Average tenure years is workers of all ages in an occupation. Fraction in unlimited contracts is fraction with term limit in their employment contract. Labor market tightness is calculated as the vacancies-unemployment ratio at the state-year level. In panels A to D, the red line is fitted by a univariate occupation-level regression whose slope coefficient are also shown with robust standard errors in parantheses. The occupation-level data is weighted by group size for both bin calculation and regressions. In panel E, the red line is fitted by a univariate individual-level regression whose slope coefficient is shown with standard error clustered at the pathway × month level in parantheses.
Data sources: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold; SOEP v30i

Table 1.1: Pathways into Retirement

Pathway	Required contributions	Other requirements	FRA (after 1990s reforms)	ERA (after 1990s reforms)
Regular	5 years	-	65	65
Long-term insured	35 years	-	65	63
Women	15 years 10 years full	female	65	60
Unemployed/part-time	15 years 8 years full	unemployed or in old-age part-time work before retirement	65	60
Disability	35 years	disability status	63	60
Invalidity	5 years 3 years full	stricter disability status	-	-

Note: This table presents an overview of pathways into retirement, including eligibility requirements and the Full Retirement Age (FRA) and Early Retirement Age (ERA) in each pathway. "1990s reforms" denotes any reforms phased up until cohort 1945. For the unemployed/part-time pathway, unemployment for at least 1 year or old-age part-time work for at least 2 years after age 58 is required. For the disability pathway, an officially recognized disability of a certain degree is required; invalidity entails a stricter disability requirement such that the worker is not able to work more than 3 hours a day in any job. Full contribution years excludes periods where contributions were paid voluntarily.

Table 1.2: Summary Statistics

	(1) individual sample	(2) occupation-matched sample	(3) bunching sample
job exit age	60.87	61.89	60.87
	(2.79)	(2.67)	(1.46)
benefit claiming age	62.03	62.61	62.12
	(2.34)	(2.12)	(1.44)
career length	43.57	44.18	43.60
	(6.54)	(6.94)	(2.35)
contribution points	37.00	38.99	36.76
	(17.21)	(18.08)	(10.67)
lifetime wealth	1,082,887	1,120,252	1,074,722
	(420,416)	(434,983)	(258,461)
female	0.45	0.45	0.48
	(0.50)	(0.50)	(0.43)
east	0.17	0.20	0.18
	(0.38)	(0.40)	(0.09)
married	0.76	0.76	0.76
	(0.42)	(0.43)	(0.06)
sick leave (years)	0.075	0.056	0.07
	(0.26)	(0.21)	(0.04)
schooling (years)	10.60	10.74	10.64
	(1.58)	(1.79)	(0.28)
small firm		0.27	
		(0.18)	
large firm		0.44	
		(0.18)	
tenure		8.95	
		(2.80)	
unlimited contract		0.83	
		(0.09)	
Obs. (individuals)	8,880,619	3,955,574	
Obs. (discontinuities)			644

Note: This table summarizes selected variables in the samples used. The individual and occupation-matched samples are at the worker level, while the bunching sample collects bunching observations at the discontinuity level. Job exit and benefit claiming ages are in years. "Career length" is time between first and last contribution. "Lifetime wealth" is in net present value terms as in equation (1.1). "East" is a dummy for residence in East Germany. "Small firm" and "large firm" are indicators for firms with less than 20 employees and more than 200 employees, respectively. Firm size, tenure and fraction in unlimited contract are at the occupation level. Standard deviations in parantheses. See appendix 1.A.3.1 for further details of variable definitions. *Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold*

Table 1.3: Summarizing Discontinuities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Statutory ages			Pure financial incentives		
	all	Early	Full	Normal	all	kinks	notches
Mean kink size $\frac{\Delta\tau}{1-\tau}$	0.08	0.32	0.41	-0.35	0.80	0.47	0.94
s.d. across groups	0.42	0.28	0.34	0.15	0.61	0.21	0.67
s.d. within group	0.06	0.05	0.08	0.05	0.28	0.13	0.34
No. discontinuities	386	117	257	93	258	78	180

Note: This table shows summary statistics of discontinuities in the bunching sample by type of discontinuity. "Kink size" is the percentage reduction in the net-of-tax rate at the discontinuity. "s.d. across groups" is standard deviations across discontinuities of a given type. "s.d. within group" is standard deviation within a group of workers facing the same discontinuity. Note that the number of discontinuities in columns (2) to (4) are larger than the total in column (1) because some kinks are linked to more than one type of statutory age. All statistics weighted by group size.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 1.4: Bunching across all Discontinuities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Statutory ages			Pure financial incentives		
	all	Early	Full	Normal	all	kinks	notches
Excess mass b	21.8	13.8	20.6	31.8	2.99	0.09	4.31
	(0.88)	(0.96)	(0.87)	(1.99)	(0.27)	(0.04)	(0.34)
Observed	1.64	1.91	1.39	4.12	0.15	0.009	0.22
elasticity $\hat{\varepsilon}$	(0.07)	(0.12)	(0.08)	(0.55)	(0.01)	(0.003)	(0.02)

Note: This table summarizes bunching responses by type of discontinuity in the bunching sample. Excess mass and observed elasticities are computed as described in appendix 1.A.4. All statistics are weighted by group sizes. Standard errors in parantheses. Observed elasticities are only calculated only at convex kinks, that is excluding non-convex NRA kinks.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 1.5: Reduced-Form Estimation

	(1)	(2)	(3)	(4)	(5)
	Dependent variable: Excess mass $b/\hat{R}$				
kink size $\frac{\Delta\tau}{1-\tau}$	0.11***	0.092***	0.087	0.040	0.074
	(0.027)	(0.027)	(0.10)	(0.15)	(0.19)
Statutory age at kink:					
Early Retirement Age	0.23***	0.15***	0.17**	0.18	0.18
	(0.029)	(0.024)	(0.085)	(0.11)	(0.14)
Full Retirement Age	0.23***	0.27***	0.34***	0.35***	0.35***
	(0.047)	(0.036)	(0.071)	(0.099)	(0.11)
Normal Retirement Age	0.77***	0.85***	0.83***	0.80***	0.82***
	(0.077)	(0.082)	(0.16)	(0.25)	(0.32)
Observations (discontinuities)	644	644	644	644	583
R-squared	0.66	0.70	0.85	0.88	0.86
Stat. age interactions	no	yes	yes	yes	yes
Worker controls	no	no	yes	yes	yes
Pathway FE, year-of-birth FE	no	no	yes	yes	yes
Pathway × year-of-birth FE	no	no	no	yes	yes
Occupation-level controls	no	no	no	no	yes

Note: This table shows results from group-level regressions of excess mass normalized by the retirement age b/R on kink size as well as dummies for the presence of statutory age types $s \in (ERA, FRA, NRA)$ based on equation (1.5), using the bunching sample. Statutory age interactions are interactions between dummies for each statutory age type. Worker controls include dummies for female, married and East Germany, last income before retirement, lifetime wealth, career length, sick leave years and education years. Occupation-level controls include firm size index, unionization rate, active union member rate, tenure in the firm, fraction in unlimited contracts, fraction receiving severance pay, fraction of involuntary job exits. Regressions weighted by group size. Bootstrapped standard errors in parantheses. *** p<0.01, ** p<0.05, * p<0.1.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 1.6: Oaxaca-Blinder Bunching Decomposition

reference category:	(1) statutory ages	(2) pure financial incentives	(3) average
Excess mass difference	-18.9	-18.9	-18.9
Explained by			
financial incentives	2.84	-4.38	-0.77
%	-15.0%	23.1%	4.07%
worker variables	-2.26	0.41	-0.93
%	11.9%	-2.16%	4.88%
firm variables	0.08	-2.92	-1.42
%	-0.42%	15.4%	7.52%
Unexplained	-19.6	-12.0	-15.8
%	103.5%	63.6%	83.5%
Obs. (discontinuities)	629	629	629

Note: This table shows results from a Oaxaca-Blinder decomposition, where differences in excess mass between statutory ages and pure financial incentive discontinuities are attributed to differences in explanatory variables and an unexplained component. The bunching sample is used. "Financial incentives" includes only kink size. "Worker variables" includes dummies for female, married and East Germany, last income before retirement, lifetime wealth, career length, sick leave years and education years. "Firm variables" includes the following occupation-level variables: firm size index, unionization rate, active union member rate, tenure in the firm, fraction in unlimited contracts, fraction receiving severance pay, fraction of involuntary job exits. Columns differ according to which group is chosen as a reference group, i.e. the coefficients of which group are used to calculate explained shares. In column (1), the reference group are statutory ages, in column (2) the reference group are pure financial incentive discontinuities, and in column (3) coefficients based on the two groups averaged.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 1.7: Reduced-Form Estimation: Heterogeneous Coefficients

	(1)	(2)	(3)	(4)
	Dependent variable: Excess mass $b/\hat{R}$			
kink size $\frac{\Delta\tau}{1-\tau}$	0.092***	0.29***	0.16***	0.26***
	(0.026)	(0.056)	(0.039)	(0.035)
Statutory age at kink:				
Early Retirement Age	0.15***	0.21***	0.26***	0.30***
	(0.027)	(0.015)	(0.036)	(0.036)
Full Retirement Age	0.27***	0.31***	0.34***	0.42***
	(0.038)	(0.037)	(0.039)	(0.056)
Normal Retirement Age	0.85***	1.05***	1.00***	1.00***
	(0.078)	(0.063)	(0.097)	(0.082)
Observations (discontinuities)	644	627	627	627
R-squared	0.69	0.87	0.81	0.96
Stat. age interactions	yes	yes	yes	yes
Heterogeneous coefficients:				
by pathway	no	yes	no	yes
by year of birth	no	no	yes	yes
by pathway × year of birth	no	no	no	yes

Note: This table shows results from group-level regressions of excess mass normalized by the retirement age b/R on kink size as well as dummies for the presence of statutory age types $s \in (ERA, FRA, NRA)$ based on equation (1.6), using the bunching sample. Column (1) reports coefficients from a regression according to equation (1.5) without controls. Columns (2) to (4) report weighted averages of heterogeneous coefficients estimated according to equation (1.6), where column (2) defines groups by pathway, (3) defines groups by year of birth, and (4) by pathway × year of birth. Groups with no variation in D^s are excluded from the within-group estimation in columns (2) to (4) since group-specific coefficients cannot be estimated in this case. Regressions weighted by group size. Bootstrapped standard errors in parantheses. *** p<0.01, ** p<0.05, * p<0.1.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 1.8: Worker Characteristics

	(1)	(2)	(3)	(4)	(5)
	Dependent variable: Dummy for job exit at...				
	any statutory age	ERA/FRA	NRA	other round age	pure financial incentive
schooling	0.007***	-0.009***	0.015***	-0.002***	0.0001***
	(0.0005)	(0.0005)	(0.0006)	(0.0002)	(0.0000)
economic training	0.015***	0.053***	-0.038***	0.023***	0.001***
	(0.002)	(0.003)	(0.002)	(0.002)	(0.000)
female	-0.020***	0.11***	-0.13***	0.026***	-0.004***
	(0.005)	(0.004)	(0.005)	(0.003)	(0.0005)
married	-0.17***	-0.017***	-0.15***	-0.0000	0.005***
	(0.004)	(0.003)	(0.001)	(0.002)	(0.0003)
female × married	-0.009**	0.035***	-0.044***	0.004***	0.0002
	(0.002)	(0.002)	(0.003)	(0.001)	(0.0002)
life earnings	0.26***	0.095***	0.17***	0.011***	-0.017***
	(0.006)	(0.006)	(0.007)	(0.003)	(0.0006)
last earnings	0.089***	0.021***	0.068***	0.021***	0.002***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.0001)
pension wealth/annual earnings	0.039***	-0.002*	0.041***	-0.001**	-0.0003***
	(0.006)	(0.001)	(0.001)	(0.000)	(0.0001)
Mean dep. var.	0.32	0.18	0.14	0.10	0.005
Observations	3,557,890	3,557,890	3,557,890	3,557,890	3,557,890
R-squared	0.15	0.08	0.20	0.10	0.02
Add. worker controls	yes	yes	yes	yes	yes
Occ.-level controls	yes	yes	yes	yes	yes
Year of birth FE	yes	yes	yes	yes	yes
Pathway FE	yes	yes	yes	yes	yes

Note: The table shows results from an individual-level regression of dummies job exits at some ages of interest on worker characteristics. Additional worker controls include dummy for East Germany, career length, sick leave years. Occupation-level controls include firm size index, unionization rate, active union member rate, tenure in the firm, fraction in unlimited contracts, fraction receiving severance pay, fraction of involuntary job exits. Standard errors clustered at the pathway × month of birth level.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 1.9: Firm Incentives

	(1)	(2)	(3)	(4)	(5)
	Dependent variable: Dummy for job exit at...				
	any statutory age	ERA/FRA	NRA	pure financial incentive	other round age
firm size index	0.032***	0.023***	0.009***	0.018***	-0.0003**
	(0.002)	(0.002)	(0.001)	(0.001)	(0.0002)
union	-0.055***	0.019***	-0.073***	-0.004	0.002***
	(0.005)	(0.005)	(0.004)	(0.003)	(0.0007)
tenure	-0.0007***	-0.0009***	0.0002	-0.0009***	0.0000
	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0000)
unlimited	-0.032***	0.002	-0.034***	-0.011***	0.001***
contracts	(0.005)	(0.004)	0.003	(0.004)	(0.0005)
labor market	0.41***	-0.093*	0.50***	-0.062***	0.036***
tightness	(0.05)	(0.05)	(0.057)	(0.023)	(0.005)
Mean dep. var.	0.32	0.18	0.14	0.07	0.005
Observations	3,537,802	3,537,802	3,537,802	3,537,802	3,537,802
R-squared	0.14	0.08	0.17	0.09	0.02
Worker controls	yes	yes	yes	yes	yes
Year of birth FE	yes	yes	yes	yes	yes
Pathway FE	yes	yes	yes	yes	yes

Note: The table shows results from an individual-level regression of dummies job exits at some ages of interest on variables related to firm incentives, using the occupation-matched sample. Firm size index, unionization, tenure and fraction in unlimited contracts are at occupation level. Labor market tightness is at state-year level. Worker controls include dummies for female, married and East Germany, last income before retirement, lifetime wealth, career length, sick leave years and education years. Standard errors clustered at the pathway × month of birth level.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

1.A Appendix

1.A.1 Appendix Figures and Tables

Figure 1.A1: Framing

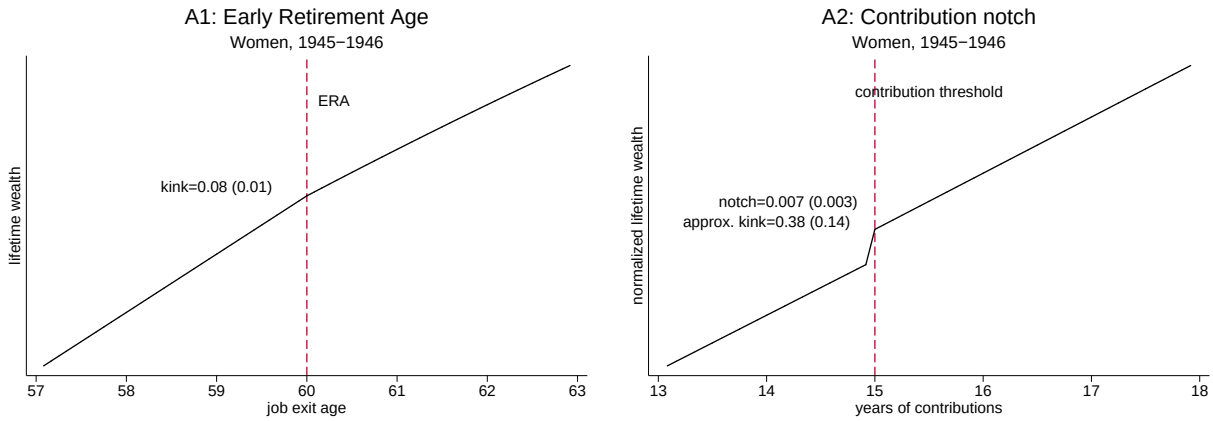
Note: This figure shows excerpts of an information leaflet that informs workers about a pension reform. Explanation/translation of the main points is provided in the red boxes on the right. Source: `http://www.deutsche-rentenversicherung.de/cae/servlet/contentblob/232636/publicationFile/49694/rente_mit_67.pdf`

Figure 1.A0: Framing (continued)

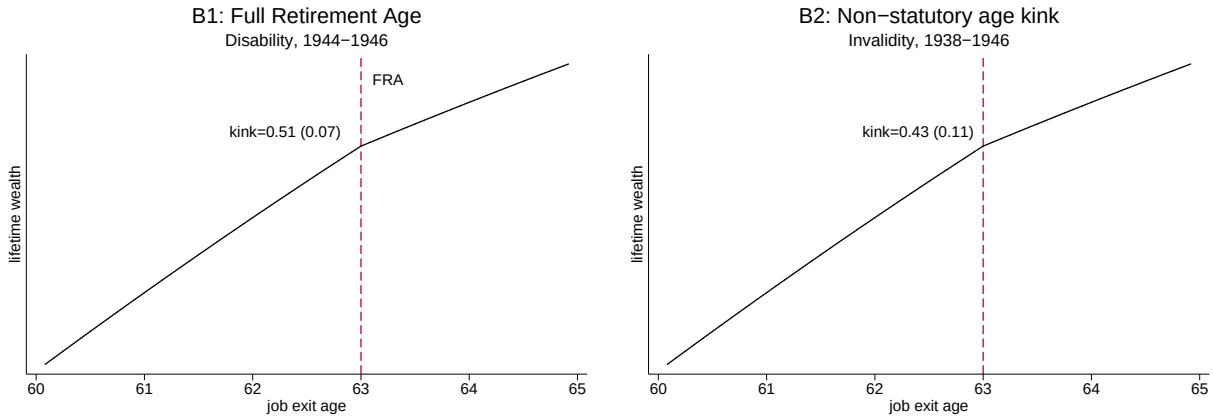
Note: This figure shows excerpts of an information leaflet that informs workers about a pension reform. Explanation/translation of the main points is provided in the red boxes on the right. The bottom right panel is taken from an information leaflet on invalidity pensions. Sources: `http://www.deutsche-rentenversicherung.de/cae/servlet/contentblob/232636/publicationFile/49694/rente_mit_67.pdfhttp://www.deutsche-rentenversicherung.de/cae/servlet/contentblob/232616/publicationFile/49858/erwerbsminderungsrente_das_netz_fuer_alle_faelle.pdf`

Figure 1.A1: Budget Constraint Discontinuities

Panel A: Statutory age vs. contribution notch

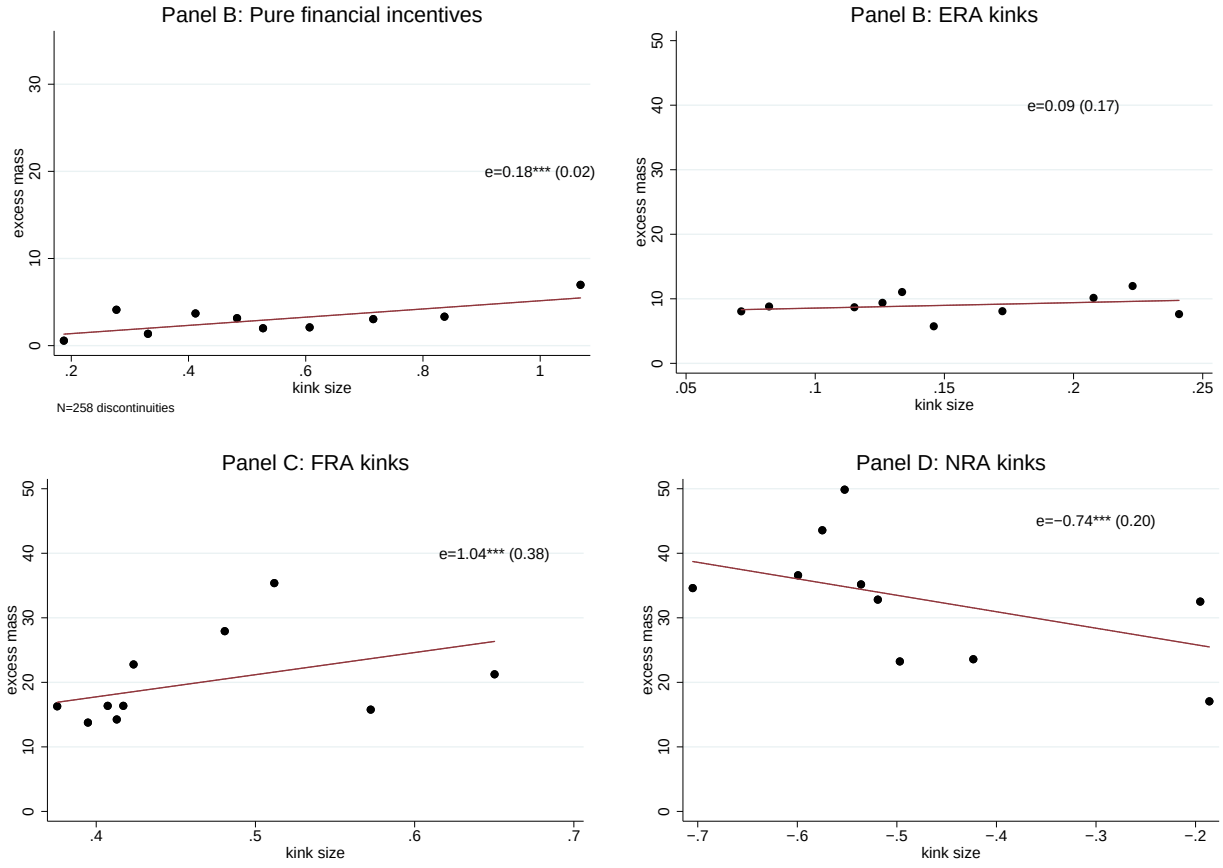


Panel B: Statutory age vs. pure financial incentive kink



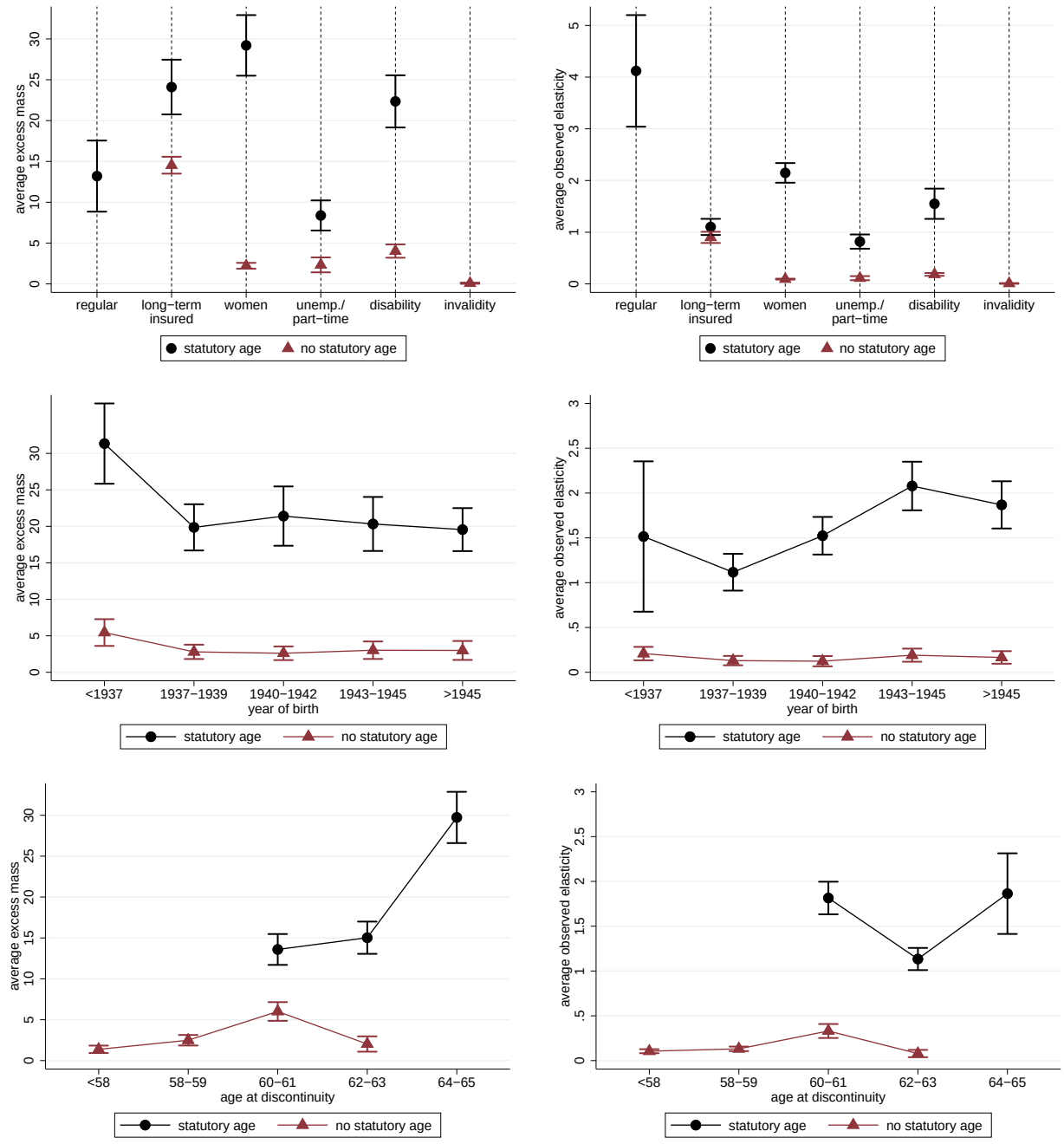
Note: This figure shows the lifetime budget constraint discontinuities around which bunching is estimated in figure 1.4. Panels A1 and B1 show kinks linked to the FRA in the disability pathway and the ERA in the women's pathway, respectively. Panel A2 shows a notch arising from the 15-year contribution requirement in the women's pathway, and panel B2 shows a kink due to financial adjustment of pensions in the invalidity pathway. Note that the size of the discontinuity varies across workers, and graphs show actual slopes for an average-income worker. "kink" denotes average kink size $\Delta\tau/(1-\tau)$ defined the proportional change in the implicit marginal net-of-tax rate at the discontinuity, with standard deviations in parantheses. "notch" denotes average notch size defined as the implicit average net-of-tax rate just below the kink relative to above. In panel A2, "approximate kink size" is calculated by approximating the notch as a kink for the marginal buncher following Kleven and Waseem (2013).

Figure 1.A2: Bunching and Financial Incentives



Note: The figure shows binned scatterplots of the excess mass at pure financial incentive discontinuities (panel A), ERAs (panel B), FRAs (panel C) and NRAs (panel D)against kink size. Each panel also includes the coefficient from a regression of normalized excess mass $b/\hat{R}$ on kink size, which can be interpreted as a difference-in-bunching elasticity, with bootstrapped standard error in parantheses.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

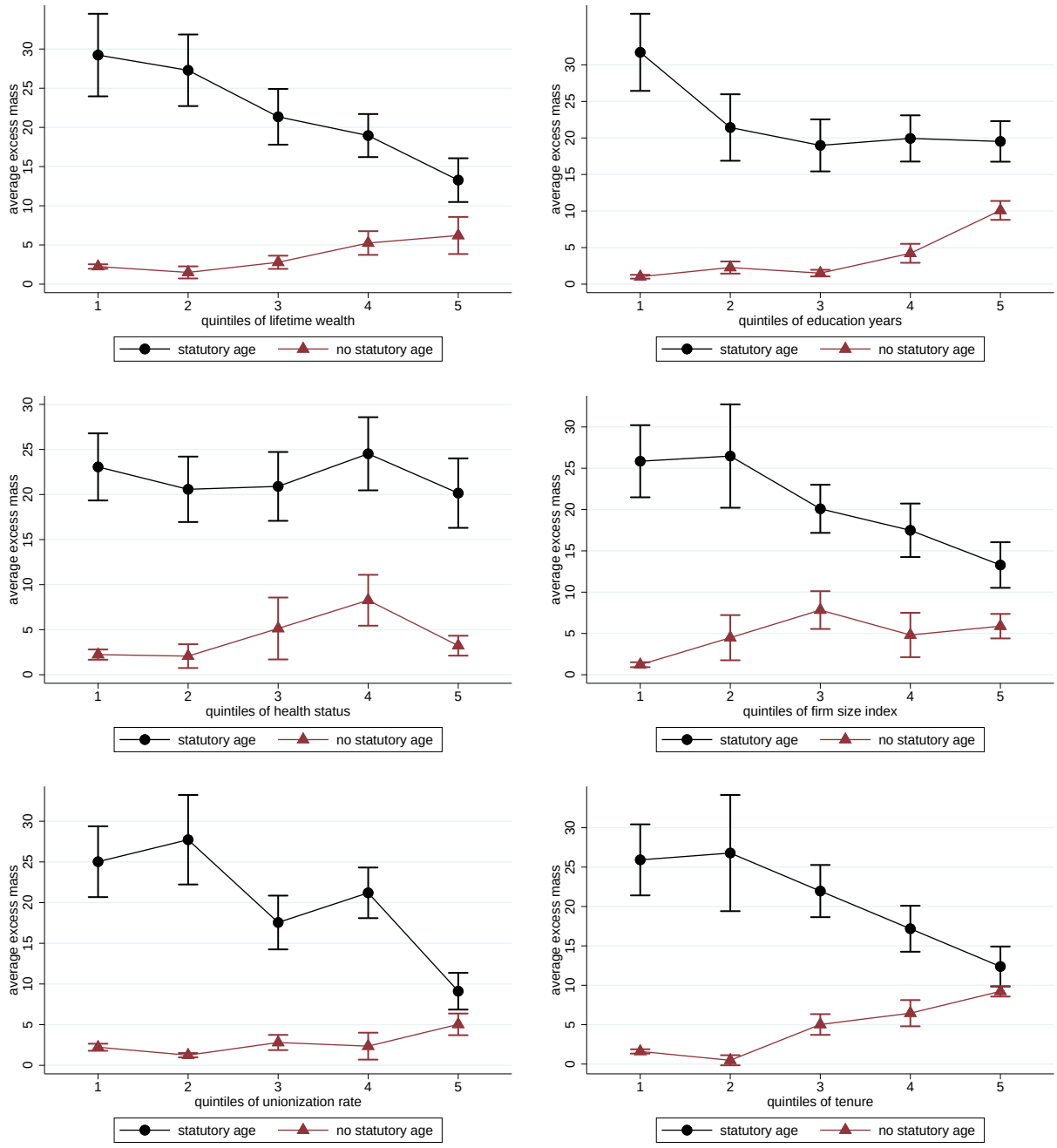
Figure 1.A3: Heterogeneity: Groups and Discontinuities



Note: The figure shows average excess mass (left panels) and average observed bunching elasticities (right panels) by retirement pathway, year of birth and the retirement age at the discontinuity. Black dots indicate bunching at statutory ages, whereas red triangles are for bunching at pure financial incentive discontinuities. The horizontal bars around the point estimates mark confidence intervals based on bootstrapped standard errors. Observed elasticities in the right panels are only calculated at convex kinks.

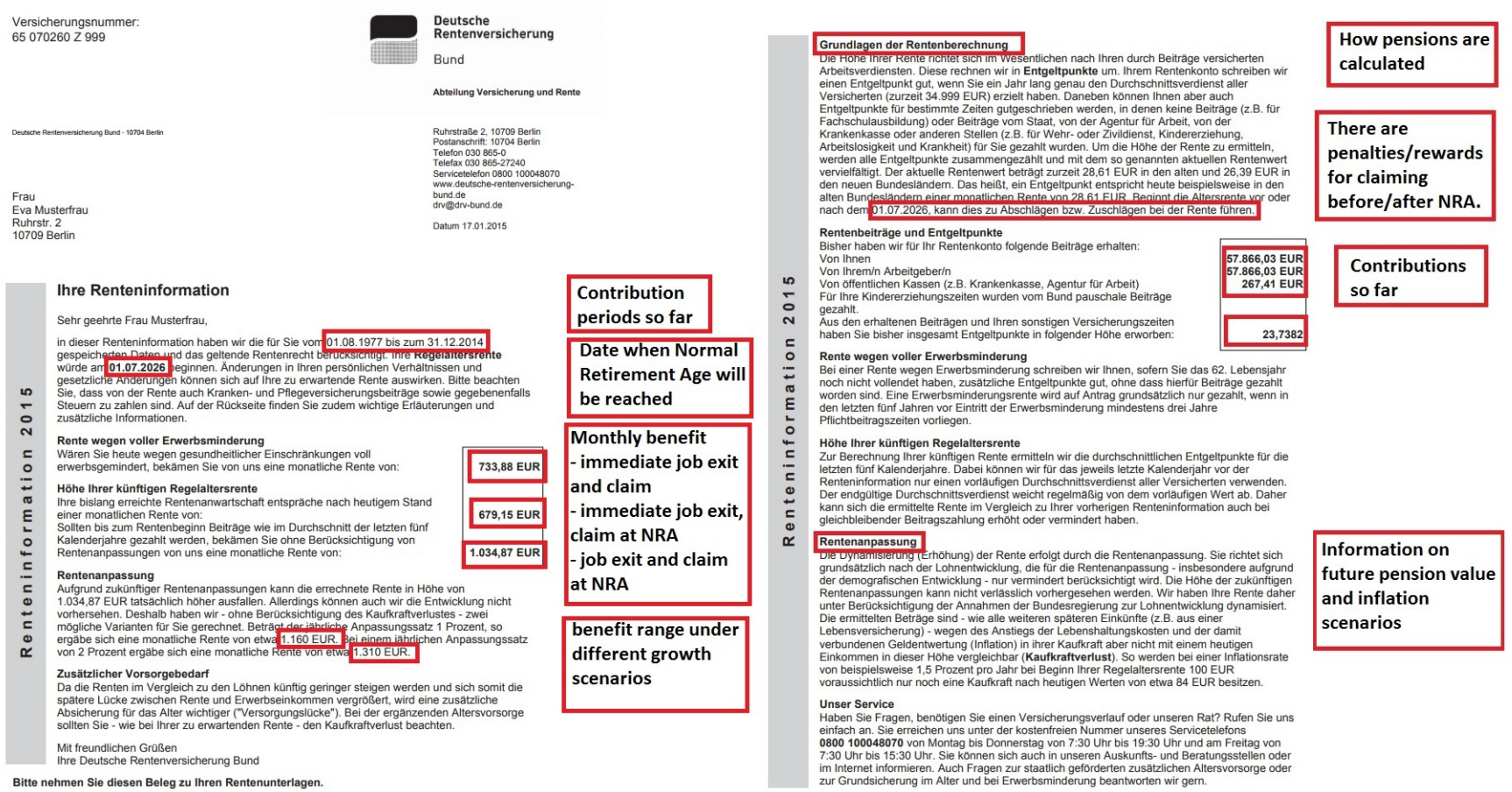
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Figure 1.A4: Heterogeneity: Worker and Firm Characteristics



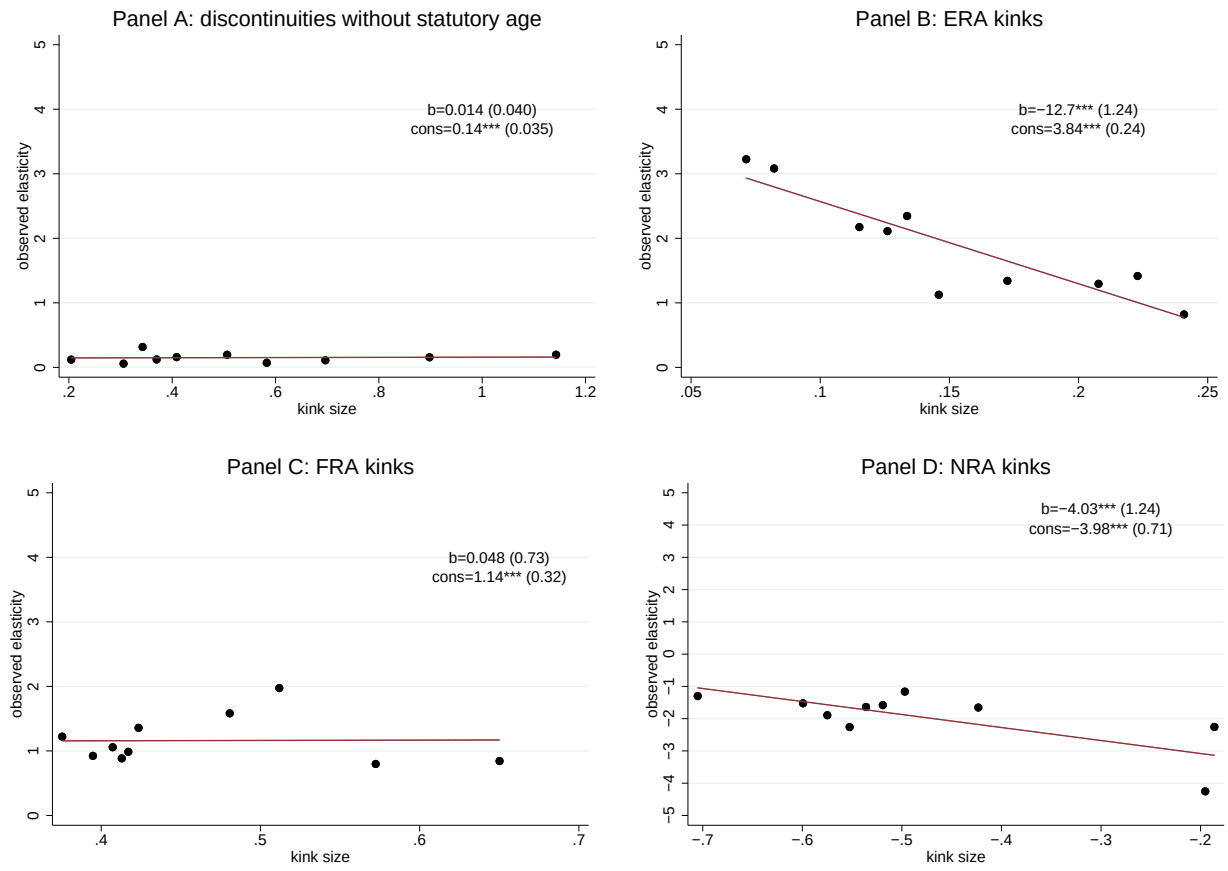
Note: The figure shows average observed bunching elasticities by quintiles of estimated lifetime wealth, schooling periods, sick leave periods, a firm size index computed from discrete size categories, unionization rate and tenure. The latter three variables are at the occupation-level. Black dots indicate bunching at statutory ages, whereas red triangles are for bunching at pure financial incentive discontinuities. The horizontal bars around the point estimates mark confidence intervals based on bootstrapped standard errors. Observed elasticities in panel B are only calculated at convex kinks. The corresponding excess mass is in figure 1.6.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Figure 1.A5: Information Letters



Note: This figure shows an information letter (*Renteninformation*) sent annually to each worker from 2004. Red boxes highlight the main information provided by the letter, with explanation/translation on the right.

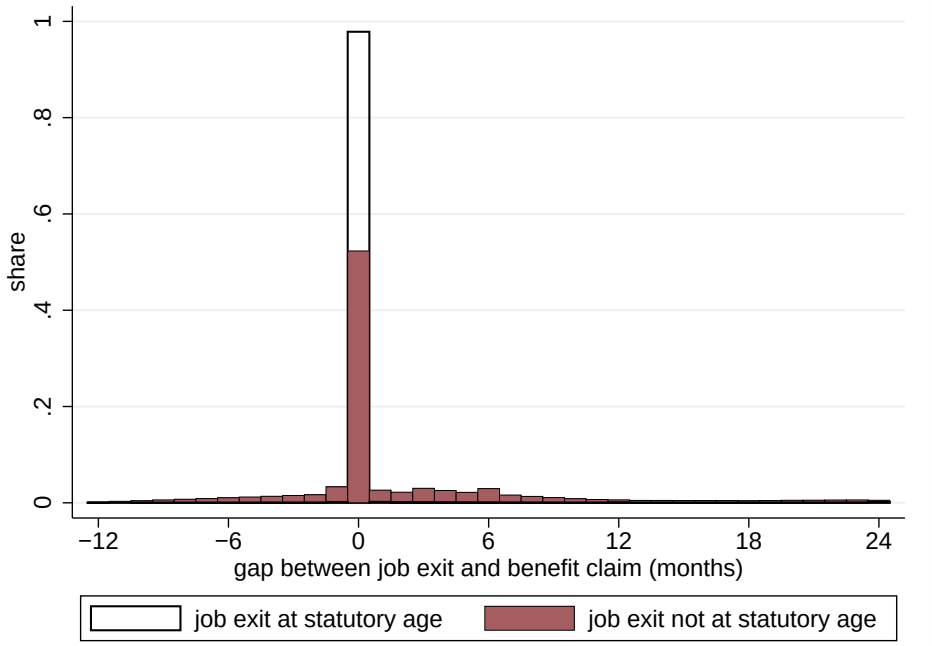
Figure 1.A6: Larger responses at larger kinks?



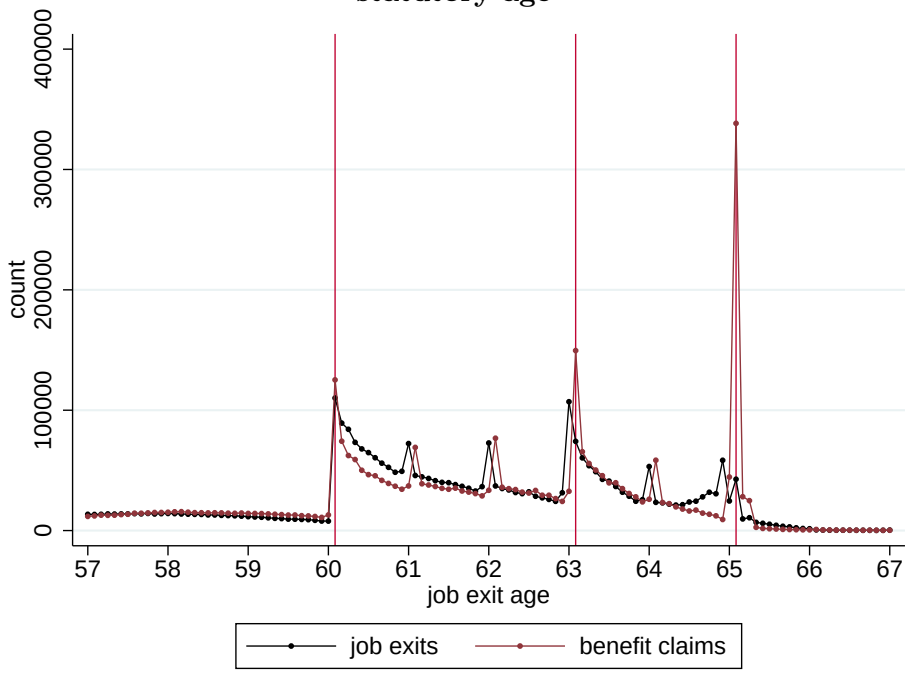
Note: The figure shows binned scatterplots of the observed elasticity at pure financial incentive discontinuities (panel A), ERAs (panel B), FRAs (panel C) and NRAs (panel D)against kink size. Each panel also includes the coefficient from a regression of observed elasticities on kink size, with bootstrapped standard error in parantheses.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Figure 1.A7: Benefit Claiming Patterns

Panel A: Claiming gaps of workers eligible to claim immediately



Panel B: Claiming ages of workers eligible to claim immediately who do not exit at statutory age



Note: Panel A of the figure shows a histogram of the gap in months between workers' job exit and benefit claim among those exiting their job at a statutory age (white bars) and those exiting at other ages (red bars). Only workers who are already eligible to claim are included in both groups. Panel B plots the distribution of job exit ages (black connected dots) and benefit claiming ages (red connected dots) among workers who are eligible to claim immediately.
Data sources: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold; SOEP v30i

Table 1.A1: Reduced-Form Estimation: Heterogeneous Coefficients

Panel A: by pathway

	(1) Long-term Insured	(2) Women	(3) Unemp./ part-time	(4) Disability	(5) Invalidity
kink size $\frac{\Delta\tau}{1-\tau}$	0.62***	0.13***	0.44*	0.20***	0.0064**
	(0.085)	(0.044)	(0.24)	(0.011)	(0.0030)
Statutory age at kink:					
Early Retirement Age	0.12***	0.24***	0.094**	0.38***	
	(0.028)	(0.019)	(0.042)	(0.016)	
Full Retirement Age	0.14*	0.57***	0.12	0.13***	
	(0.078)	(0.052)	(0.11)	(0.018)	
Normal Retirement Age	1.11***	1.16***	0.53***	1.34***	
	(0.070)	(0.12)	(0.17)	(0.050)	
Discontinuities	98	127	159	165	78
Stat. age interactions	yes	yes	yes	yes	yes

Panel B: by year of birth (selected)

	(1) 1933	(2) 1936	(3) 1939	(4) 1942	(5) 1945	(6) 1948
kink size $\frac{\Delta\tau}{1-\tau}$	0.082	0.078	0.086**	0.23***	0.25***	0.18*
	(0.61)	(0.34)	(0.036)	(0.080)	(0.059)	(0.10)
Statutory age at kink:						
Early Retirement Age	0.56	0.67**	0.15	0.25***	0.25***	0.30***
	(0.44)	(0.33)	(0.17)	(0.064)	(0.057)	(0.070)
Full Retirement Age			0.17***	0.64***	0.14	0.31***
			(0.037)	(0.11)	(0.40)	(0.073)
Normal Retirement Age	1.02**	0.87***	0.66**	2.12***	1.22***	0.80***
	(0.49)	(0.22)	(0.31)	(0.84)	(0.47)	(0.22)
Discontinuities	15	15	46	58	23	37
Stat. age interactions	yes	yes	yes	yes	yes	yes

Note: This table shows heterogeneous coefficients whose weighted averages are presented in table 1.7. Panel A presents heterogeneous coefficients by pathway, where the regular pathway is excluded because there is no variation in the presence of statutory ages. Panel B shows heterogeneous coefficients cohort for selected years of birth. Regressions weighted by group size. Bootstrapped standard errors in parantheses. *** p<0.01, ** p<0.05, * p<0.1.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 1.A2: Reduced-Form Estimation: Effects of Information Letters

	(1)	(2)	(3)	(4)
	Dependent variable: Excess mass $b/\hat{R}$			
kink size $\frac{\Delta\tau}{1-\tau}$	0.085**	0.029	0.094***	0.040
	(0.041)	(0.099)	(0.016)	(0.097)
Statutory age at kink:				
Early Retirement Age	0.070*	0.13	0.15***	0.18*
	(0.039)	(0.090)	(0.032)	(0.098)
Full Retirement Age	0.14***	0.30***	0.18***	0.33***
	(0.048)	(0.085)	(0.058)	(0.083)
Normal Retirement Age	0.78***	0.68***	0.82***	0.70***
	(0.074)	(0.16)	(0.11)	(0.15)
Interactions:				
annual letters × kink size	0.085	-0.054	-0.003	0.042
	(0.12)	(0.091)	(0.054)	(0.12)
annual letters × statutory age	0.20***	0.11*		
	(0.054)	(0.061)		
annual letters × Early Retirement Age			0.003	0.021
			(0.054)	(0.062)
annual letters × Full Retirement Age			0.14*	0.038
			(0.08)	(0.069)
annual letters × Normal Retirement Age			0.044	0.17
			(0.18)	(0.15)
Observations (Discontinuities)	644	644	644	644
R-squared	0.71	0.86	0.70	0.86
Statutory age interactions	yes	yes	yes	yes
Worker controls	no	yes	no	yes
Year of birth FE	no	yes	no	yes
Pathway FE	no	yes	no	yes

Note: This table shows results from group-level regressions analogous to table 1.5, allowing for additional interactions of the presence of statutory ages with an indicator for annual information letters in 2004 and later. Excess mass normalized by the retirement age b/R is regressed on kink size as well as dummies for the presence of statutory age types $s \in (ERA, FRA, NRA)$ based on equation (1.5). Regressions weighted by group size. Bootstrapped standard errors in parantheses. *** p<0.01, ** p<0.05, * p<0.1.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 1.A3: Reduced-Form Estimation: Larger responses at larger kinks?

	(1)	(2)	(3)
	Dependent variable: Observed elasticity $\hat{\varepsilon}$		
kink size $\frac{\Delta\tau}{1-\tau}$	-0.012	0.001	-0.76***
	(0.058)	(0.13)	(0.27)
Statutory age at kink:			
Early Retirement Age	1.60***	1.94***	0.92***
	(0.16)	(0.067)	(0.32)
Full Retirement Age	0.59***	1.06***	0.17
	(0.091)	(0.093)	(0.25)
Normal Retirement Age	1.82***	1.12***	6.03
	(0.48)	(0.24)	(10.2)
Observations (Discontinuities)	568	568	568
R-squared	0.69	0.75	0.86
Statutory age interactions	no	yes	yes
Worker controls	no	no	yes
Year of birth FE	no	no	yes
Pathway FE	no	no	yes

Note: This table shows results from group-level regressions analogous to table 1.5, allowing for additional interactions of the presence of statutory ages with an indicator for reform periods. Excess mass normalized by the retirement age b/R is regressed on kink size as well as dummies for the presence of statutory age types $s \in (ERA, FRA, NRA)$ based on equation (1.5). Regressions weighted by group size. Bootstrapped standard errors in parantheses. *** p<0.01, ** p<0.05, * p<0.1.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 1.A4: Reduced-Form Estimation Excluding NRA Discontinuities

	(1)	(2)	(3)	(4)
	Dependent variable: Excess mass $b/\hat{R}$			
kink size $\frac{\Delta\tau}{1-\tau}$	0.14***	0.26***	0.18***	0.26***
	(0.012)	(0.027)	(0.039)	(0.037)
Statutory age at kink:				
Early Retirement Age	0.23***	0.21***	0.26***	0.30***
	(0.015)	(0.012)	(0.039)	(0.036)
Full Retirement Age	0.37***	0.32***	0.33***	0.34***
	(0.024)	(0.028)	(0.027)	(0.056)
Observations (discontinuities)	551	551	551	551
R-squared	0.74	0.87	0.77	0.88
Stat. age interactions	yes	yes	yes	yes
Heterogeneous coefficients:				
by pathway	no	yes	no	yes
by year of birth	no	no	yes	yes
by pathway × year of birth	no	no	no	yes

Note: This table shows results from group-level regressions analogous to table 1.7, excluding all discontinuities linked to a NRA. Column (1) reports coefficients from a regression according to equation (1.5) without controls. Columns (2) to (4) report weighted averages of heterogeneous coefficients estimated according to equation (1.6), where column (2) defines groups by pathway, (3) defines groups by year of birth, and (4) by pathway × year of birth. Groups with no variation in D^s are excluded from the within-group estimation in columns (2) to (4) since group-specific coefficients cannot be estimated in this case. Regressions weighted by group size. Bootstrapped standard errors in parantheses. *** p<0.01, ** p<0.05, * p<0.1.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 1.A5: Reduced-Form Estimation with Salience Effects

	(1)	(2)	(3)
	Dependent variable: Excess mass $b/\hat{R}$		
kink size $\frac{\Delta\tau}{1-\tau}$	0.16**	0.16***	0.17***
	(0.066)	(0.062)	(0.022)
Statutory age at kink:			
Early Retirement Age	0.36***	0.35***	0.22***
	(0.052)	(0.050)	(0.034)
Full Retirement Age	0.57***	0.80***	0.35**
	(0.12)	(0.12)	(0.14)
Normal Retirement Age	0.30*	0.14	-0.0018
	(0.16)	(0.15)	(0.22)
Interactions:			
kink size × any statutory age	-0.89***	-1.34***	
	(0.29)	(0.27)	
kink size × Early Retirement Age			0.052
			(0.19)
kink size × Full Retirement Age			0.039
			(0.32)
kink size × Normal Retirement Age			-1.72***
			(0.38)
Observations (discontinuities)	644	644	644
R-squared	0.66	0.70	0.75
Stat. age interactions	no	yes	yes

Note: This table shows results from group-level regressions analogous to table 1.5, allowing for additional interactions of kink size with the presence of statutory ages. Excess mass normalized by the retirement age b/R is regressed on kink size as well as dummies for the presence of statutory age types $s \in (ERA, FRA, NRA)$ based on equation (1.5). Statutory age interactions are interactions between dummies for each statutory age type. Regressions weighted by group size. Bootstrapped standard errors in parantheses. *** p<0.01, ** p<0.05, * p<0.1.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

1.A.2 Institutional Details

1.A.2.1 Pathways and Statutory Ages

Pensions in the German public pension system (*gesetzliche Rentenversicherung*) are legally defined in German Social Law, vol. 6 (*Sozialgesetzbuch (SGB) VI*), where a section is devoted to each of the six pathways. First, the *regular pathway* is defined in SGB VI §235. Workers are eligible for this pathway with at least 5 years of contributions (*Wartezeit*, lit. waiting time). A regular pension can only be claimed from the NRA. Hence, the implicit ERA and FRA of the regular pathway coincide with the NRA. The NRA is 65 for workers born until 1946, but for cohorts 1947 to 1964 it increases gradually by one month for each year of birth from 65 to 67 (§235(2)).

Second, the *long-term insured pathway* is defined in §236. Workers are eligible with at least 35 years of contributions. The ERA is 63 throughout the sample period. The FRA is 63 until 1936, is raised gradually by 1 month for each month of birth from 63 to 65 during birth cohorts 1937 and 1938 (SGB VI appendix 21) where it remains until cohort 1948. The FRA increases to 65 and 3 months for cohort 1949 and further increases gradually by one month for each year of birth from 65/3 to 67 for cohorts 1950 to 1964 (§236(2)).

Third, the *women's pathway* is defined in §237a. Women with at least 15 years of contributions are eligible. At least 10 years have to be full contributions, i.e. excluding voluntary contributions, made after their 40th birthday. The ERA is 60 throughout the sample period. The FRA is 60 until 1939, is raised to 65 during cohorts 1940 to 1944 (SGB VI appendix 20) and remains 65 for women born until the end of the sample period. For cohorts born 1952 and later, the pathway will be abolished.

Fourth, the *unemployed/part-time pathway* is defined in §237. Eligibility requires at least 15 years of contributions, and at least 8 out of the 10 years before retirement have to be full contributions. Moreover, the workers must be either unemployed for at least 1 year after age 58 years and 6 months, or in old-age part-time work. Old-age part-time work is a program where workers aged 55 and older reduce their hours to part-time while the decrease in earnings is partly compensated by a government subsidy to the worker. Note that the program has been terminated in 2009. The ERA of this pathway is 60 for workers born until 1945, rises gradually by 1 month for each month of birth from 60 to 63 during birth cohorts 1946 to 1948 (SBG VI appendix 19), and remains 63 until the end of the sample period. The FRA is 60 until 1936, increases gradually by 1 month for each month of birth from 60 to 65 during birth cohorts 1937 to 1940 (SBG VI appendix 19) and remains 65 until the end of the sample period. For cohorts born 1952 and later, the pathway will be abolished.

Fifth, the *disability pathway* is defined in §236a. Workers with at least 35 years of contributions and with an officially recognized disability of at least degree 50% are eligible. The degree of disability is an index factoring in all types of permanent physical and mental conditions. The ERA is 60 throughout the sample period. The FRA is 60 for workers born until 1940, is raised gradually by 1 month for each month of birth from 60 to 63 during birth cohorts 1941 to 1943 (SGB VI appendix 22), and remains 63 until the end of the sample period.

All these pathways are introduced in conjunction with the relevant statutory ages. The NRA (*Regelaltersgrenze*) is defined in §235 as the age from which a regular pension can be claimed. For the remaining pathways, the FRA (*Altersgrenze*) and the ERA (*Alter der frühestmöglichen Inanspruchnahme*) are specified along with the pathways themselves. The FRA is further described as the "age from which an insured person is eligible", while the ERA is the"age from which early claiming is possible".

The sixth pathway, the *invalidity* is defined in §43. Workers are required at least 5 years of

contributions, and at least 3 out the 5 years before retirement must be full contributions. Moreover, workers must have been officially recognized as "low earnings potential", which entails permanently not being able to work more than 3 hours per day in any job. A partial invalidity pension may be available if the worker is deemed to be able to work more than 3 but less than 6 hours per day. Invalidity pensions can be claimed at any age and there is no ERA or FRA in this pathway. Earned points are "filled up" (*Zurechnungszeit*) as if the worker had kept on earning their average pre-retirement income until age 60. Hence, invalidity pensions feature an additional insurance element compared to other pathways since benefits are less dependent on lifetime contributions.

1.A.2.2 Pension Adjustment

Explicit pension adjustment for a worker's retirement age was introduced into the pension formula (*Rentenformel*) in 1997 along with the ERA and FRA reforms described above. The adjustment factor (*Zugangsfaktor*) is defined in §77 SGB and is 100% if a worker claims their pension at the FRA of their pathway. For each month of claiming before the FRA, the adjustment factor (and hence the benefit paid) is reduced by 0.3%, with the maximum negative adjustment implied by the distance between the ERA (the earliest claiming age) and FRA. The adjustment factor remains 100% between the FRA and the NRA. Only after the NRA, there are rewards for late retirement: the adjustment factor increases by 0.5% for each month of claiming after the NRA.

Since 2001, invalidity pensions are also subject to an adjustment factor defined in §77(2)3. Until the end of the sample period, invalidity pensions are decreased by 0.3% for each month of claiming before age 63. There is a maximum negative adjustment of 10.8% that applies to claims below age 60. Moreover, there was a transition period between 2001 and 2003 according to SBG VI appendix 23, where the maximum negative adjustment was gradually increased from 0 to 10.8%. This was done to avoid a notch in the budget set of invalidity workers that would have created a strong incentive to retire before 2001. The end of the filling period of earned points was gradually extended from 55 to 60 at the same time.

1.A.2.3 Benefit Calculation

Upon submitting her pension claim, a worker's benefits B_i are computed according to the following "pension formula":

$$B_i(R_i) = V \cdot \alpha(\max(R_i, ERA)) \cdot \sum_{t=0}^{R_i-1} \frac{w_{it}}{\bar{w}_t} \tag{1.7}$$

The formula has three components. The first component is the *sum of earned points*. In the Bismarckian system, the points a worker earns in a year are equal to her earnings w_{it} relative to the average income among the insured population $\bar{w}_t$. Points are then summed across all years in which contributions were paid. Hence, additional contributions always increase the worker's benefits and pensions become roughly proportional to lifetime income. Second, the worker is assigned an *adjustment factor* α as a function of her benefit claiming age. The benefit claiming age $\max(R_i, ERA)$ is the job exit age if the job exit occurs no earlier than the the ERA, or the ERA otherwise. Adjustment is framed around the FRA, where a worker can claim her *full pension*, i.e. $\alpha(FRA) = 100\%$. The adjustment function α follows a kinked schedule, with a penalty of 0.3% for each month of retirement before the FRA, a reward of 0.5% for each month of retirement after the NRA, and no adjustment between the FRA and the NRA. The third component is the *pension value* V which translates adjusted earned points into monthly benefits. V is indexed to annual nominal wage growth (€26.39 in 2014).

1.A.2.4 Information Provision

In addition to online material and a service hotline, the main way the German pension fund provides information about pensions and retirement is via information letters, whose content is defined in §109. Before June 2002, a detailed information letter (*Rentenauskunft*) was sent to each enrolled worker in the month they turned 55 years old. The frequency of information letters was drastically increased between June 2002 and December 2003. During this transition period, the pension fund conducted surveys of workers and the design of letters was optimized in order to provide information in a more concise and easily comprehensible way. Under the new information provision regime from January 2004, workers are sent a detailed letter every three years from age 55, and a basic letter (*Renteninformation*) is sent annually from age 27.

The basic letter contains information information on contributions paid and points earned so far, the benefit amount the worker is currently eligible for, a projection of her benefit amount if she keeps working until the NRA, and the date on which she will reach the NRA. There is also an explanation of how benefits are calculated, in particular how contributions translate into benefit eligibility, and the tax treatment of pension benefits. Moreover, workers are cautioned about potential losses of purchasing power under different inflation scenarios, and the potential need to supplement public pensions with private savings. In addition to this, the detailed letter provides a more extensive account of the worker's contribution payments so far, and informs about possible retirement dates before and after the NRA with corresponding pension adjustment.

1.A.3 Data

1.A.3.1 Variable Definitions

Job exit ages. A worker's age at benefit claiming and the age of the last contribution can be observed in the data as the distance between the month of birth and the month of claiming or the last contribution. Job exit ages cannot be directly observed, but correspond to the age at the last contribution for most workers. However, for some workers their last month of work does not entail any contributions, or their last month of contributions stems from a status other than employment. To account for this, additional information on the insurance status in the last three years before a worker's benefit claim is used. This status is coded into four categories, 1=work/contributions, 2=no work/no contributions, 3=work/no contributions, 4=no work/contributions. If a worker's last known status is 1 or 2, the last contribution coincides with the job exit. This is the case for 87% of workers in the sample. Categories 3 and 4 pose the problem that the job exit cannot be inferred from the last contribution. However, the timing of job exits can be bounded by the information on workers' status in the three years before retirement. For instance, if a worker is known to be in category 1 20 months before benefit claiming and category 4 8 months before retirement, her job exit is age must have been between 20 months and 8 months before the benefit claiming age. Hence, job exit ages of the remaining workers are imputed via a uniform distribution between the closest known bounds. This imputation is mostly relevant for job exits before the ERA since gaps between job exits and benefit claiming occur in these cases. At the ERA or later, most workers claim benefits right after their job exit so last contributions are not typically confounded by a status other than work.

Years of Contributions. Pathway eligibility is partly determined by a worker's years of contributions (*Wartezeit*, lit. waiting time). Besides contribution periods (*Beitragszeiten*) from employment, a number of other periods such as voluntary contributions of self-employed individuals and "substitute periods" (*Ersatzzeiten*, e.g. due to political imprisonment in the former GDR)

count towards the 15-year threshold. In addition, some periods of education, childcare, sick leave, receipt of some types of unemployment benefits and the invalidity filling period (*Berücksichtigungs-, Anrechnungszeiten*) count towards the 35-year threshold. The contribution periods actually used for pension calculation cannot be observed directly in the data, but they can be reconstructed from a variables related to workers' earnings histories. Around the 15-year threshold, contributions are calculated as the sum of contribution (both full and partial) and substitute periods. For the 35-year threshold, other relevant periods are added as far as they are observed.

Lifetime budget constraints. Lifetime budget constraints are simulated based on the formulas presented in section 1.2.3. First, a pension benefit calculator is constructed according to equation (1.7) using a sample period average pension value V, a worker's observed sum of earned points $\sum_{t=0}^{R_i-1} \frac{w_{it}}{\bar{w}_t}$ and the adjustment factor function $\alpha(R_i, ERA)$ that applies to their specific pathway and birth cohort. Individual lifetime wealth at the worker's actual job exit age is then computed according to equation (1.1) with a discount factor of 3% and remaining life expectancies at age 55 taken from mortality tables by the German Federal Statistics Office taking into account heterogeneity by gender and year of birth. Lifetime gross wage earnings are approximated as the sum of earned points multiplied by an average of mean annual incomes across the sample period. Net earnings are calculated from gross earnings using an tax simulator taking into account personal income tax and social insurance contributions, and income splitting is applied to married individuals. Since the budget constraint abstracts from periods of inactivity, the starting age is set to 25 years, a value that would generate roughly the observed average earned points if all workers had uninterrupted earnings careers.

In order to simulate lifetime wealth across a range of job exit ages, an approximation of annual earnings w_{it} is needed. A lifetime average of gross annual earnings is computed as lifetime wage earnings divided by the hypothetical uninterrupted career length from age 25 until the observed job exit age. Net annual earnings are calculated using the income tax simulator. A worker's lifetime wealth can then be simulated across a range of job exit ages by extrapolating additional income from work based on annual earnings and simulating pensions across claiming ages, the latter taking into account additional contributions and changing adjustment. Monthly implicit net wages are calculated as the increment in simulated lifetime wealth, and the implicit net-of-tax rate is the implicit net wage divided by gross income.

1.A.3.2 Group Assignment

Pathway eligibility. As explained in section 1.2.4, workers choose the pathway from which to claim a pension, and reforms induce some partly mechanical switching between pathways. In particular, when FRAs are increased to 65 in a certain pathway, an increase in the number of workers eligible for that pathway claiming regular pensions can usually be observed. This occurs because there is no difference in benefits across pathways at the NRA and beyond, and workers may perceive claiming a regular pension as easier or more natural than claiming a special, non-regular pension. To account for this, pathway assignment is based on eligibility in order to keep group composition as stable as possible.

Pathway eligibility is based on observable characteristics where possible, with some imputation to account for unobservables. Workers with at least 35 years of contributions are eligible for the long-term insured pathway. For the women's pathway, women with at least 15 years of contributions are deemed eligible. The additional requirement of full contributions in 8 out of the last 10 years cannot be used since the exact timing of contributions is insufficiently observable. Workers are defined as eligible for the the unemployed/part-time pathway if they have at least 15 years of contributions, and they are observed to be unemployed or in part-time work within the last 3 years

before benefit claiming. Unfortunately, disability cannot be observed in the data, but a subset of workers satisfying the contribution requirements of the disability and invalidity pathways of 35 and 5 years, respectively, is identified.

If a worker is eligible for only one pathway, assignment is unambiguous. Moreover, workers who are observed to claim from one of the non-regular pathways are assumed to have chosen their "best" pathway and are thus assigned. Among the remaining workers who are found eligible for more than one pathway, assignment is based on a notion of which of those pathways is most advantageous. For instance, if a woman is eligible for the women's pathway, she must also be eligible for the regular pathway, but the feasible set of retirement age/consumption combinations in the women's pathway dominates that of the regular pathway because both ERA and FRA are lower. Besides, she may be eligible for the unemployed/part-time and/or long-term insured pathways, but those are also dominated by the women's pathway. Hence, women claiming a regular pension who are eligible for the women's pathway (and possibly unemployed or long-term insured) are assigned to the women's pathway rather than regular. Unemployed/part-time is assigned analogously.

Both long-term insured and disability pathways require at least 35 years of contributions, but among the workers satisfying this, only those with an official disability can choose the disability pathway. Since counterfactual disability status cannot be observed, the share of workers satisfying the requirement has to be imputed. In particular, it is assumed that the relative shares of disabled individuals among those potentially eligible for both pathways is the same as the shares among those actually claiming in the pathways at a given age. Hence, the ratio of disability/long-term insured claimants is computed for each integer retirement age in each year of birth, and ambiguous cases are assigned based on the corresponding ratio. Similarly, invalidity and regular pensions both require only 5 years of contributions, and the ratio of actual claimants by year of birth and integer retirement age is used to impute eligibility in ambiguous cases.

As shown in table 1.2, the most important difference between the number of actual claimers and eligible workers arises in the regular pathway where eligibility is largely overestimated by claiming. Hence, many regular claimers are eligible for more advantageous pathways, particularly long-term insured and women's pathways. The vast majority of these switchers are workers retiring at the NRA and beyond, where they receive the same benefits from the regular pathway as they would from other pathways.

Groups and Discontinuities. Workers are grouped into cells by year of birth and pathway, since this split accounts for most of the variation in statutory ages and lifetime budget constraints faced by workers, while still preserving sufficiently large group sizes for the purpose of bunching estimation. During the cohorts where reforms change statutory ages at the month-of-birth level, workers around the statutory age in the affected pathway are grouped by pathway and month of birth instead. This split yields a total number of 420 groups of whom 108 are at the year-of-birth and 312 at the month-of-birth level.

Moreover, there are seven types of notches created by pathway contribution thresholds. At 5 years of contributions, workers switch from no pension at all to either regular or invalidity. At 15 years of contributions, women switch from the regular pathway to the women's pathway. Moreover, workers who are unemployed or in old-age part-time work before retirement switch from regular to that pathway at 15 years of contributions. At 35 years of contributions, regular workers switch to the long-term insured or disability pathway. Finally, workers previously eligible for the women's or unemployed pathway may switch to the disability pathway at 35 years. For each year of birth, workers around a notch are identified based on pathway eligibility as described above. Restrictions in terms of years of contributions are relaxed in order to observe workers to the left of the notch who are close to the threshold but, by definition of the threshold, cannot yet be observed to claim the corresponding pathway. In order to account for variation in the notch size depending on retirement

ages, each year of birth and type of notch is further divided into two ranges of retirement ages, 55 to 60 and 60 to 65. This yields a total of 78 groups each of whom faces one notch.

1.A.3.3 Survey Data

Survey Sample and Variables. The German Socioeconomic Panel (SOEP) is a panel household survey, of which the waves 1984 to 2013 are used. In total, there are 175,224 working individuals whose occupation is reported. To maximize power, all age groups are used to compute occupation-level averages. There are an average of 475 workers in each 3-digit occupation cell. The following variables of interest can be directly observed in the survey: union membership, active union membership, currently in unlimited contract, severance paid upon job exit, involuntary job exit. A firm size index is computed based on the size categories <20 employees, 20 to 200, 200 to 2000, and >2000 employees. Tenure on the job can be computed as the time from the month of job start to the month of interview.

Matching at Occupation Level. In the administrative data, occupations are reported at the 3-digit level according to the *KldB 1988* classification. The survey data reports occupations according to the slightly updated *KldB 1992* classification. A mapping between the two classifications is created manually. Among the 337 3-digit KldB 1988 occupations, 90% have a unique match in KldB 1992. 10% have two or more matches, and 4% have three or more matches. To get occupation-level values, the occupation-level average from the survey data is taken if the occupation has a unique match. If there is more than one match, an average weighted by the size of each occupation cell among the matches is taken.

1.A.4 Empirical Methodology

1.A.4.1 Bunching Estimation

The bunching estimation is based on Chetty et al. (2011) where a counterfactual density is fitted to the observed distribution of job exit ages around each discontinuity, excluding the data in the bunching region around the discontinuity. The counterfactual C_j is estimated as a regression of the form

$$C_j = \sum_{i=0}^{p} \beta_i (R_j)^i + \sum_{r \in \Gamma} \delta_r \mathbb{1}(R_j = r) + \sum_{k=R^-}^{R^+} \gamma_k \mathbb{1}(R_j = k) + \varepsilon_j$$

where C_j is the number of individuals in monthly job exit age bin j, Γ is a set of round retirement age types, and $[R^-, R^+]$ is the excluded range of job exit ages around the discontinuity. Hence, the regression fits a p-th order polynomial to the distribution of job exit ages, while allowing for additional round-number bunching through the coefficients δ_r. The counterfactual density at the discontinuity is then predicted as

$$\hat{C}_j = \sum_{i=0}^{p} \hat{\beta}_i (R_j)^i + \sum_{r \in \Gamma} \hat{\delta}_r \mathbb{1}(R_j = r)$$

thus omitting the contribution of the dummies in the excluded range. The bunching mass $\hat{B} = \sum_{k=R^-}^{R^+} C_j - \hat{C}_j$ is the difference between the observed and the counterfactual distribution in the bunching region. Finally, the excess mass is defined as bunching relative to the counterfactual

density:

$$\hat{b} = \frac{\hat{B}}{\sum_{k=R^-}^{R^+} \hat{C}_j/(R^+ - R^- + 1)}$$

In practice, the order of the polynomial is chosen as $p = 7$ and the excluded range $[R^-, R^+]$ as well as the set of round ages Γ to control for are determined separately for each type of discontinuity. Around statutory ages, the bunching region is generally defined as the discontinuity and one additional month on either side. Round-age dummies are included for each full-year age above 55, where additional dummies for full-year ages above 60 and 64 allow for heterogeneity in round-number bunching by age. Other statutory ages that may fall in the estimation range are also netted out of the counterfactual by dummies. Between 24 and 36 bins are included on both sides of the discontinuity for the estimation of the polynomial, with the exception of ERAs where only 12 bins are included to the left. In the regular pathway, invalidity and some cohorts of unemployed/part-time, round-number dummies are not included because there is no visible round-number bunching. In invalidity, bunching is restricted to the month of the discontinuity itself as there is no visible diffuse bunching mass. For groups at the month-of-birth level, dummies for job exit ages that fall in the calendar month of December are additionally included in Γ. December effects are also allowed to be heterogeneous across 5-year age ranges. The estimation around the pathway switching notches includes 120 bins on each side of the notch in order to increase statistical power, and has no round-number dummies. The month of the notch itself and 12 months to the left are excluded to account for missing mass. Bunching is estimated sharply at the month of the notch. The missing mass is extended to 24 months in the long-term insured pathway to line up with the relatively larger bunching mass.

Observed elasticities are calculated at each discontinuity according to equation (1.2). Kink sizes are computed as the marginal implicit net-of-tax rate just before the kink divided by the rate at the kink. Notches are approximated as kinks faced by the marginal buncher: The average net-of-tax rate between the location of the marginal buncher and the notch is used as the rate before the kink, and divided by the actual marginal net-of-tax rate after the kink. Standard errors for individual bunching mass estimates are bootstrapped by re-sampling the individual data within the respective group. Standard errors for regressions based on bunching estimates are block bootstrapped, that is the data is re-sampled at the discontinuity level.

1.A.4.2 Discontinuities Used for Bunching

The following table lists all discontinuities where bunching is estimated. Note that 11 out of the 655 discontinuities where the local density is too low to estimate a stable counterfactual are excluded from the main analysis.

Pathway	Cohorts	Age Group	Frequency	Source of Discontinuity	Type	Number
Regular	1933-1949	55-67	annual	ERA=FRA=NRA	kink	17
Long-term insured	1933-1936	55-67	annual	ERA=FRA	kink	4
Long-term insured	1937-1949	55-67	annual	ERA	kink	13
Long-term insured	1939-1946	55-67	annual	FRA=NRA	kink	8
Long-term insured	1947-1948	55-67	annual	FRA	kink	2
Long-term insured	1933-1938 1947-1949	55-67	annual	NRA	kink	9
Long-term insured	1937-1938 1949	55-67	monthly	moving FRA	kink	36
Women	1933-1939	55-67	annual	ERA=FRA	kink	7
Women	1940-1949	55-67	annual	ERA	kink	10

Women	1945-1946	55-67	annual	FRA=NRA	kink	2
Women	1947-1949	55-67	annual	FRA	kink	3
Women	1933-1944	55-67	annual	NRA	kink	15
	1947-1949					
Women	1940-1944	55-67	monthly	moving FRA	kink	60
Unemp./part-time	1933-1936	55-67	annual	ERA=FRA	kink	4
Unemp./part-time	1937-1945	55-67	annual	ERA	kink	9
	1949					
Unemp./part-time	1942-1946	55-67	annual	FRA=NRA	kink	5
Unemp./part-time	1947-1949	55-67	annual	FRA	kink	3
Unemp./part-time	1933-1941	55-67	annual	NRA	kink	12
	1947-1949					
Unemp./part-time	1937-1941	55-67	monthly	moving FRA	kink	60
Unemp./part-time	1946-1948	55-67	monthly	moving ERA	kink	36
Disability	1933-1940	55-67	annual	ERA=FRA	kink	8
Disability	1941-1949	55-67	annual	ERA	kink	9
Disability	1944-1949	55-67	annual	FRA	kink	6
Disability	1933-1949	55-67	annual	NRA	kink	17
Disability	1941-1943	55-67	monthly	moving FRA	kink	36
Invalidity	1938-1949	55-67	annual	pension adjustment around age 63	kink	12
Invalidity	1938-1943	55-67	monthly	adjustment introduction in 2001	kink	72
Long-term insured	1937-1949	55-63/0	annual	35 year contribution threshold (from regular)	notch	13
Long-term insured	1938-1943	63/1-65	annual	35 year contribution threshold (from regular)	notch	17
Women	1937-1949	55-60/0	annual	15 year contribution threshold (from regular)	notch	13
Women	1933-1949	60/1-65	annual	15 year contribution threshold (from regular)	notch	17
Unemp./part-time	1937-1949	55-60/0	annual	15 year contribution threshold (from regular)	notch	13
Unemp./part-time	1933-1949	60/1-65	annual	15 year contribution threshold (from regular)	notch	17
Disability	1937-1949	55-60/0	annual	35 year contribution threshold (from regular)	notch	13
Disability	1933-1949	60/1-65	annual	35 year contribution threshold (from regular)	notch	17
Disability	1937-1949	55-60/0	annual	35 year contribution threshold (from unemp.)	notch	13
Disability	1933-1949	60/1-65	annual	35 year contribution threshold (from unemp.)	notch	17
Disability	1937-1949	55-60/0	annual	35 year contribution threshold (from women)	notch	13
Disability	1933-1949	60/1-65	annual	35 year contribution threshold (from women)	notch	17
total						655

Chapter 2

Bunching Responses to Reference Points: Theory and Applications

2.1 Introduction

The notion of reference-dependent preferences has gained increasing empirical support in recent studies. Across different contexts, results suggest individuals seem to evaluate outcomes relative to reference points. A classic example is Camerer et al. (1997) who show that taxi drivers evaluate their earnings relative to daily targets. DellaVigna et al. (2018), Rees-Jones (2018) and Allen et al. (2016) show evidence of reference-dependent behavior among job seekers, tax filers, and marathon runners, respectively. Similarly, the first chapter of this thesis argues that framing pensions and retirement decisions around statutory retirement ages induces workers to perceive those thresholds as reference points.

The defining feature of reference dependence is some form of discontinuity in the way individuals evaluate an outcome. For instance, loss aversion, a commonly used type of reference dependence, entails a discontinuous change in marginal utility at the reference point. A direct consequence of such a discontinuity in preferences is bunching of the outcome at the reference point. Indeed, Rees-Jones (2018) and Allen et al. (2016) argue in favor of reference dependence based on bunching that occurs at a threshold with no extrinsic incentives. At the same time, there is a large and growing literature on bunching (see Kleven 2016 for a survey) exploiting discontinuities in extrinsic incentives to measure responses to these incentives. The bunching literature provides methods to estimate excess bunching and to link observed responses neatly to underlying parameters, such as the elasticity of labor supply with respect to the net-of-tax rate.

This chapter argues that bunching methods are naturally suited to quantify reference-dependent preferences and to recover underlying parameters. To do so, I exploit the direct analogy between responses to reference points and responses to discontinuities in extrinsic incentives. Using a standard labor supply model, the workhorse of the bunching literature, I first show that the different types of reference dependence considered in previous studies have a key prediction in common:

They imply sharp bunching of the outcome at the reference point. Moreover, observed bunching responses can be linked to the parameters governing reference dependence. On this basis, I propose both structural and reduced-form estimation methods to implement a bunching approach to reference dependence. Finally, the chapter presents two empirical applications of these ideas in the context of retirement decisions. The first application estimates responses to a type of pure reference point, namely round retirement ages. The second application extends the analysis from chapter 1, recovering reference dependence parameters from bunching at statutory retirement ages, and the resulting estimates are used to simulate policy counterfactuals.

I divide the analysis in this chapter into two parts. The first part characterizes bunching responses to reference points in a standard labor supply model. This type of model is commonly used to analyze bunching at a budget constraint kink where the marginal tax rate changes. Two different types of reference dependence are then incorporated into the model. The first type is "loss aversion", a discontinuity in marginal utility at the reference point dividing the choice set into gains and losses (Kahneman and Tversky 1979). Loss aversion can occur relative to a consumption level, or as a discontinuity in marginal disutility from work, which can also be interpreted as loss aversion in leisure. Interpretations put forward for such "utility kinks" include workers' expectations or endowment effects with respect to a status quo. The second type of reference dependence is a "utility notch", where the level of utility jumps at the reference point (Allen et al. 2016). Utility notches can be one-sided, for instance when workers perceive a certain level of consumption or labor supply as a goal. It could also be that individuals derive discontinuously higher utility only exactly at the reference point, and deviations in any direction are costly. Such a two-sided utility notch can be motivated by a perceived norm in favor of the threshold, for example.[1]

Next, I show that all types of reference dependence have a key prediction in common: sharp bunching of labor supply at the reference point. Analogously to a budget constraint discontinuity, a number of individuals choose to locate at the reference point as a corner solution. There are however important differences in the direction of labor supply responses and the implied shape of the density around the reference point. For instance, loss aversion in consumption induces workers to increase labor supply towards the reference point, such that bunching occurs from below. Loss aversion in the labor supply/leisure dimension, on the other hand, implies that workers reduce labor supply towards the reference point, and bunching occurs from above. Moreover, utility notches produce "holes", i.e. missing mass in the density, but utility kinks do not.

The theoretical results can be useful for several types of empirical applications. The first and most straightforward possibility is a "pure" reference point without any economic incentives. In this case, parameters can be directly estimated from observed bunching at the reference point. A second, somewhat more challenging case for estimation is when a potential reference point coincides with economic incentives, such as in the context of chapter 1. Here, bunching must be observed at this combined threshold, and at at least one other threshold varying in the underlying economic

[1] Another possible type of reference dependence is "diminishing sensitivity", a discontinuity in the second derivative of utility. Since this feature is mostly relevant for choice under uncertainty, it is not considered in this chapter.

incentive and/or the presence of a reference point. It is then possible to jointly estimate reference dependence parameters and an elasticity with respect to the incentive. A final task for estimation may be to empirically distinguish between different types of reference dependence. One way to do this is to rely on the shape of the distribution around the reference point. For instance, to distinguish between a kink in utility from consumption and a kink in disutility from work, it is sufficient to estimate the share of responses originating from below vs. above in addition to bunching at the reference point.

The above estimation strategies can be implemented via a structural approach, parametrically specifying the type(s) of reference dependence at work. However, I also suggest a reduced-form method to estimate the quantitative importance of reference dependence without such an assumption. In particular, a simple linear decomposition of bunching at a potential reference point can be interpreted as an approximation of the structural bunching equations from the theoretical framework. For both structural and reduced-form methods, identification requires observing at least as many bunching moments as parameters are to be estimated. They key identification assumption is that bunching at different thresholds is generated by the same underlying parameters.

The second part of this chapter presents two empirical applications in the context of retirement decisions, using the empirical setting from chapter 1. The first application estimates responses at a pure reference point. Bunching at round numbers has been observed in different contexts and has been attributed to reference dependence (Kleven 2016). Similarly, there is clear bunching at round ages in the distribution of job exits of German workers.[2] While most existing bunching applications treat round-number effects as a confounder of the effect of incentives, I aim at quantifying round-number effects and estimating implied reference dependence parameters. The sharp bunching at round retirement ages translates into a strongly significant excess mass between 0.6 and 1.1. Strikingly, this implies that workers bunch at arbitrary round numbers at a similar magnitude to bunching in response to some substantial financial incentives from chapter 1. In terms of magnitudes, the estimated kinks in disutility from work at round ages correspond to kinks in the implicit tax rate of 11% to 17%.

The second application extends the analysis of statutory retirement ages presented in this thesis. The evidence from the first chapter suggests that workers perceive statutory ages as reference points[3], and the estimation in this chapter has two additional goals. First, the quantitative importance of reference dependence vs. financial incentives in explaining bunching at statutory ages

[2]The analysis focuses on round ages other than those framed as statutory retirement ages in order to isolate the effect of round numbers.

[3]Reference dependence may not be the only way to rationalize the strong documented responses at statutory ages, but the approach has several advantages. First, workers perceiving statutory ages as reference points is a plausible notion from the viewpoint of the institutional setting, where they are framed as such. Second, there is complementary evidence from surveys and experiments, suggesting that retiring at statutory ages may be perceived as a norm, and providing support for reference-dependent behavior around statutory ages (Shoven et al. 2017, Merkle et al. 2017). Third, the reference dependence model replicates key features of observed retirement behavior, including sharp bunching at statutory ages even if there is a disincentive to retire. Fourth, reference-dependent decision utility is a fairly general way to model the effect of interest, capturing a number of potential sources. The evidence suggests that framing effects may be at work, but this may not exclude alternative sources such as social norms or "deep" internal preferences leading to the same type of behavior.

can be gauged. I find large and significant reference point effects, and the estimated parameters imply that 50% to 80% of actual statutory age retirements in the data are attributed to reference dependence. The second goal is to empirically distinguish between different types of reference dependence. Since statutory ages are framed by policy as reference points both in terms of consumption and in terms of the retirement age itself, reference dependence in these two dimensions may be plausible. Results suggest that bunching at the Early Retirement Age seems to be due to workers postponing retirement and is mainly driven by reference dependence in consumption. Bunching at the Normal Retirement Age is driven by reference dependence in work/leisure and occurs due to workers moving retirement forward.

In addition, a key advantage of the structural approach relative to the reduced-form estimation from chapter 1 is that counterfactuals can be simulated in order to highlight policy implications. First, an increase in the Normal Retirement Age from 65 to 66 is predicted to lead to an increase in average actual retirement ages by 4 months. Second, I simulate the effect of providing more financial incentives for late retirement. Although both policies are calibrated to have the same effect on average retirement ages, the fiscal impact is very different: The increase in the Normal Retirement Age entails a net fiscal gain +€700m, whereas the increase in financial rewards would lead to a net loss of -€200m. The difference in fiscal effects arises because the increase in late retirement rewards more than offsets additional contributions due to longer working lives, while shifting a statutory age is virtually costless to the government.

Hence, reforms shifting statutory ages are effective in influencing retirement behavior and can generate a positive fiscal impact. The overall welfare effects of the policies crucially depend on the extent to which the planner values reference point effects, however. In particular, the government may not want to fully "exploit" reference dependence with respect to statutory ages if it represents a pure bias. In this case, increasing financial rewards may even be the preferred policy, since its negative fiscal effect can be more than offset by the gain in workers' utility from moving away from their inefficient choice of retiring at the reference point.

This chapter relates mainly to two strands of literature. First, it contributes to the bunching literature reviewed by Kleven (2016). Bunching methods have been pioneered bySaez (2010) and Chetty et al. (2011) in the context of taxable income responses. Applications to retirement decisions include Brown (2013) and Manoli and Weber (2016*a*). Bunching is generally used to estimate responses to extrinsic incentives, and some studies extend the method to account for optimization frictions (Kleven and Waseem 2013, Gelber et al. 2017). This chapter is the first to incorporate reference dependence into a standard bunching model, highlighting that bunching methods can be used to estimate responses to intrinsic, non-financial factors.

Second, the chapter contributes to the literature on the role of reference points. Despite its prominence in behavioral models such as prospect theory (Kahneman and Tversky 1979), there are relatively few studies on the impact of reference dependence in field settings (Barberis 2013). DellaVigna et al. (2018) demonstrate loss aversion among job seekers, and Rees-Jones (2018) and Allen et al. (2016) suggest reference dependence among tax filers and marathon runners, respec-

tively.[4] Showing bunching at reference points, the last two studies come closest to the methods used in this chapter. This chapter links bunching responses to reference dependence parameters and systematically develops estimation strategies for this purpose. Moreover, the applications in this chapter add a new important context by measuring reference dependence in retirement behavior.

The remainder of this chapter is organized as follows. Section 2.2 sets up the labor supply model and characterizes bunching at reference points,, section 2.3 develops estimation strategies based on the theoretical results, section 2.4 presents the two empirical applications in the retirement context, and section 2.5 concludes.

2.2 Reference Dependence and Bunching in a Labor Supply Model

2.2.1 Basic Setup and Bunching at a Budget Constraint Kink

This section sets out the basic labor supply model used to analyze bunching responses. Following the bunching literature (Saez 2010, Kleven 2016) the analysis is framed in terms of taxable income as the running variable. Taxable income z is generated from labor supply h paid at wage rate w. It is straightforward to adapt the model to account for other choice variables. For instance, section 2.4.1 presents an application to retirement ages.

Workers maximize utility

$$U = u(c) - v(z, n)$$

c is consumption, z is taxable income and n is an ability parameter. Utility is increasing and concave in consumption and disutility from work is convex such that $u'(c) > 0$, $u''(c) < 0$, $v_z > 0$, and $v_{zz} > 0$. Moreover, low earnings ability increases disutility from work such that $v_{zn} > 0$ The budget constraint is

$$c = z - T(z) \tag{2.1}$$

where w is the wage rate, and $T(z)$ is a function capturing the tax schedule that earnings are subject to.

Consider first the case of a linear budget constraint with a constant net-of-tax rate $1 - \tau$, and as is standard in the bunching literature, a utility function that is quasi-linear in consumption and iso-elastic in labor supply such that

$$U = (1 - \tau)z - \frac{n}{1 + \frac{1}{\varepsilon}} \left(\frac{z}{n}\right)^{1+\frac{1}{\varepsilon}} \tag{2.2}$$

where ε is the elasticity of taxable income with respect to the implicit net-of-tax rate. Workers' utility maximization yields

$$z = n(1 - \tau)^{\varepsilon}$$

[4] In addition, there is some empirical evidence on reference dependence from the behavioral finance literature, for example on the "disposition effect" (Barberis and Xiong 2009).

If the distribution of ability $F(n)$ is smooth, this implies a smooth distribution of taxable income with density $h_0(z)$.

Suppose now that there is a kink in the budget constraint such that the net-of-tax rate decreases by $\Delta\tau$ at some level of taxable income $\hat{z}$. Figure 2.1 illustrates the effect of the budget set kink in a budget set diagram and density diagram following Saez (2010) and Kleven (2016). Whilst an individual with ability $\hat{n}$ initially chooses taxable income $\hat{z}$, there is a marginal buncher with ability n^* whose indifference curve is tangent to the initial budget set at z^* and to the upper part of the new budget set at $\hat{z}$. All workers initially located between $\hat{z}$ and z^* bunch at the kink, while all individuals initially to the left of the kink leave their labor supply unchanged and all individuals initially to the right of z^* stay above the kink.

Total bunching is

$$B = \int_{\hat{z}}^{z^*} h_0(z)dz \approx h_0(\hat{z})(z^* - \hat{z})$$

where $h_0(\hat{z})$ is the pre-kink density and the approximate equality holds if $h_0(z)$ is constant on $[\hat{z}, z^*]$. The two tangency conditions for the marginal buncher imply $z^* = n^*(1-\tau)^\varepsilon$ and $\hat{z} = n^*(1-\tau-\Delta\tau)^\varepsilon$ and thus

$$\frac{z^*}{\hat{z}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^\varepsilon \tag{2.3}$$

or, in terms of the excess mass $b = B/h_0(\hat{z})$,

$$\frac{b}{\hat{z}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^\varepsilon - 1 \tag{2.4}$$

2.2.2 Reference Dependence

Throughout this chapter, reference dependence is defined as a situation where workers evaluate outcomes relative to some threshold level of consumption or labor supply. In the language of the bunching literature, this corresponds to a discontinuity in utility at the reference point.[5] This section considers four versions of reference-dependent preferences, all of which arguably have some plausibility or have been discussed in previous literature (e.g. Allen et al. 2016). There are two dimensions along which the variants differ. First, there can be a reference point in terms of consumption or in terms of labor supply. Second, it can be parametrized as a kink (i.e. a discontinuity in marginal utility) or a notch (i.e. a discontinuity in the level of utility).

Version 1a: Kink in utility from consumption

$$U = u(z) - v(z, n) - \mathbb{1}(c \leq \hat{c}) \cdot \lambda_c(\hat{c} - c) \tag{2.5}$$

[5] This relatively general definition may also include notions not typically subsumed under reference dependence, including norms and other utility costs of deviating from certain reference levels. The common feature that creates discontinuities in utility are that they lead to discontinuous responses for intrinsic reasons, as opposed to the extrinsic incentives usually considered in the bunching literature.

Version 1b: Kink in disutility from work

$$U = u(c) - v(z, n) - \mathbb{1}(z \geq \hat{z}) \cdot \lambda_l (z - \hat{z}) \tag{2.6}$$

The last term in equation (2.5) introduces a discrete jump in marginal utility from consumption at $\hat{c}$ where the parameters $\lambda_c > 0$ captures the size of this kink. Thus, workers' marginal utility gain from approaching $\hat{c}$ from below is greater than their marginal utility gain from increasing consumption beyond the level at this reference point. This corresponds to the loss aversion property from prospect theory (Kahneman and Tversky 1979), where the choice set is divided into two domains, gains and losses. Panel A of figure 2.2 shows the impact on the utility function and indifference curves. Indifference curves exhibit a convex kink at $\hat{c}$ in this case. On the other hand, the last term in equation (2.6) introduces a discrete jump in the marginal disutility from work at $\hat{z}$. Panel B of figure 2.2 shows that this introduces a convex kink in indifference curves at $\hat{z}$. This can also be interpreted as loss aversion with respect to a reference level of leisure. Interpretations put forward for such utility kinks include workers' expectations or endowment effects with respect to a status quo (Kleven 2016).

Version 2a: One-sided utility notch

$$U = u(c) - v(z, n) + \mathbb{1}(z \geq \hat{z}) \cdot \delta \tag{2.7}$$

Version 2b: Two-sided utility notch

$$U = u(c) - v(z, n) + \mathbb{1}(z = \hat{z}) \cdot \delta_2 \tag{2.8}$$

Equations (2.7) and (2.8) introduce a discrete jump in utility at $\hat{z}$ where the parameters $\delta > 0$ and $\delta_2 > 0$ capture the size of this notch. Equation (2.7) describes a situation where the worker's satisfaction jumps when they attain at least $\hat{z}$.[6] As shown in Panel C of figure 2.2, this implies a notch in indifference curves at $\hat{z}$. Reasons for a one-sided notch may include workers perceiving $\hat{z}$ as a level of aspiration or a goal. On the other hand, equation (2.8) corresponds to a case where utility increases only when labor supply is exactly $\hat{z}$. Panel D shows that such a two-sided notch implies that indifference curves are unchanged, except a point originally below the curve is now part of a given indifference set. For instance, this could be motivated by a norm in favor of $\hat{z}$, or $\hat{z}$ serving as a default option where any deviations are costly to workers.

2.2.3 Bunching Responses at Reference Points

This section shows how sharp bunching occurs at reference points of any of the four types described above. The budget constraint is assumed to be linear with net-of-tax rate $1 - \tau$ throughout.

[6] Unlike in the case of utility kinks, it does not matter whether utility notches are written in terms of labor supply or consumption.

2.2.3.1 Kink in Utility from Consumption

Suppose first that there is a consumption reference point as in equation (2.5). The left panels of figure 2.3 illustrates the impact in the usual budget set and density diagrams. In the absence of the reference point, workers locate along the budget line depending on their abilities. An individual with ability $\hat{n}$ is initially located at $\hat{z}$ and n^* is located at z^*. When the reference point is introduced, the individual whose indifference curve was initially tangent to the budget line at z^* sees a clockwise rotation of their indifference curves below the reference point, and the new point of tangency is at $\hat{z}$. This individual is the marginal buncher: All workers initially located between z^* and $\hat{z}$ now bunch at the reference point, while all individuals initially to the right of the reference point leave their labor supply unchanged and all individuals initially to the left of z^* stay below the reference point. Like a kink in the budget constraint, a kink in utility does not produce a hole in the density of taxable incomes, since workers initially below z^* also work more, causing a rightward shift in the density below $\hat{z}$ that fills the hole.

Bunching at the reference point is

$$B = \int_{z^*}^{\hat{z}} h_0(z)dz \approx h_0(\hat{z})(\hat{z} - z^*)$$

With quasi-linear and iso-elastic utility, the two tangency conditions for the marginal buncher imply $z^* = n^*(1-\tau)^\varepsilon$ and $\hat{z} = n^*[(1+\lambda_c)(1-\tau)]^\varepsilon$. Hence

$$\frac{z^*}{\hat{z}} = \left(\frac{1}{1+\lambda_c}\right)^\varepsilon \tag{2.9}$$

or

$$\frac{b}{\hat{z}} = 1 - \left(\frac{1}{1+\lambda_c}\right)^\varepsilon \tag{2.10}$$

Bunching in response to a kink in utility from consumption only depends on the strength of the reference point λ_c and the the elasticity and is independent of the net-of-tax rate. Intuitively, the marginal gain in consumption is valued more by workers below the reference point, and the elasticity determines by how much they are willing to increase labor supply in response. Equation (2.10) implies that the parameter λ_c can be estimated can be calculated based on observed bunching, given an estimate of the elasticity.[7]

2.2.3.2 Kink in Disutility from Work

Suppose instead that there is a labor supply reference point as in (2.6). In the right panel of figure 2.3, an individual with ability $\hat{n}$ is initially located at $\hat{z}$ and n^* is located at z^*. The marginal buncher whose indifference curve was initially tangent to the budget line at z^* sees a counter-clockwise rotation of their indifference curves to the right of the reference point, and the new point of tangency is at $\hat{z}$. All workers initially located between $\hat{z}$ and z^* now bunch at the reference point,

[7] Section 2.3 describes in more detail how reference dependence parameters can be estimated.

while all individuals initially to the left of the reference point leave their labor supply unchanged and all individuals initially to the right of z^* stay above the reference point. Note that bunching occurs from the right in this case, while bunching in response to the consumption reference point is from the left. Again, the kink in utility does not produce a hole in the density of taxable incomes, since workers initially above z^* also work less, causing a leftward shift in the density above $\hat{z}$.

Bunching at the reference point is

$$B = \int_{\hat{z}}^{z^*} h_0(z)dz \approx h_0(\hat{z})(z^* - \hat{z})$$

The two tangency conditions for the marginal buncher imply $z^* = n^*(1-\tau)^\varepsilon$ and $\hat{z} = n^*(1-\tau-\lambda_l)^\varepsilon$. Hence

$$\frac{b}{\hat{z}} = \left(\frac{1-\tau}{1-\tau-\lambda_l}\right)^\varepsilon - 1 \tag{2.11}$$

Equations (2.4) and (2.11) implies that a kink in disutility from work has the same bunching effect as a kink in the budget set. Indeed, a local change in the net-of-rax rate $\Delta\tau = \lambda_l$ would produce exactly the same bunching response.

2.2.3.3 One-Sided Utility Notch

Consider next a one-sided utility notch as in equation (2.7). In the left panel of figure 2.4, an individual with ability $\hat{n}$ is initially located at $\hat{z}$ and n^* is located at z^*. When the utility notch is introduced, indifference curves become discontinuous at $\hat{z}$, and the individual initially located at z^* is now indifferent between z^* and $\hat{z}$. This worker is the marginal buncher. All workers initially located between z^* and $\hat{z}$ bunch at the reference point, while individuals initially to the left of z^* or to the right of $\hat{z}$ do not alter their labor supply. Hence, all bunching originates from the left. Moreover, the utility notch produces a hole in the density since no individual is willing to locate between z^* and $\hat{z}$.

Bunching at the reference point is

$$B = \int_{z^*}^{\hat{z}} h_0(z)dz \approx h_0(\hat{z})(\hat{z} - z^*)$$

Based on the utility specification in equation (2.2), utility of the marginal buncher at $\hat{z}$ is

$$\hat{U} = (1-\tau)\hat{z} - \frac{n^*}{1+\frac{1}{\varepsilon}}\left(\frac{\hat{z}}{n^*}\right)^{1+\frac{1}{\varepsilon}} + \delta$$

Using the first-order condition $z^* = n^*(1-\tau)^\varepsilon$, utility at the initial interior solution z^* can be expressed as

$$U_I = \frac{1}{1+\varepsilon} n^*(1-\tau)^{1+\varepsilon}$$

The indifference condition $\hat{U} = U_I$ then implies

$$\frac{1}{1+\varepsilon}\frac{z^*}{\hat{z}} + \frac{\varepsilon}{1+\varepsilon}\left(\frac{z^*}{\hat{z}}\right)^{-\frac{1}{\varepsilon}} = 1 + \frac{\delta}{c(\hat{z})} \tag{2.12}$$

where $c(\hat{z}) = (1-\tau)\hat{z}$. Equation (2.12) defines bunching in response to a given utility notch, although there is generally no closed-form solution for b or $z^*/\hat{z}$. The parameter δ can be directly calculated from observed bunching, given an estimate of the elasticity.

Note that the solution given by equation (2.12) is conceptually similar to bunching at a downward tax notch (a discrete increase in the average net-of-tax rate). Furthermore, the equation implies that $\lim_{\varepsilon\to 0+} \frac{z^*}{\hat{z}} = 1$. Hence, there is no bunching when labor supply is perfectly inelastic, which rules out the existence of a dominated region as with a downward tax notch.

2.2.3.4 Two-Sided Utility Notch

Suppose finally that there is a two-sided utility notch according to equation (2.8). In the right panel of figure 2.4, An individual with ability $\hat{n}$ is initially located at $\hat{z}$, n_+^* is located at z_+^*, and n_-^* at z_-^*. When the utility notch is introduced, indifference curves become discontinuous at $\hat{z}$. The individual initially located at z_+^* is now indifferent between z_+^* and $\hat{z}$ and the individual initially at z_-^* is now indifferent between z_-^* and $\hat{z}$. These two are the the upper marginal buncher and the lower marginal buncher, respectively. All workers initially between z_-^* and z_+^* bunch at the reference point, while individuals initially to the left of z_-^* or the right of z_+^* do not alter the choice. Hence, bunching at the reference point occurs from both sides, and there is a hole in the density between z_-^* and z_+^*.

Bunching at the reference point is

$$B = \int_{z_-^*}^{z_+^*} h_0(z)dz \approx h_0(\hat{z})(z_+^* - z_-^*)$$

Analogously to equation (2.12), the indifference condition for the lower marginal buncher implies

$$\frac{1}{1+\varepsilon}\frac{z_-^*}{\hat{z}} + \frac{\varepsilon}{1+\varepsilon}\left(\frac{z_-^*}{\hat{z}}\right)^{-\frac{1}{\varepsilon}} = 1 + \frac{\delta_2}{c(\hat{z})}$$

Similarly, the upper marginal buncher is indifferent between the interior solution at z_+^* and the reference point such that

$$\frac{1}{1+\varepsilon}\frac{z_+^*}{\hat{z}} + \frac{\varepsilon}{1+\varepsilon}\left(\frac{z_+^*}{\hat{z}}\right)^{-\frac{1}{\varepsilon}} = 1 + \frac{\delta_2}{c(\hat{z})}$$

Hence, z_+^* and z_-^* are two solutions to the same non-linear equation, where $z_+^* \geq \hat{z}$ and $z_+^* \leq \hat{z}$. Note that, as with the one-sided utility notch $\lim_{\varepsilon\to 0+} \frac{z_-^*}{\hat{z}} = 1$, i.e. there is no bunching from the

left with a zero elasticity. However,

$$\lim_{\varepsilon \to 0+} \frac{z_+^*}{\hat{z}} = 1 + \frac{\delta_2}{c(\hat{z})}$$

Thus, there is bunching from the right even as ε converges to zero. This implies that the two-sided utility notch creates a dominated region to the right of the reference point for a given value of δ_2, a situation similar to an upward tax notch.

2.2.4 Bunching Responses when Reference Points Coincide with Economic Incentives

Certain types of reference points are unrelated to financial incentives, such as round numbers or endowment effects where gains and losses are relative to a status quo. However, a reference point may also coincide with economic incentives - for instance when certain incentives are framed as reference points as argued in chapter 1. This section analyzes such a situation, where there is both a budget constraint kink and a reference point attached to some threshold $\hat{z}$. For the sake of brevity, the analysis focuses on the "utility kink" versions of reference dependence. Appendix section 2.A.2 repeats the analysis for utility notches.

2.2.4.1 Reference Dependence in Consumption

Suppose that a consumption reference point according to equation (2.5) coincides with a budget set kink. In order to compute total bunching, this situation needs to be compared to an initial one without any discontinuity. The left panel of figure 2.A1 illustrates the joint effect. There is an individual whose initial indifference curve is tangent at z_-^* and her new, kinked indifference curve is tangent to the lower part of the kinked budget set at $\hat{z}$. This worker is the lower marginal buncher. In addition, there is an upper marginal buncher exactly like in the case of only a budget set kink. This individual is tangent to the original budget set at z_+^* and tangent to the upper part of the kinked budget set at $\hat{z}$. All individuals to the left of z_-^* work more due to the flatter indifference curves, and all individuals to the right of z^* work less due to the flatter budget line. Hence, there is no hole in the density from either side.

Bunching at the threshold is

$$B = \int_{z_-^*}^{z_+^*} h_0(z)dz \approx h_0(\hat{z})(z_+^* - z_-^*)$$

The lower marginal buncher is determined analogously to equation (2.9)

$$\frac{z_-^*}{\hat{z}} = \left(\frac{1}{1+\lambda_c}\right)^{\varepsilon}$$

The upper marginal buncher, on the other hand, follows the standard formula with a budget set

kink only in equation (2.3)

$$\frac{z_+^*}{\hat{z}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^{\varepsilon}$$

Total bunching is

$$\frac{b}{\hat{z}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^{\varepsilon} - \left(\frac{1}{1+\lambda_c}\right)^{\varepsilon} \tag{2.13}$$

Hence, bunching has two additive components in this case. All bunching due to the reference point character of the threshold is from below, and occurs independently of the size of the budget set kink. On the other hand, all bunching due to the budget set kink occurs from above and is independent of the strength of the reference point.

2.2.4.2 Reference Dependence in Labor Supply

Suppose now instead that utility is reference-dependent according to equation (2.6). The right panel of figure 2.A1 illustrates the joint effect of the utility kink and the budget set kink. The marginal buncher's initial indifference curve is tangent to the initial budget set at z^* and his new, kinked indifference curve is tangent to upper part of the kinked budget set at $\hat{z}$. Again, all individuals to the right of z^* decrease their labor supply due to the joint effect of the utility kink and the budget set kink, so that there is no hole in the density.

The two tangency conditions for the marginal buncher imply $z^* = n^*(1-\tau)^{\varepsilon}$ and $\hat{z} = n^*(1-\tau-\Delta\tau-\lambda_l)^{\varepsilon}$. Hence

$$\frac{z^*}{\hat{z}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau-\lambda_l}\right)^{\varepsilon} \tag{2.14}$$

or

$$\frac{b}{\hat{z}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau-\lambda_l}\right)^{\varepsilon} - 1 \tag{2.15}$$

The equations show that the budget set kink and the utility kink are "perfect substitutes" in terms of producing bunching. Hence, the additional effect of a reference point on bunching at an existing budget set kink is as if the budget set was made more discontinuous. Indeed, when comparing equation (2.15) to equation (2.11), the response $z^*/\hat{z}$ increases by a factor $\left(\frac{1-\tau-\Delta\tau}{1-\tau-\Delta\tau-\lambda_l}\right)^{\varepsilon}$. Hence, the additional bunching from the reference point scales bunching from the budget set kink by a factor that relates the size additional utility kink to the existing post-kink slope of the budget set.

2.2.5 Extensions

Heterogeneous Parameters. The preceding analysis assumes homogenous preferences. However, parameter heterogeneity can be incorporated into the bunching approach. Kleven (2016) shows that in the presence of heterogeneous elasticities bunching at a pure budget set kink can be related to a local average elasticity. Consider a joint distribution $\hat{f}(n,\varepsilon)$ and a joint counterfactual density of labor supply $\tilde{h}_0(z,\varepsilon)$, such that $h_0(z) = \int_\varepsilon \tilde{h}_0(z,\varepsilon)d\varepsilon$. Denoting by Δz_ε^* the response of the

marginal buncher at ε, total bunching can be written as

$$B = \int_{\varepsilon} \int_{\hat{z}}^{z^*_\varepsilon} \tilde{h}_0(z, \varepsilon)\, dz\, d\varepsilon \approx h_0(\hat{z}) E[\Delta z^*_\varepsilon]$$

where the approximate equality holds if $\tilde{h}_0(z, \varepsilon)$ is constant on $[\hat{z}, z^*_\varepsilon]$ for each ε. Hence, z^* can be replaced by $E[\Delta z^*_\varepsilon]$ in equation (2.3) to account for the local average response.

Similarly, joint distributions of $(n, \varepsilon, \lambda_c)$ or $(n, \varepsilon, \lambda_l)$ can be incorporated into the bunching quantities leading to equations (2.9), (2.11) and (2.24). For instance, with heterogeneity in reference dependence in consumption,

$$B = \int_{\lambda_c} \int_{\varepsilon} \int_{\hat{z}}^{z^*_{\varepsilon,\lambda_c}} \tilde{h}_0(z, \varepsilon, \lambda_c)\, dz\, d\varepsilon\, d\lambda_c \approx h_0(\hat{z}) E[\Delta z^*_{\varepsilon,\lambda_c}]$$

where $\tilde{h}_0(z, \varepsilon, \lambda_c)$ is the counterfactual and $\Delta z^*_{\varepsilon,\lambda_c}$ is the response of the marginal buncher at (ε, λ_c). The approximate inequality holds if $\tilde{h}_0(z, \varepsilon, \lambda_c)$ is constant on $[\hat{z}, z^*_{\varepsilon,\lambda_c}]$ for each (ε, λ_c). Thus, equation (2.9) is identified off the average response $E[\Delta z^*_{\varepsilon,\lambda_c}]$.

Income/Wealth Effects. The standard bunching formula (2.3) applies to small kinks where income effects are small (Saez 2010). Equivalently, the formula can be derived from a quasi-linear utility function as above. For larger kinks, however, there may be income effects arising from the change in the implicit net wage. Kleven (2016) argues that in this case, bunching recovers a weighted average between a compensated and an uncompensated elasticity. In other words, if one views the bunching elasticity as an estimator of a compensated elasticity, it is downward biased towards the uncompensated elasticity (assuming leisure is a normal good). The intuition behind this result is that income effects attenuate responses to price changes, since they work in the direction opposite to the substitution effect.

A similar intuition applies to bunching in response to reference points: The presence of income or wealth effects attenuate the response of the marginal buncher. For instance, the marginal buncher responding to a consumption reference point by increasing their labor supply described by equation (2.9), is willing to decrease labor supply by less if the marginal utility of the additional consumption available at higher labor supply is lower. In other words, with income effects, the bunching equations (2.3), (2.9), (2.11) and (2.24) overstate the response at given parameter values. Therefore, estimated parameters can be interpreted as lower bounds on the "compensated" ε, λ_c and λ_l in the presence of income effects.

2.3 Estimation

The conceptual framework allows for identification of reference dependence parameters based on observed bunching. This section first discusses identification assumptions and illustrates structural estimation methods for some cases of interest. In addition, a reduced-form decomposition method is

proposed that allows for quantifying the importance of reference dependence as a source of bunching without parametrically specifying the type of reference point.

2.3.1 Identification

In order to estimate the underlying utility parameters, the first step is to quantify bunching responses at the reference point and possibly at other thresholds. Following the standard methods laid out in section 1.3 of chapter 1, the bunching mass B can be measured as the observed density spike at a threshold $\hat{z}$, assuming that the density would have been smooth in the absence of any discontinuity in preferences or incentives. In terms of identification, bunching must be estimated at at least as many thresholds as parameters are to be estimated.[8] For instance, to identify the two parameters ε and λ_c in equation (2.13), bunching must be observed at at least two thresholds that vary in the financial incentive and/or the presence of a reference point.[9]

The second step of the estimation is to calculate parameters based on bunching at the available thresholds. Indexing $i = 1, ..., n$ the discontinuities available for estimation, and k the number of parameters to be estimated, parameters are "just identified" if $n = k$. As in chapter 1, bunching at discontinuity B_i can be written as a function of the elasticity, a vector of additional parameters $\boldsymbol{\omega}$, an indicator for the presence of a reference point D_i and additional variables at the threshold $\mathbf{x}_i$.

$$B_i = B(\varepsilon, \boldsymbol{\omega}(D_i), \mathbf{x}_i)$$

yields a system of equations that can be solved for ε and $\boldsymbol{\omega}$. The key identification assumption is that bunching at each discontinuity is generated by the same underlying parameters ε and $\boldsymbol{\omega}$. For instance, equations (2.4) and (2.10) can be solved to yield implied values of ε and λ_c when bunching is observed at two thresholds.

In some cases, parameters are "over-identified", i.e. $n > k$. Bunching at threshold i can be written as

$$B_i = B(\varepsilon, \boldsymbol{\omega}(D_i), \mathbf{x}_i) + \nu_i$$

including an error term ν_i. Instead of finding an exact solution to the system of equations, parameters can be estimated, for instance by minimizing squared deviations. This is the case for the reduced-form estimation in chapter 1, where there are many more discontinuities than parameters. The identification assumption becomes that of chapter 1: Parameters should not be correlated with D_i or $\mathbf{x}_i$ across discontinuities.

[8] Other applications of this idea include Gelber et al. (2017) who use two bunching observations to solve for two parameters, an elasticity and an adjustment cost parameter.

[9] For simplicity, most of the arguments in this section are presented using the "kink in utility from consumption" version of reference dependence. Assuming other types of reference dependence, ε and λ_l, δ or δ_2 can be estimated analogously.

2.3.2 Structural Estimation

In the following, I discuss how the structural bunching equations from the conceptual framework can be used to estimate reference dependence parameters. Three potential applications are considered. First, a pure reference point can be estimated if the type of reference dependence is known. Second, reference dependence vs. economic incentives can be quantified as sources of bunching. Third, different types of reference dependence can be empirically distinguished when additional moments of the distribution are available.

2.3.2.1 Estimation of a Pure Reference Point

First, suppose a threshold serves as a (suspected) reference point, and there is no change in incentives at this threshold. In this case, finding significant bunching at the threshold provides a direct test of whether there is indeed a reference point. In order to estimate utility parameters, it is necessary to assume a particular type of reference dependence. For instance, assuming there is loss aversion in consumption, equation (2.10) can be used to estimate λ_c. A caveat is that the elasticity has to be known or estimated as well, since responses also depend on ε. If ε is known from other sources, equation (2.10) can be directly used to estimate λ_c from observed bunching at the reference point.

Aa arguably more compelling alternative is to estimate both parameters from the same context. Denoting by b_1 observed bunching at the reference point $\hat{z_1}$ and b_2 bunching at a pure budget set kink $\Delta\tau/(1-\tau)$ at $\hat{z_2}$, the two parameters can be calculated by solving

$$\frac{b_1}{\hat{z_1}} = 1 - \left(\frac{1}{1+\lambda_c}\right)^{\varepsilon}$$

$$\frac{b_2}{\hat{z_2}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^{\varepsilon} - 1$$

The underlying identification assumption is that ε does not differ across the two thresholds. Standard errors on the parameter estimates can be obtained via the usual bootstrap method.

2.3.2.2 Reference Dependence vs. Economic Incentives

A second case of interest may be when one observes bunching at a threshold that features economic incentives, but it may also serve as a reference point exacerbating bunching. In this case, the methods proposed in this chapter can be used to empirically distinguish between reference dependence and incentives as sources of bunching. This may be useful for two reasons. First, reference dependence itself may be the object of interest, but it is confounded by incentives. Second, standard bunching applications aiming at estimating a "true" labor supply elasticity need to consider that certain thresholds may serve as reference points, and estimate the elasticity "net" of this effect.

To estimate both parameters of interest, it is necessary to observe bunching at at least two thresholds that vary in the underlying economic incentives and/or the presence of a reference

point. For instance, one could use the threshold $\hat{z}_1$ where both a budget set kink and a consumption reference point is present, and a pure budget set kink at threshold $\hat{z}_2$. Then, ε and λ_c can be calculated based on

$$\frac{b_1}{\hat{z}_1} = \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^{\varepsilon} - \left(\frac{1}{1+\lambda_c}\right)^{\varepsilon}$$

$$\frac{b_2}{\hat{z}_2} = \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^{\varepsilon} - 1$$

where again the key assumption is that ε is constant across $\hat{z}_1$ and $\hat{z}_2$.

2.3.2.3 Distinguishing Different Types of Reference Dependence

Finally, it may be desirable to empirically distinguish between different types of reference dependence. The difficulty is that observing bunching alone is not sufficient to distinguish between the types, as they all lead to sharp bunching at the threshold. Thus, additional bunching moments are needed to make progress. One possibility is to use the fact that the different types of reference dependence have different implications for the shape of the distribution around the reference point. Suppose for instance that there is a threshold $\hat{z}$ that serves as a reference point but it is unkown whether it is in terms of consumption or labor supply. It may be possible to use the density around the threshold to infer whether bunching originates from below vs. above, which in turn implies which type is at work.

One way to implement this idea is by estimating the share of bunching from below α. The excess mass originating from below $b_- = \alpha b$ can be written as

$$\frac{b_-}{\hat{z}} = 1 - \left(\frac{1}{1+\lambda_c}\right)^{\varepsilon}$$

and bunching from above $b_+ = (1-\alpha)b$ is

$$\frac{b_+}{\hat{z}} = \left(\frac{1-\tau}{1-\tau-\lambda_l}\right)^{\varepsilon} - 1$$

The two equations can be solved for λ_c and λ_l, given that ε is known or estimated from another discontinuity. Section 2.4.3.2 provides an application of this idea, using the observed density around thresholds to distinguish between reference dependence in consumption vs. labor supply.

2.3.3 Reduced-Form Estimation

The approaches discussed above yield estimates of reference dependence and its underlying parameters, given an assumption on the type of reference dependence. One may be able to empirically distinguish different types, but in some instances an attractive alternative may be an approach where reference point effects can be quantified without specifying the type of reference dependence.

The basis for such an approach is the observation that bunching at a budget set kink is approximately linear in the financial incentive at the threshold (Saez 2010, Kleven 2016). Note that

$\log(z^*/\hat{z}) \approx \Delta z^*/\hat{z}$, and $\log(1-\tau-\Delta\tau)/(1-\tau)) \approx -\Delta\tau/(1-\tau)$ when $\Delta\tau$ is small and hence $\Delta z^* = z^* - \hat{z}$ is small. Take logs on equation (2.3)

$$\log \frac{z^*}{\hat{z}} = -\varepsilon \log\left(1 - \frac{\Delta\tau}{1-\tau}\right)$$

which implies

$$\frac{b}{\hat{z}} \approx \varepsilon \frac{\Delta\tau}{1-\tau} \tag{2.16}$$

Thus, bunching is approximately linear in the kink size. This idea can be extended to the presence of a reference point at the threshold. For instance, if there is a budget set kink and a consumption reference point at the threshold, and when $\Delta\tau$ is small and hence Δz_+^* is small, equation (2.13) implies

$$\frac{b}{\hat{z}} \approx \varepsilon \frac{\Delta\tau}{1-\tau} + 1 - \left(\frac{1}{1+\lambda_c}\right)^{\varepsilon}$$

Hence, bunching is approximately linear in kink size and additive in a component related to the financial incentive and one related to reference dependence. Appendix 2.A.3 shows that this insight also applies to the other types of reference dependence.[10]

This motivates a simple linear, additive decomposition of observed bunching across thresholds.

$$\frac{b_i}{\hat{z}_i} = \varepsilon \frac{\Delta\tau_i}{1-\tau_i} + \beta D_i \tag{2.17}$$

where i indexes thresholds and D_i is an indicator for a reference point being attached to threshold i. The parameter β can be interpreted as the additional bunching due to a reference point of any type. The identification assumption is the same as for structural estimation: ϵ should not differ across thresholds, such that additional bunching can be attributed to reference dependence.

Moreover, the conceptual framework implies a mapping between β and underlying utility parameters under each type of reference dependence. For instance, in the case of a kink in utility from consumption, $\beta = 1 - \left(\frac{1}{1+\lambda_c}\right)^{\varepsilon}$ or $\beta \approx \varepsilon\lambda_c$ for small λ_c. Hence, the specification used in section 1.4 of chapter 1 can be viewed as a reduced-form decomposition of the effects of financial incentives vs. reference points under the model presented in this chapter.

2.4 Applications: Reference Dependence in Retirement Behavior

2.4.1 Conceptual Framework

The remainder of this chapter presents applications of the methods presented above to the context of retirement behavior. To begin with, this section considers a simple static model of retirement decisions in order to derive bunching equations analogous to those from sections 2.2.3 and 2.2.4.

[10] In the case of a kink in utility from consumption and a one-sided utility notch, the approximation holds if $\Delta\tau$ is small. In the case of a kink in disutility from work, $\Delta\tau + \lambda_l$ is required to be small. The only exception is the two-sided utility notch, where the additional bunching due to a reference point is not fully independent of the kink size, which may introduce some bias into an estimation using equation (2.17).

The static model considered in this section corresponds to the "lifetime budget constraint" model of retirement suggested by Burtless (1986). Similar static models are used in recent retirement bunching applications such as Brown (2013) and Manoli and Weber (2016*b*). Section 2.4.1.3 provides an outlook on the relationship with a full dynamic model.

2.4.1.1 Basic Setup

Workers maximize lifetime utility

$$U = u(C) - v(R, n)$$

where C is lifetime consumption, R is the worker's retirement age relative to a career starting age normalized to 0, and n is a parameter capturing earnings ability at old age. In the retirement context, a stylized lifetime budget constraint can be written as the sum of wage earnings until retirement subject to payroll tax $\tilde{\tau}$, and pension benefits $B(R)$ that are paid from R to time of death T:

$$C = w(1 - \tilde{\tau})R + B(R)(T - R) \tag{2.18}$$

The *implicit* net-of-tax rate $1 - \tau$ is defined via

$$\frac{dC}{dR} = w(1 - \tau)$$

where $\frac{dC}{dR}$ is the marginal gain in consumption from postponing retirement by one period, or the implicit net wage. the implicit net wage. Consider first the case of a linear budget constraint $C = w(1 - \tau)R$, and assume as in section 2.2.1, that utility is quasi-linear in consumption C and iso-elastic in labor supply R. Then workers' utility maximization yields $R = n\left[w(1 - \tau)\right]^{\varepsilon}$, where ε is the elasticity of the retirement age with respect to the implicit net-of-tax rate. If the distribution of ability $F(n)$ is smooth, this implies a smooth distribution of retirement ages with density $h_0(R)$.

2.4.1.2 Retirement Bunching Responses

Now the model can be used analogously to the standard labor supply model in order to quantify retirement bunching. There is bunching in the running variable R at discontinuities in marginal incentives captured by the implicit net-of-tax rate as well as bunching at reference points.

Bunching at a Budget Set Kink. Analogously to figure 2.1, there is bunching at a lifetime budget constraint kink where the marginal implicit tax rate increases by $\Delta\tau$ at some retirement age threshold $\hat{R}$. There is a marginal buncher with ability n^* whose indifference curve is tangent to the initial budget set at R^* and to the upper part of the new budget set at $\hat{R}$. All workers initially retiring between $\hat{R}$ and R^* bunch at the kink, while all individuals initially to the left of the kink leave their retirement age unchanged and all individuals initially to the right of R^* stay above the kink.

Total retirement bunching is

$$B = \int_{\hat{R}}^{R^*} h_0(R)dR \approx h_0(\hat{R})(R^* - \hat{R})$$

The two tangency conditions for the marginal buncher imply $R^* = n^*[w(1-\tau)]^\varepsilon$ and $\hat{R} = n^*[w(1-\tau-\Delta\tau)]^\varepsilon$ and thus

$$\frac{b}{\hat{R}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^\varepsilon - 1 \tag{2.19}$$

In addition to the preferences described above, workers may evaluate outcomes relative a threshold retirement age $\hat{R}$. Previous work on the retirement context has pointed out that such reference dependence may be present in both utility from consumption and disutility from work (e.g. Behaghel and Blau 2012, Merkle et al. 2017). Both are considered here and modeled as a kink in the utility function.[11]

Bunching at a Consumption Reference Point. Consider first a change in utility from consumption at the reference point:

$$U = u(C) - v(R,n) - \mathbb{1}(C \leq \hat{C}) \cdot \lambda_c(\hat{C} - C) \tag{2.20}$$

The last term in equation (2.20) introduces a discrete jump in the marginal utility from consumption at the reference level $\hat{C} = C(\hat{R})$. For example, such loss aversion in consumption may arise due to gain-loss framing of pension adjustments around age thresholds or expectations set by benefit or consumption levels linked to "full" or "normal" retirement ages.

Analogously to section 2.2.3.1, it can be shown that retirement bunching at the reference age is

$$\frac{b}{\hat{R}} = \left(\frac{1}{1+\lambda_c}\right)^\varepsilon - 1 \tag{2.21}$$

Bunching at a Labor Supply Reference Point. The second type of reference dependence is captured by

$$U = u(C) - v(R,n) - \mathbb{1}(R \geq \hat{R}) \cdot \widetilde{\lambda}_l(R - \hat{R}) \tag{2.22}$$

where marginal disutility from labor supply jumps by $\widetilde{\lambda}_l > 0$ at $\hat{R}$. Higher disutility from continuing work after the threshold is consistent with an interpretation where workers perceive continuing work after the threshold as a loss in lifetime leisure, for instance relative to "full" or "normal" retirement ages, or other round ages.

As in section 2.2.3.2, retirement bunching at the reference point is

$$\frac{b}{\hat{R}} = \left(\frac{1-\tau}{1-\tau-\lambda_l}\right)^\varepsilon - 1 \tag{2.23}$$

[11] The "loss aversion" functional form of reference dependence is most commonly used in the literature (e.g. Rees-Jones 2018 and DellaVigna et al. 2018).

where $\lambda_l = \widetilde{\lambda}_l / w$ is the parameter normalized by the wage.

2.4.1.3 Extension: Dynamics

Bunching applications typically use static models, but retirement decisions are dynamic problems and often modeled as such in other contexts. Appendix 2.A.4 sets out a dynamic life-cycle model of retirement, and shows how it linked to the static model considered in this section. In particular, the static model can be viewed as a reduced form of the full dynamic model under two assumptions: First, all uncertainty in earnings capacity is realized at the "beginning" of old age when the retirement age is decided, and second, there are no liquidity constraints.

This chapter focuses on the static model for several reasons. First, simple and transparent bunching equations can be derived from the static version. Second, the static model is directly analogous to a standard labor supply model and thus results can be easily compared to those from existing bunching models. Third, the sharp bunching responses documented in this chapter 1 may indicate that dynamic uncertainty does not play a large role for the response of retirement to different discontinuities. Fourth, as long as uncertainty attenuates responses at statutory ages and pure financial incentives in the same way, the relative magnitude of the parameters of interest can still be identified.

2.4.2 Application 1: Bunching at Round Retirement Ages

The first application analyzes a type of "pure" reference point in the context of retirement decisions, namely round retirement ages. Bunching at round numbers has been observed in different contexts including taxable income (Kleven and Waseem 2013) and house prices (Best and Kleven 2018) and has been attributed to reference dependence (Kleven 2016). In typical bunching applications, round-number effects are treated as a confounder of the effect of incentives and are netted out of the bunching analysis by allowing for round-number effects in the counterfactual.[12] In contrast, this section aims at estimating the reference dependence parameters implied by observed bunching at round retirement ages.

Specifically, round ages other than those framed as statutory retirement ages are considered. Other round ages plausibly represent pure reference points since they do not entail any change in financial incentives, nor any regulation that would facilitate firm responses. The most plausible type of reference dependence at round ages is arguably in terms of the retirement age itself (or equivalently lifetime leisure), where they may serve as goals or expectations that workers strive for. It is unlikely that round retirement ages serve as consumption reference points since, in contrast to statutory retirement ages, the consumption or benefit level associated with them is not framed in a particular way. Assuming that they serve as reference points in terms of labor supply, bunching at a round age is given by equation (2.23). Hence, the parameter of interest λ_l can be estimated from observed bunching at the round age, given an estimate of ε from the same context.

[12] An exception is Allen et al. (2016) where the object of interest is bunching at round marathon finishing times.

Panel A of figure 2.6 shows the retirement age distribution among workers born between 1933 and 1948, that is the full sample from chapter 1.[13] Vertical red lines indicate the location of the main statutory retirement ages where very large bunching occurs. However, bunching at other round ages is also clearly visible, in particular at ages 61, 62 and 64. Panels B to D zoom into the retirement age distribution around these ages, where all birth cohorts are again pooled. The connected black dots show the actual distribution and the fitted red curve is a counterfactual estimated as a 7th-degree polynomial. Although the density has different shapes around the three ages, there is clear and sharp bunching at each round age. Table 2.1 shows bunching estimates and implied parameters. The excess mass is 0.61 at age 61, 0.89 at age 62 and 1.12 at age 64. An advantage of pooling all the full sample is that sample sizes are large: There are around 1 to 1.5 million individuals around each threshold. Bootstrapped standard errors are small and all estimates are highly significant. The magnitudes of responses are substantially smaller than bunching observed at statutory retirement ages. However, it is striking that the bunching mass is similar or even larger than that at sizeable financial incentives in figure 1.4 of chapter 1.

The implied values of λ_l, the parameter governing the kink in marginal disutility from work, can be calculated based on equation (2.23). The results are included in the respective panel of figure 2.6 for each round age. The estimate of λ_l is 0.05 at age 61 and around 0.07 at ages 62 and 64, and all estimates are significantly different from zero. An intuitive interpretation of the magnitudes arises from to the analogy of a kink in disutility from work and a budget set kink discussed in section 2.2.4.2. The estimates imply that the effect of round ages as reference points corresponds to a kink in the implicit tax rate $\Delta\tau/(1-\tau)$ between 11% and 17%.

2.4.3 Application 2: Bunching at Statutory Retirement Ages

The second application estimates reference dependence at statutory retirement ages analyzed in chapter 1. Adding structural estimation to the reduced-form estimation of chapter 1 serves several purposes. First, more externally valid utility parameters governing reference dependence can be recovered. Second, while the reduced-form estimation is a linear approximation, the structural approach can estimate an exact relationship under the parametric assumptions from section 2.4.1.[14] Third, bunching can be simulated under some counterfactual assumptions regarding parameters and policy variables.

This section considers statutory ages as potential reference points both in terms of consumption and in terms of labor supply. This is motivated by the fact that the framing discussed in chapter 1 is consistent with loss aversion in consumption and labor supply. Moreover, there is independent empirical support for both types of reference dependence. For example Merkle et al. (2017) show that gain/loss framing of pension benefits around statutory ages induces loss averse behavior in survey respondents, and Shoven et al. (2017) find that statutory ages may be perceived as norms

[13] See section 1.2 of chapter 1 for a detailed description of the empirical setting and data.

[14] Based on the arguments in section 2.3.3, the reduced-form specification from chapter 1 can be interpreted as a linear approximation of the structural equation (2.24) for small $\Delta\tau$ and λ_l. Appendix 2.A.5.4 shows in more detail how the reduced-form coefficients can be linked to the model parameters.

in terms of the retirement date itself. In the following, a bunching equation nesting both types is derived, and estimation results are presented from focusing on either type, as well as estimating both jointly.

At statutory age thresholds, a potential reference point coincides with a change in incentives. Appendix figure 2.A2 illustrates the joint effect on retirement bunching. There is an individual whose initial indifference curve is tangent at R^*_- and her new, kinked indifference curve is tangent to the lower part of the kinked budget set at $\hat{R}$. This worker is the *lower marginal buncher* who bunches due to the consumption reference point. In addition, individuals bunch from the right due to the combination of the budget set kink and reference dependence in disutility from work. There is an *upper marginal buncher* whose original indifference curve is tangent to the original budget set at R^*_+ and whose kinked indifference curve is tangent to the upper part of the kinked budget set at $\hat{R}$. Hence, all individuals initially located between R^*_- and R^*_+ bunch, and there is no hole in the density because workers to the left of R^*_- also delay retirement due to their flatter indifference curves, and workers to the right of R^*_+ retire earlier due to the flatter budget line and the steeper indifference curves.

Bunching at the threshold is

$$B = \int_{\hat{R}}^{R^*} h_0(R)dR \approx h_0(\hat{R})(R^*_+ - R^*_-)$$

The lower marginal buncher is determined analogously to equation (2.21)

$$\frac{R^*_-}{\hat{R}} = \left(\frac{1}{1+\lambda_c}\right)^{\varepsilon}$$

The two tangency conditions for the upper marginal buncher imply $R^*_+ = n^*_+[w(1-\tau)]^{\varepsilon}$ and $\hat{R} = n^*_+[w(1-\tau-\Delta\tau-\lambda_l)]^{\varepsilon}$. Hence

$$\frac{R^*_+}{\hat{R}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau-\lambda_l}\right)^{\varepsilon}$$

Total bunching at the combined threshold can be expressed as

$$\frac{b}{\hat{R}} = \left[\left(\frac{1-\tau}{1-\tau-\Delta\tau-\lambda_l}\right)^{\varepsilon} - 1\right] + \left[1 - \left(\frac{1}{1+\lambda_c}\right)^{\varepsilon}\right] \tag{2.24}$$

Hence, bunching is additive in two components. The first term in equation (2.24) captures bunching from the right due to the combination of the budget set kink and reference-dependent disutility from work. The second term in the equation captures bunching from the left due to reference-dependent utility from consumption.

2.4.3.1 Basic Estimation: Upper Bounds

Taking the model to the discontinuity-level data in the bunching sample from chapter 1, equations (2.19) and (2.24) imply that bunching at discontinuity i is given by

$$\frac{b_i}{\hat{R}_i} = \left[\left(\frac{1-\tau_i}{1-\tau_i-\Delta\tau_i-\sum_s \lambda_l^s D_i^s}\right)^{\varepsilon} - 1\right] + \left[1 - \left(\frac{1}{1+\sum_s \lambda_c^s D_i^s}\right)^{\varepsilon}\right] + \xi_i \tag{2.25}$$

where ξ_i is an error term. Recall that the discontinuities in the bunching sample vary in the kink size $\Delta\tau_i$ and in the presence of statutory ages of types $s \in (ERA, FRA, NRA)$. However, a key issue with the estimation is that λ_c^s and λ_l^s cannot be separately identified for a given statutory age type based solely on equation (2.25). Intuitively, both reference dependence in consumption and leisure lead to sharp bunching at the threshold $\hat{R}$ such that a given amount of excess mass could be rationalized by a range of combinations of λ_c^s and λ_l^s. In other words, one bunching equation per discontinuity is not sufficient to separately identify the two parameters. A first solution to this problem is to estimate upper bounds on both types of reference dependence, assuming that all reference point bunching is only driven by loss aversion in consumption *or* loss aversion in leisure. This can be done by restricting λ_l^s and λ_c^s to zero, respectively. The method is based on the estimation strategy from section 2.3.2.2, where one type of reference dependence can be distinguished from the effect of financial incentives.

Appendix table 2.A2 presents the full set of non-linear least squares estimates based on equation (2.25). Column (1) reports upper bounds on λ_c^s obtained from estimating the model with all λ_l^s set to zero. The elasticity of 0.15 is precisely estimated and very similar to the reduced-form results. Upper bounds on the λ_c^s parameters are positive and highly significant. While λ_c^{ERA} can be bounded around 5, λ_c^{FRA} is larger and the estimation does not seem able to bound λ_c^{NRA} in an informative way. Column (2) reports results from the reverse exercise, setting all λ_c^s to zero. Upper bounds on the λ_l^s parameters are also positive and highly significant, with magnitudes varying between 0.1 and 0.4. As expected, the implied strength of reference dependence at statutory ages is substantially larger than that at other round ages.

2.4.3.2 From Parameter Ranges to Point Estimates

Next, a range of possible λ_c-λ_l combinations can be estimated. To see this, it is useful to recall the discussion in section 2.3.2.3 on how two types of reference dependence can be distinguished when the shares of bunching from the two sides are known. Bunching from the right at discontinuity i is

$$\frac{b_i^+}{\hat{R}_i} = \left[\left(\frac{1-\tau_i}{1-\tau_i-\Delta\tau_i-\sum_s \lambda_l^s D_i^s}\right)^{\varepsilon} - 1\right] + \xi_i \tag{2.26}$$

and bunching from the left is

$$\frac{b_i^-}{\hat{R}_i} = \left[1 - \left(\frac{1}{1+\sum_s \lambda_c^s D_i^s}\right)^{\varepsilon}\right] + \xi_i \tag{2.27}$$

where $b_i = b_i^+ + b_i^-$. Now the full range of parameter combinations consistent with equation (2.25) can be estimated by varying the bunching shares from each side between their minimum and maximum possible values. Denoting α_i the share of excess mass originating from the right, this share ranges between a minimum $\hat{\alpha}_i$ and 1. The minimum right bunching share $\hat{\alpha}_i$ is given by the fraction of bunching that would persist of workers only bunch due to the budget constraint kink. Appendix 2.A.5 provides for further details of the implementation of this and the other estimation strategies. Figure 2.7 plots all possible combinations of λ_c^s and λ_l^s for each statutury age type, obtained from a simulation where the bunching share from the right at each discontinuity is gradually moved from its minimum to 1 as described above. The negative slope of the relationship illustrates the intuition that the two types of reference dependence are substitutes in terms of rationalizing observed excess mass. However, the simulated ranges are wide, including large positive values of λ_c^{NRA}, and negative values of λ_l^{FRA}, which indicates the need for further narrowing down the parameter estimates.

In order to make progress and obtain point estimates of both λ_c^s and λ_l^s, i.e. to quantify the relative importance of reference dependence in consumption and leisure, empirical estimates of α_i are needed. Under some additional assumptions, the empirical density around thresholds can be used to this avail. To illustrate the idea, figure 2.8 plots pooled distributions around the different statutory age types, omitting the bunching region. There seems to be some "missing" density to the left of the ERA, and a clear drop in the density to the right of the NRA, while the situation is less clear around the FRA. As a first attempt to use this information, column (3) of table 2.A2 presents parameter estimates under an ad-hoc assumption based on this visual inspection, where all reference point bunching is assumed to be due to λ_c^s at the ERA and FRA and due to λ_l^s at the NRA. Results remain highly significant and suggest that this combination of different types of reference dependence produces tighter bounds on the underlying parameters than previous specifications.

Finally, bunching shares from both sides can be computed based on empirical estimates of the corresponding density shifts. Intuitively, the counterfactual density is assumed to be continuous around the threshold, and the relative number of bunchers from the left and from the right is inferred from the vertical difference between the counterfactual density and the one actually observed on both sides of the threshold. In addition, an augmented method measures the horizontal density shift on both sides, taking into account that observed vertical shifts may be confounded by different gradients of the density. In general, this estimation requires a stronger assumption about the true relative density shifts being reasonably well approximated by locally observed relative shifts.

Panel A of table 2.2 shows estimates of relative bunching from the two sides and appendix table 2.A1 presents additional robustness checks. Between 64% and 87% of ERA bunching is estimated to originate from the left, and with 80% to 99% a large majority of NRA bunching is from the right. 56% to 68% of FRA bunching originates from the right. Taking into account the density gradient seems to make little difference to the estimated shares. Columns (4) and (5) of table 2.A2 present parameter estimates based on these estimates. λ_c^{ERA} remains at a magnitude between 3 and 4, indicating that reference dependence in consumption is the most important source of bunching at the ERA. Conversely, most bunching at the NRA seems to be due to reference dependence in leisure

as λ_l^{NRA} is virtually the same as its upper bound of 0.38. Bunching at the FRA, on the other hand, seems to be driven by a mixture of the two types of reference dependence. The labeled dots in figure 2.7 mark the parameter combinations selected by the bunching share estimation, confirming that the estimation successfully bounds parameters in a narrower, positive range.

Overall, the estimated utility parameters are large in magnitude. For instance, $\lambda_c^{ERA} = 4$ implies that marginal utility from consumption just before the ERA is five times larger than just after the ERA. The λ_l^s have an additional, natural interpretation: The estimated magnitudes correspond to the increase in the implicit tax rate at the kink that would produce the same amount of bunching. For example, $\lambda_l^{NRA} = 0.38$ corresponds to the effect of a 38 percentage point kink at the NRA. Moreover, there are interesting differences in the role of reference dependence in consumption and leisure at different types of statutory ages. The estimation suggests that loss aversion in lifetime consumption plays an important role at the earlier statutory ages, which induces workers to postpone retirement until the ERA in particular. The FRA seems to set expectations regarding both consumption and lifetime leisure, implying bunching from both sides. Bunching at the NRA, on the other hand, is mostly due to strongly increased disutility from postponing retirement beyond this age.

2.4.3.3 Implications and Counterfactual Simulations

In contrast to round numbers, statutory ages are framed by policy and they can be used for potential pension reforms. This section first uses the structural estimates as the basis for some metrics to illustrate the magnitude of responses to statutory ages, and then simulates the effects of a range of counterfactual assumptions on parameters and policy variables.

Metrics Related to Bunching. First, table 2.2, panel B presents some metrics implied by the structural estimates. The figures are based on the preferred specification using gradient-corrected left/right bunching share estimates. The implied retirement age responses are large: the right marginal buncher moves their retirement date forward by around 5 months at the ERA, 11 months at the FRA and 29 months at the FRA. However, the ERA and FRA also induce some individuals to postpone retirement by between 9 and 10 months. These responses imply substantial changes in lifetime consumption. Workers moving retirement forward towards statutory ages incur lifetime consumption losses of around €5000 at the ERA, €11,000 at the FRA and €37,000 at the NRA. These correspond to 0.5%, 1% and 3.6% of estimated lifetime consumption of the affected workers, respectively. Consumption gains of workers postponing retirement towards the ERA and FRA are between €12,500 and €14,000 or just above 1% of lifetime consumption. Finally, hypothetical financial incentives that would imply the same bunching responses are very large. Kink sizes between 2 and 3.6, i.e. changes in the implicit net-of-tax rates of at least 200%, would be necessary to induce spikes in retirement similar to the ones observed at statutory ages.

Financial Incentives vs. Reference Dependence. In terms of counterfactual simulation, a natural first question may be how much bunching at statutory age retirements would prevail in

the absence of reference points. Panel A of table 2.3 shows results from a simulation of the job exit age distribution under this assumption. Column (1) reports the actually observed fraction of job exits and average excess mass at discontinuities, while column (2) shows simulated figures based on the structural estimates. Over the entire sample period, 29% of workers actually retire at statutory ages. In the counterfactual scenario with λ_c^s and λ_l^s set to zero, this fraction is estimated to decrease to only 6 percentage points. Hence, the estimation attributes around 80% of actual retirements at statutory ages to reference dependence. The average excess mass across all discontinuities is predicted to decrease even more dramatically from 19.2 to 0.99. This sharp drop is partly a consequence of the simulation predicting negative excess mass (holes) in the job exit age distribution at non-convex kinks. When attention is restricted only to convex statutory age kinks, the average excess mass is predicted to decrease from 14.7 to between 1.71. In addition, columns (3) and (4) show results from analogous simulations based on the lower bound and the upper bound of the reduced-form elasticity estimates, respectively. The fraction of retirements at statutory ages is estimated to decrease to between 5 and 10 percentage points. In other words, at most 35% of statutory age retirements are attributed to financial incentives. Figure 2.9, panel A shows a graph of the simulated job exit age distribution under the central scenario. As expected, the spikes at the main statutory ages are greatly reduced in magnitude. Moreover, the graph illustrates the predicted un-bunching patterns based on the estimated bunching shares from each side. There is a visible upward shift in the density below age 60, and even more strongly, at job exit ages above 65.

Appendix table 2.A3 relaxes the assumption that all additional effects at statutory ages can be attributed to reference dependence. In particular, it allows for the possibility that firms make use of mandatory retirement clauses linked to NRA to lay off older workers. In column (2), a mandatory retirement effect is introduced by assuming that the NRA is viewed by workers as a reference point equally to the FRA, and the remaining bunching at the NRA is attributed to mandatory retirement. Allowing for this effect in combination with financial incentives for workers increases the explained share of statutory age job exits to 26%. Column (3) aims at estimating a lower bound on the role of reference points by making the extreme assumption that all bunching at NRAs is driven by firm responses. In combination with worker incentives, this would explain 57% of all statutory age job exits, leaving a lower bound of 43% of job exits attributed to reference dependence.

Policy Reforms. Finally, the parameter estimates can be used to simulate the effects of counterfactual policy scenarios. In particular, I focus on two policies that represent realistic options for pension reform. The first reform increases the NRA from 65 to 66, keeping incentives around the NRA constant. The second reform increases rewards for late retirement from the current level of 6%, keeping the NRA at 65. Panel B of figure 2.9 and panel B of table 2.3 summarize the effects. Under the NRA increase, there is un-bunching of the spike at 65, and since most NRA bunching is from the right, the previous bunchers are dispersed mainly above age 65. A large job exit spike emerges at the new NRA of 66. Actual retirement ages increase by around 4 months, and the increase among individuals who retire at 65 and above is 11 months. In the second scenario, the increase in late retirement rewards is calibrated to match the effect on the average retirement age in

the first scenario. In order to get the same effect, an increase in rewards by 75% (from currently 6% p.a. to 10.5% p.a.) would be needed. In the graph, providing more incentives for late retirement leads to a decrease in the excess mass at the NRA by more than half, and the previous bunchers disperse along the density above age 65.

Hence, both types of policies could achieve an increase in actual retirement ages above the NRA. However, the predicted fiscal impact of the two scenarios is very different. The NRA increase has a simulated net fiscal effect of +€731m per year. This is due to the additional contributions of workers who postpone retirement, countered only by a small increase in pension liabilities since no additional pension adjustment has been introduced. On the contrary, the fiscal effect of increased rewards is negative at -€206m. Workers also contribute longer in this scenario, but this is more than offset by the large increase in pensions necessary to induce them to postpone retirement. These results highlight the potential effectiveness of using statutory age thresholds as a policy tool. If the government wants to increase average retirement ages to improve the fiscal balance of the system, increasing statutory ages may be an attractive policy option.

Welfare Effects of Policy Reforms. Whereas simulating the fiscal impact of reforms is relatively straightforward, an important but trickier issue is how to calculate a welfare effect beyond purely fiscal considerations. One reason why it is not obvious how to calculate welfare in the presence of reference dependence is because such an evaluation requires taking a stance on the extent to which workers actually experience reference point effects. Write workers' experienced utility as

$$U = u(C) - v(R, n) - \omega \left[\mathbb{1}(C \leq \hat{C}) \cdot \lambda_c(\hat{C} - C) + \mathbb{1}(R \geq \hat{R}) \cdot \widetilde{\lambda}_l(R - \hat{R}) \right] \tag{2.28}$$

where $\omega \in [0, 1]$ is the weight on the reference dependence component of the utility function. The distinction between decision utility and experienced utility is that according to the model in section 2.4.1, decision utility is as if $\omega = 1$ in the equation above. Hence, workers' decisions fully take account of reference points, but they may experience actual utility effects from deviating from reference points fully, partially, or not at all. In terms of experienced utility, there are two polar cases. $\omega = 0$ represents the case where reference point effects are a pure bias and workers only experience standard disutility from work and utility from consumption. On the other hand, $\omega = 1$ is the case where workers fully experience reference point effects, too. Based on these extreme cases, bounds on welfare effects can be computed.

Panel B of 2.3 shows welfare effects of the simulated reforms. Total welfare is calculated based on equation (2.28) as the sum of workers' utility from consumption, their disutility from work, the utility component related to reference dependence weighted by ω, and the net fiscal effect is added.[15] Worker consumption increases under both scenarios due to the increase in lifetime earnings when postponing retirement. The effect of the increase in financial rewards of around

[15] An implicit assumption behind the calculation is that worker consumption and government revenue carry the same marginal value.

+€1567m in net present value terms is much larger than the +€630m from the NRA increase since workers additionally receive higher pensions under the second reform. Worker welfare, which includes consumption, labor supply and reference point effects, hardly changes under the NRA increase. Moreover, the effect does seem to depend much on the assumption about ω. Intuitively, most workers do not deviate from the reference point both before and after the reform, so the assumption on how much they experience a utility cost of such deviations is not crucial. The increase in financial rewards, on the other hand, has large effect on worker welfare. It increases by €804m under $\omega = 1$, and by €1262m under $\omega = 0$. The difference between the two cases arises because many workers move away from the NRA with stronger financial incentives, and when a positive weight is placed on these reference point effects this causes utility cost. The increase in total welfare is around €800m under the NRA increase, regardless of the assumption on ω. The larger financial rewards, on the other hand, increase total welfare by €598m if $\omega = 1$ and by €1056m if $\omega = 0$.[16] Hence, the NRA increase has larger positive effects on welfare if workers fully experience reference point effects, but the increase in financial rewards is more beneficial if workers do not experience reference point effects. The results illustrate that the assumptions made on the utility function are crucial for evaluating welfare.

2.5 Conclusion

A growing literature uses bunching methods to identify responses to local changes in tax rates and other extrinsic variables. At the same time, the notion of reference dependence has been gaining support with recent empirical evidence from several domains. This chapter aims at linking the two literatures by using bunching methods to quantify reference-dependent preferences. Two empirical applications illustrate the idea. First, bunching at round retirement ages is estimated as an example of pure reference points. Second, responses at statutory retirement ages are estimated, where economic incentives and potential reference points coincide. This context is also used to illustrate how different types of reference dependence can be potentially distinguished.

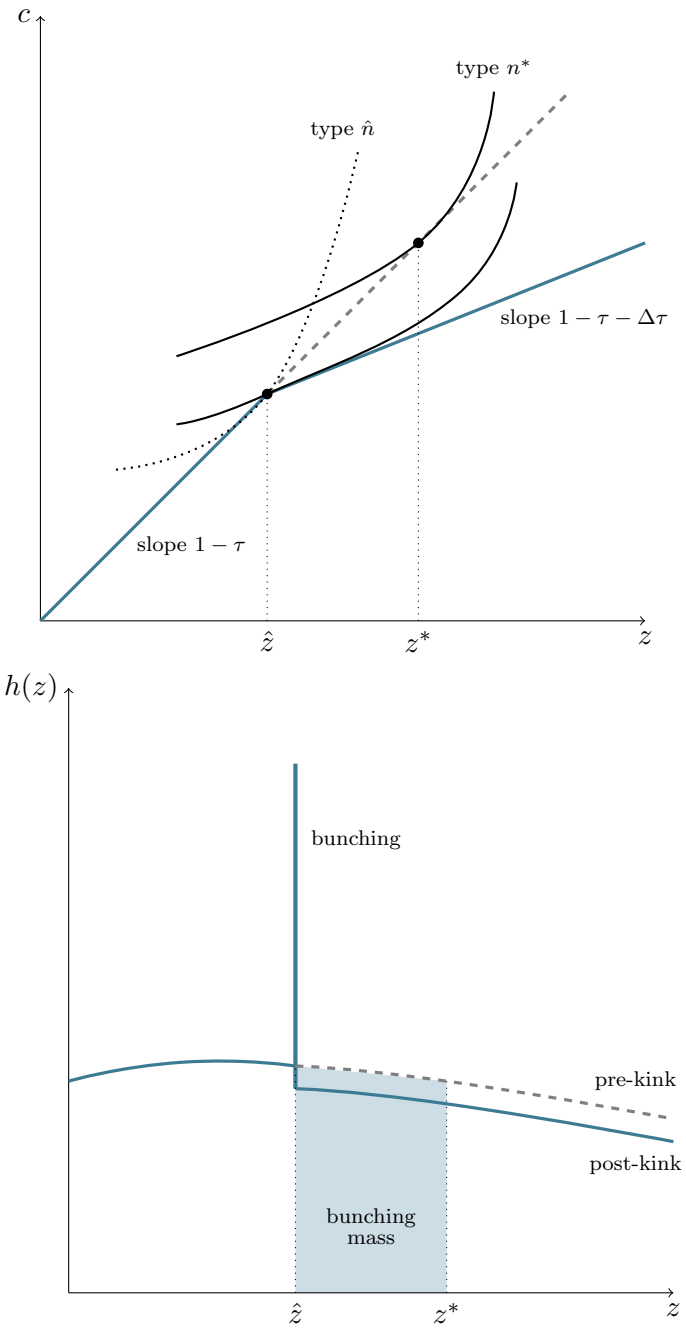
The arguments presented in this chapter have implications for the bunching literature. First, reference points may be an important confounder of the effect of economic incentives. For instance, if one were to estimate labor supply elasticities from bunching at statutory retirement ages, this would vastly exaggerate the responsiveness of retirement decisions to financial incentives. However, the methods proposed in this chapter point towards a solution to this problem. When bunching is observed at multiple thresholds that vary in the financial incentive and/or the presence of reference points, the "true" labor supply elasticity can be jointly estimated with reference dependence parameters. A second implication is that bunching methods can be used to quantify objects other than responses to extrinsic incentives. If the goal is to identify reference points, detecting bunching at such a threshold provides compelling, prima facie evidence of reference dependence and may be an alternative to other identification strategies requiring stronger assumptions. More generally, the

[16] The fact that both reforms have positive welfare effects is due to the fact that workers retire inefficiently early to begin with, since the pension system generally entails an implicit net-of-tax rate of less than 1.

concept of reference dependence used in this chapter can be extended to other non-financial factors that change discontinuously at some value of the running variable.

Two limitations of the approach may be worth pointing out. First, the methods are designed to quantify a reference point at a given location. This chapter is silent about why a reference point emerges, or how it may evolve endogenously. Second, a key assumption for bunching estimation is that the counterfactual density would be smooth in the absence of a budget constraint discontinuity or a reference point. This obviously excludes other sources of bunching such as firm responses, but it also calls for some caution in defining what is included in the notion of a reference point. For instance, in the case of statutory retirement ages, they may serve as an "intrinsic" reference point, but part of the reference point character may be an "extrinsic" social norm in favor of retiring at these ages. Hence, reference point effects may subsume several factors and it may be challenging to further disentangle these.

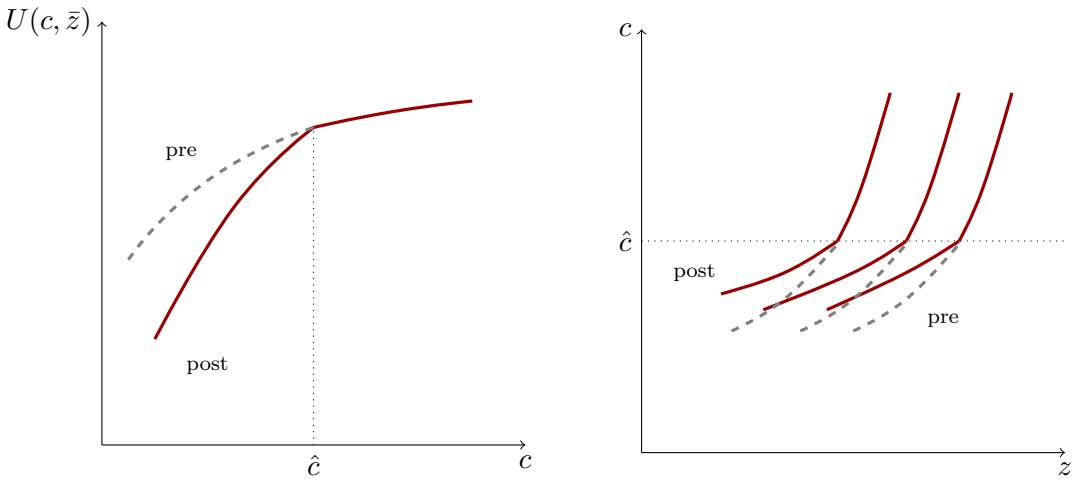
Figure 2.1: Bunching at a Budget Set Kink



Note: This figure shows an indifference curve diagram and a density diagram on bunching responses to a budget set kink. In the upper panel, the blue line is the post-kink budget set, whereas the dashed grey line is the pre-kink budget set. The dotted curve is an indifference curve of an individual with ability $\hat{n}$ who chooses $\hat{z}$ before and after the change. The solid curves are indifference curves of the marginal buncher with ability n^* who is tangent to the old budget set at z^* and tangent to the upper part of the new budget set at $\hat{z}$. In the lower panel, the solid blue line denotes the post-kink density, whereas the dotted line denotes the pre-kink density. The blue shaded area is the initial location of the mass of workers bunching in response to the kink.

Figure 2.2: Reference-Dependent Preferences

Panel A: Kink in Utility from Consumption



Panel B: Kink in Disutility from Work

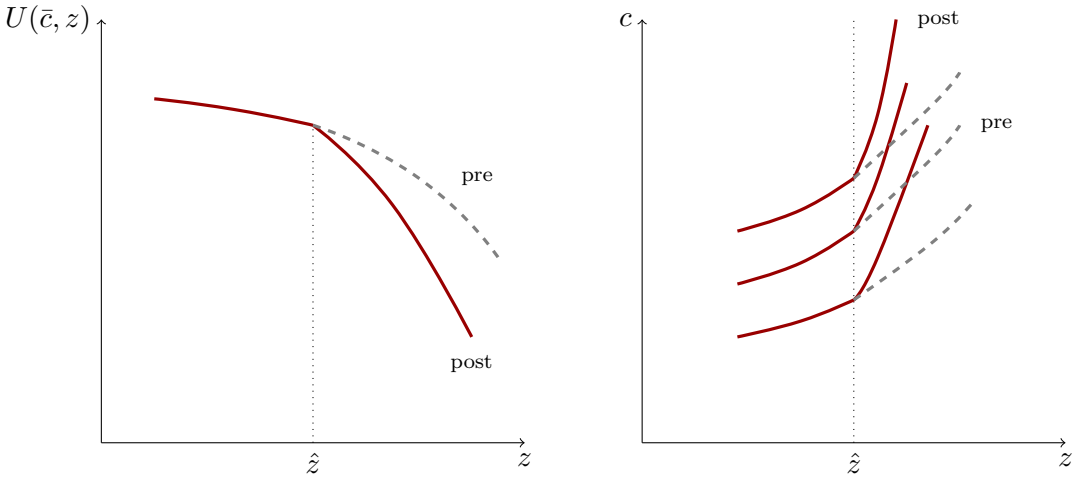
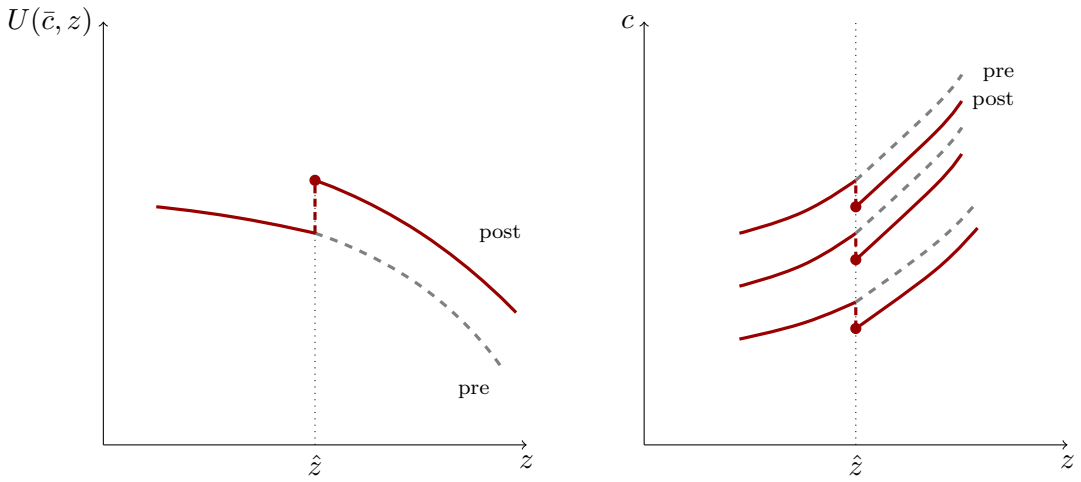
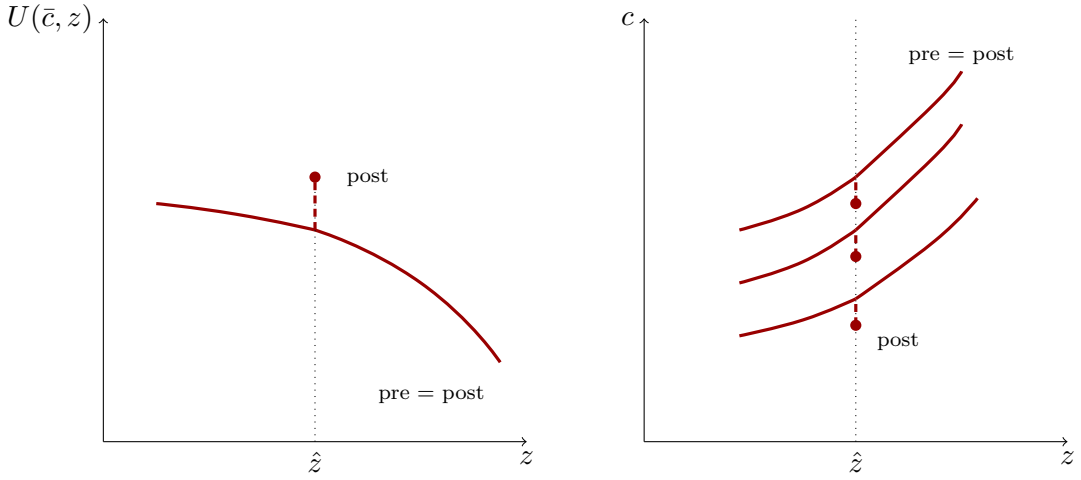


Figure 2.2 continued

Panel C: One-Sided Utility Notch

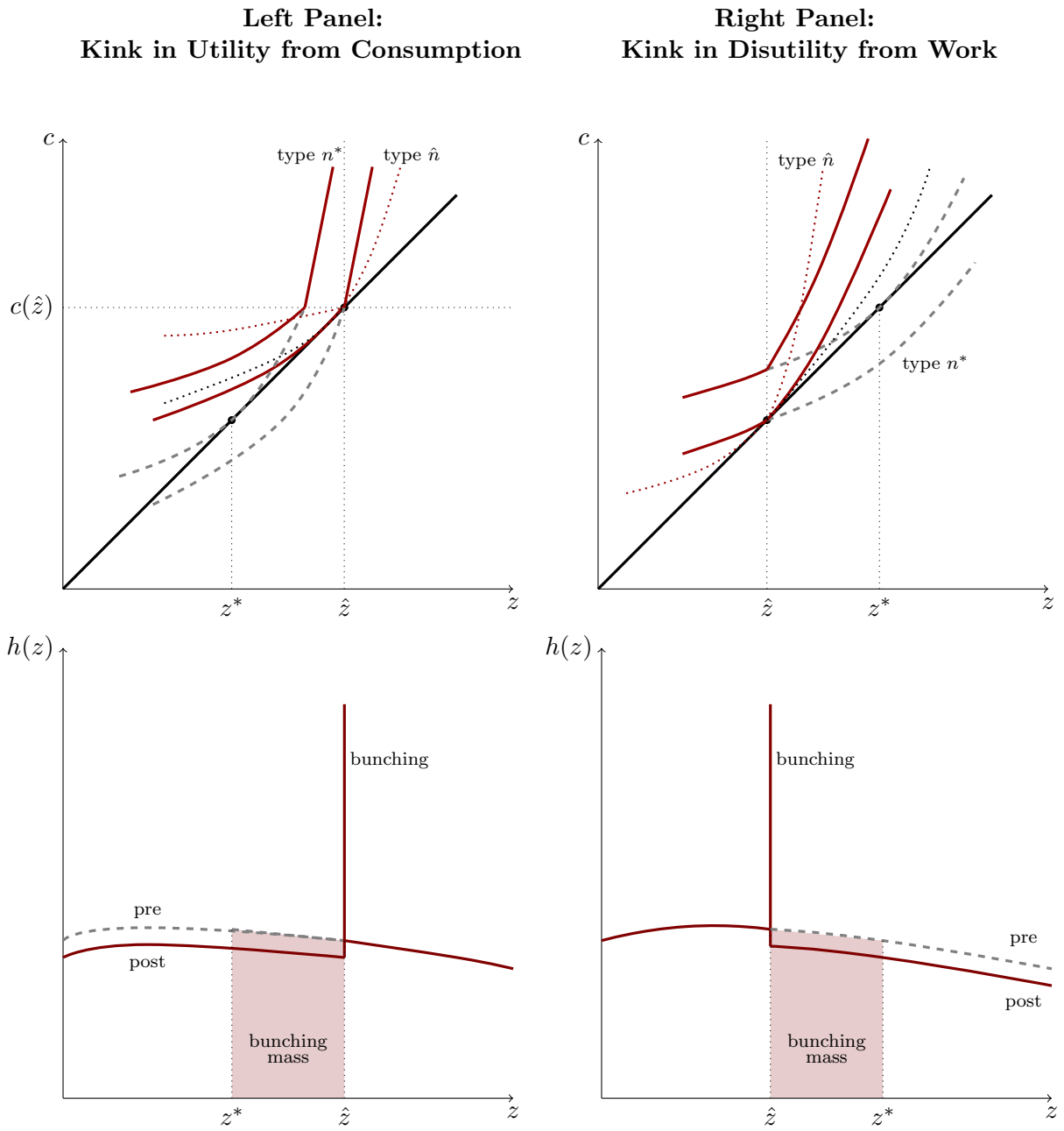


Panel D: Two-Sided Utility Notch



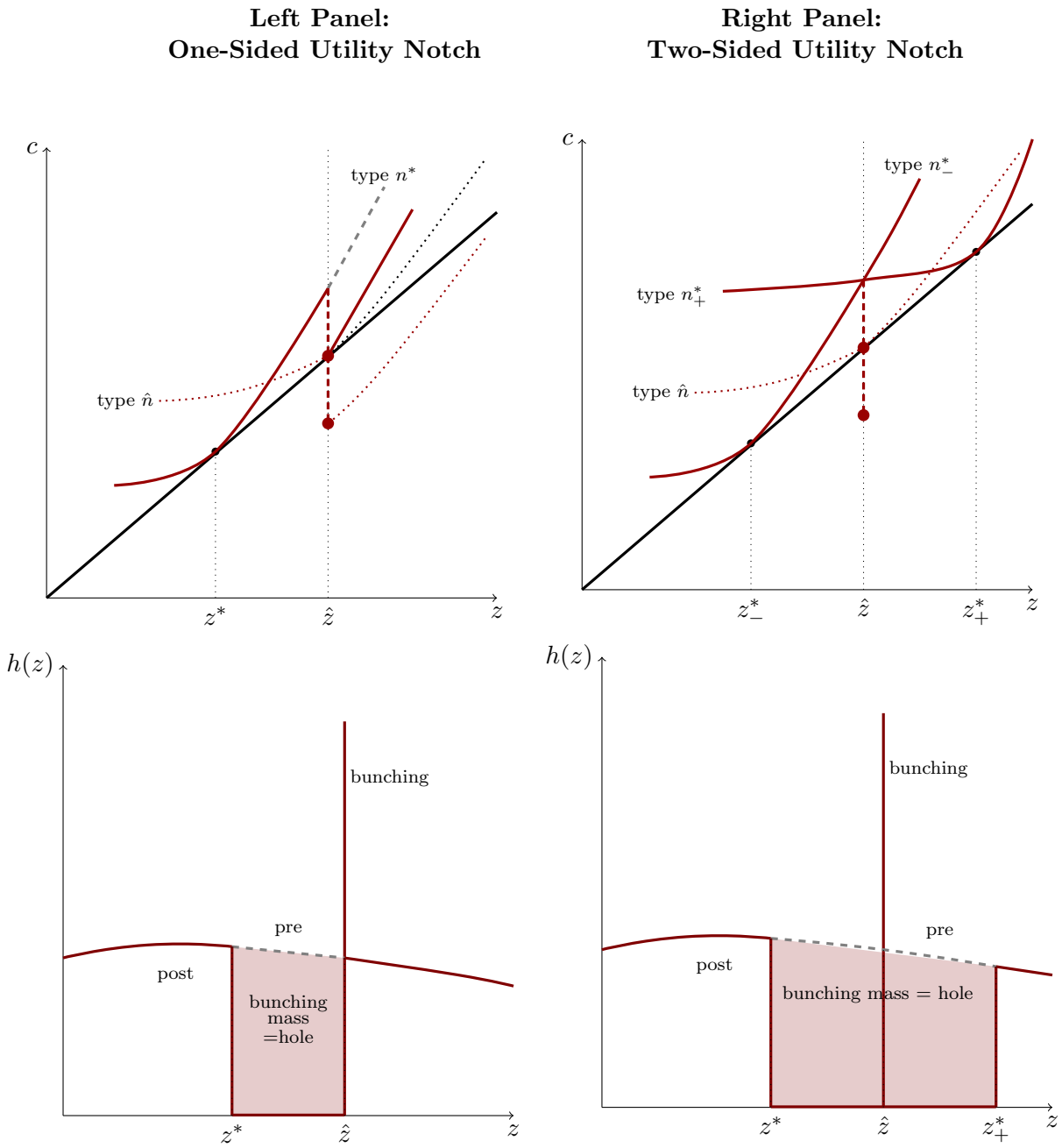
Note: This figure illustrates the effect of different types of reference-dependent preferences on the utility function (left panels) and indifference curves (right panels). "Pre" and "post" denote the situations before and after the introduction of the reference point, respectively. Panel A shows that a kink in utility from consumption according to equation (2.5) induces an increase in marginal utility from consumption below $\hat{c}$ and a kink in indifference curves at the threshold in the $z - c$ space, reflecting the discontinuous change in the marginal rate of substitution. Panel B illustrates the effect of a kink in disutility from work as in equation (2.6). Disutility from earning taxable income is steeper above the reference point $\hat{z}$, and indifference curves exhibit a kink at the threshold. Panel C shows that the one-sided utility notch from equation (2.7) implies an upward jump in the level of utility at the reference point $\hat{z}$, and a downward jump in indifference curves at $\hat{z}$ since less consumption is needed to compensate the individual once $\hat{z}$ is crossed. Panel D illustrates the two-sided utility notch from equation (2.8), where the individual derives higher utility only exactly at the reference point. Hence, indifference curves include a point below the curve at $\hat{z}$ since the individual is indifferent between a lower level of consumption at $\hat{z}$ and any point on the curve.

Figure 2.3: Bunching with Utility Kinks



Note: This figure shows bunching responses to a kink in utility from consumption (left panel) and to a kink in disutility from work (right panel) in an indifference curve diagram. In the top diagram of the left panel, the dashed grey curves are the pre-reference point indifference curves of the marginal buncher with ability n^*, whereas the solid red curves are her post-reference point indifference curves. The dotted curves are indifference curves pre-ref. (grey) and post-ref. (red) of an individual with ability $\hat{n}$ who retires at $\hat{z}$ before and after the change. The marginal buncher is tangent at z^* in the absence of the reference point, and tangent at $\hat{z}$ with the reference point. In the top diagram of the right panel, the dashed grey curves are the pre-ref. indifference curves of the marginal buncher with ability n^*, whereas the solid red curves are her post-ref. indifference curves. The dotted curves are indifference curves pre-ref. (grey) and post-ref. (red) of an individual with ability $\hat{n}$ who retires at $\hat{z}$ before and after the change. The marginal buncher is tangent at z^* in the absence of the reference point, and tangent at $\hat{z}$ with the reference point. In both of the bottom diagrams, the solid red line denotes the post-ref. density, whereas the dotted line denotes the pre-ref. density. The red shaded area is the initial location of the mass of workers bunching in response to the kink.

Figure 2.4: Bunching with Utility Notches

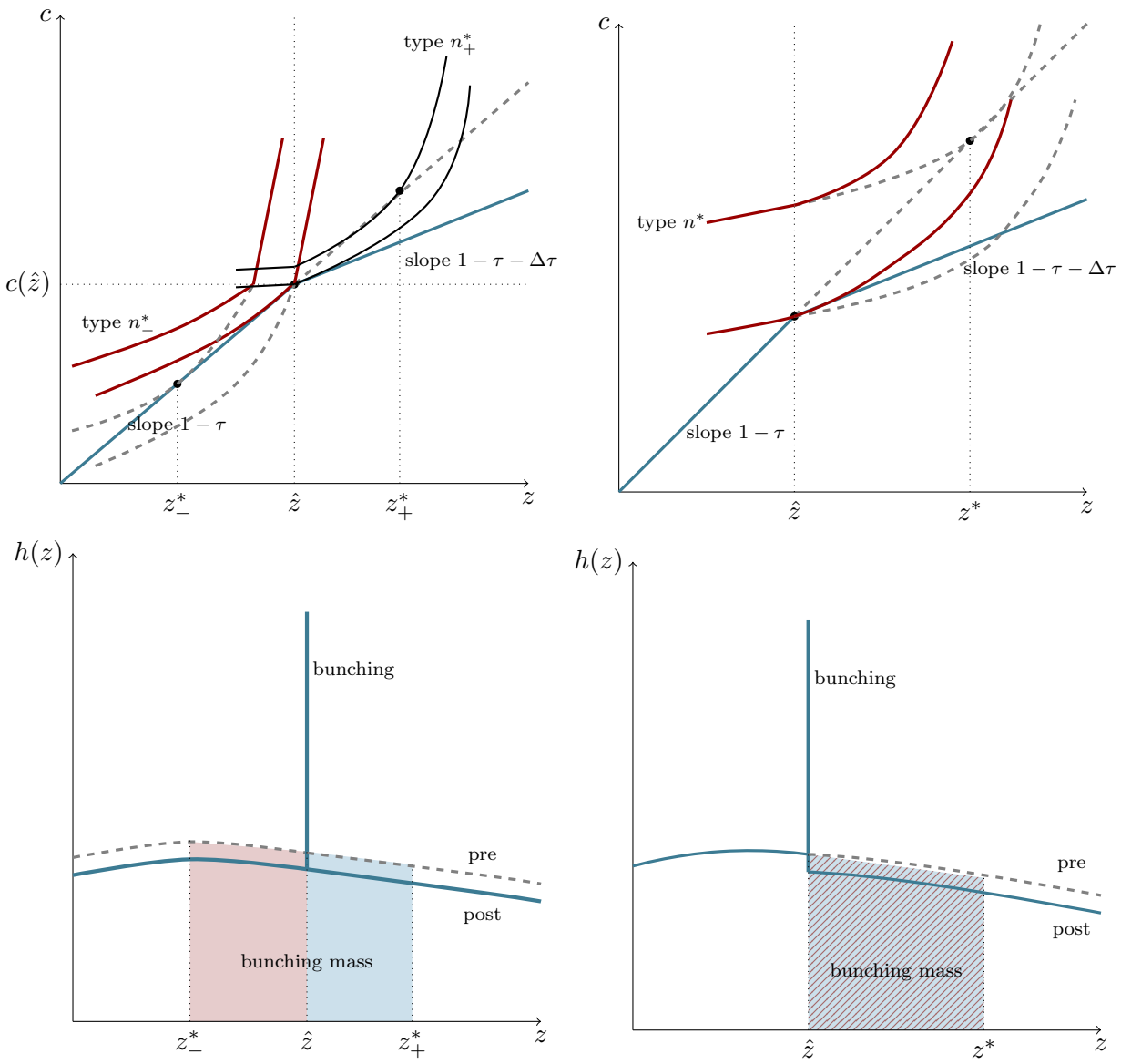


Note: This figure shows bunching responses to a one-sided utility notch (left panel) and to a two-sided utility notch (right panel) in an indifference curve diagram. In the top diagram of the left panel, the dashed grey curves are the pre-kink indifference curves of the marginal buncher with ability n^*, whereas the solid red curves are her post-kink indifference curves. The dotted curves are indifference curves pre-kink (grey) and post-kink (red) of an individual with ability $\hat{n}$ who chooses $\hat{z}$ before and after the change. The marginal buncher is tangent at z^* in the absence of the reference point and becomes indifferent between z^* and $\hat{z}$ with the reference point. In the top diagram of the right panel, the solid curves are the pre-kink indifference curves of the lower marginal buncher with ability n^*_- and the upper marginal buncher with ability n^*_+. Post-kink, the red dots become parts of the indifference sets to which they are connected with the dashed lines. The dotted curves is an indifference curves of an individual with ability $\hat{n}$ who chooses $\hat{z}$ before and after the change. The lower (upper) marginal bunchers is tangent at z^*_- (z^*_+) in the absence of the reference point and becomes indifferent between z^* (z^*_+) and $\hat{z}$ with the reference point. In both of the bottom diagrams, the solid red line denotes the post-notch density, whereas the dotted line denotes the pre-notch density. The red shaded area is the initial location of the mass of workers bunching in response to the notch.

Figure 2.5: Bunching when Reference Points Coincide with Economic Incentives

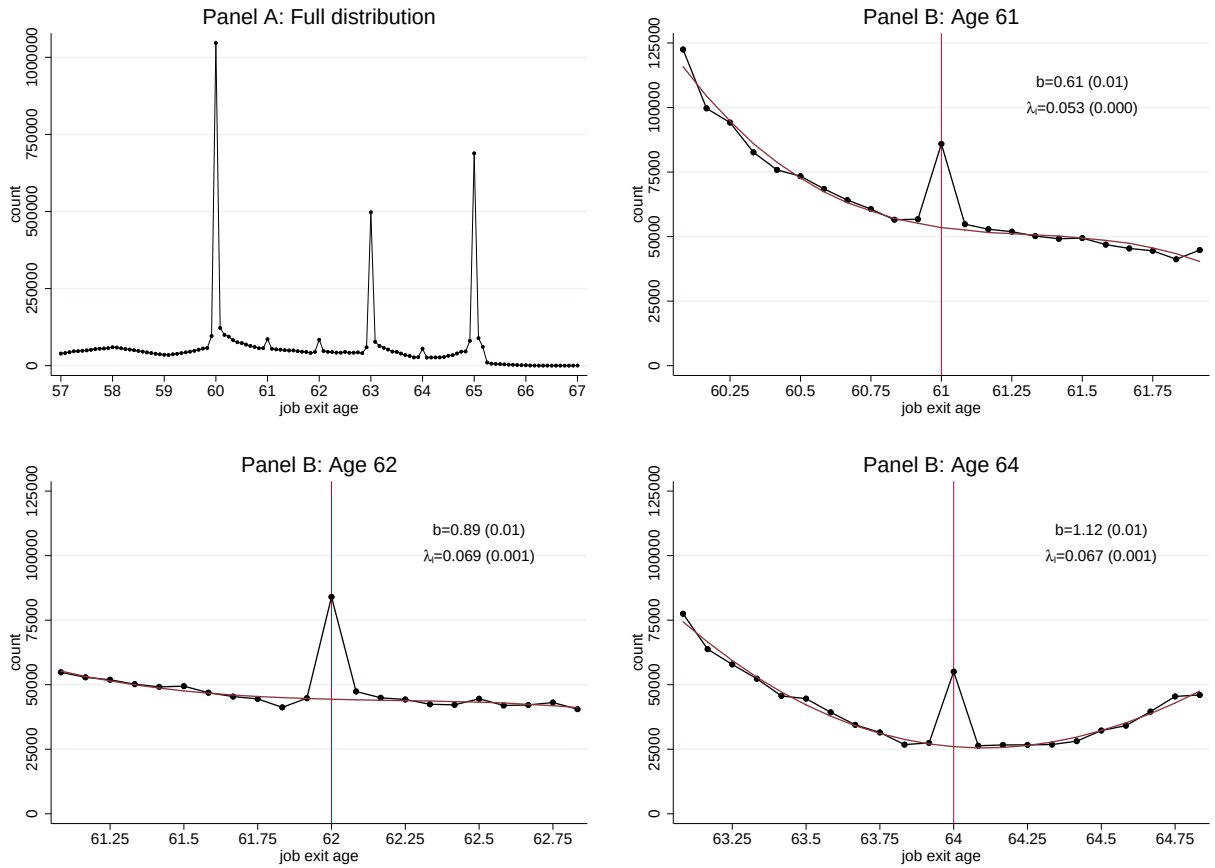
Left Panel:
Kink in Utility from Consumption

Right Panel:
Kink in Disutility from Work



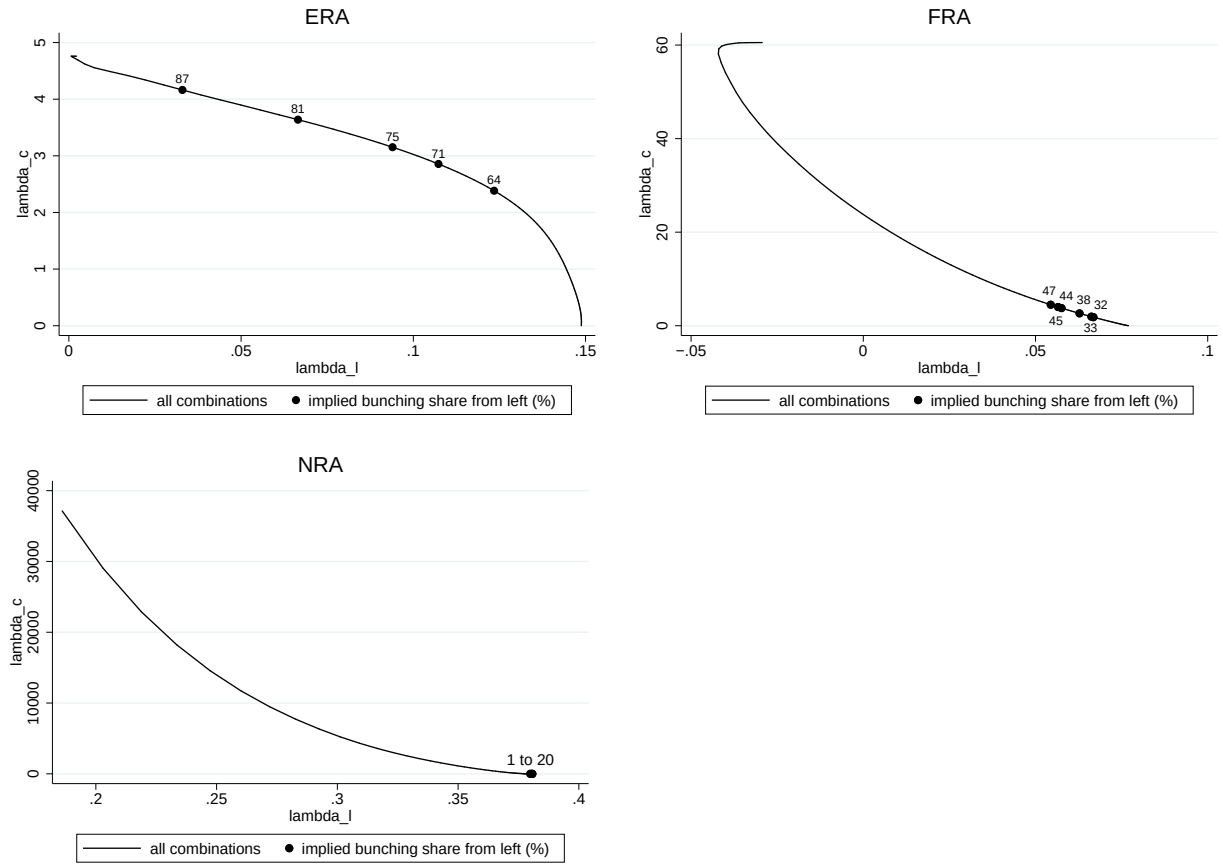
Note: This figure shows bunching responses to a kink in utility from consumption (left panel) and to a kink in disutility from work (right panel) in combination with a budget set kink. In the top diagram of the left panel, the blue line is the kinked budget set, whereas the dashed grey line is the initial budget set. The dashed grey curves are the initial indifference curves of the lower marginal buncher with ability n_-^* without the reference point, whereas the solid red curves are her indifference curves with the reference point. The lower marginal buncher is tangent at z_-^* in the absence of the reference point and the budget set kink, and tangent at $\hat{z}$ with the reference point and the budget set kink. The solid black curves are indifference curves of the upper marginal buncher who is tangent to the initial budget set at z_+^* and tangent to the kinked budget set at $\hat{z}$. In the top diagram of the right panel, the blue line is the kinked budget set, whereas the dashed grey line is the initial budget set. The dashed grey curves are the indifference curves of the marginal buncher with ability n^* without the reference point, whereas the solid red curves are her indifference curves with the reference point. The marginal buncher is tangent at z^* in the absence of the reference point and the budget set kink, and tangent at $\hat{z}$ with the reference point and the budget set kink. In both of the bottom diagrams, the solid blue line denotes the post-kinks density, whereas the dotted line denotes the pre-kinks density. In the bottom left diagram, the red shaded is the initial mass of workers bunching from below, and the blue shaded area is the initial location of the mass of workers bunching from above. In the bottom right diagram, the red-and-blue shaded area is the initial location of the mass of workers bunching in response to the combined threshold.

Figure 2.6: Bunching at Round Retirement Ages



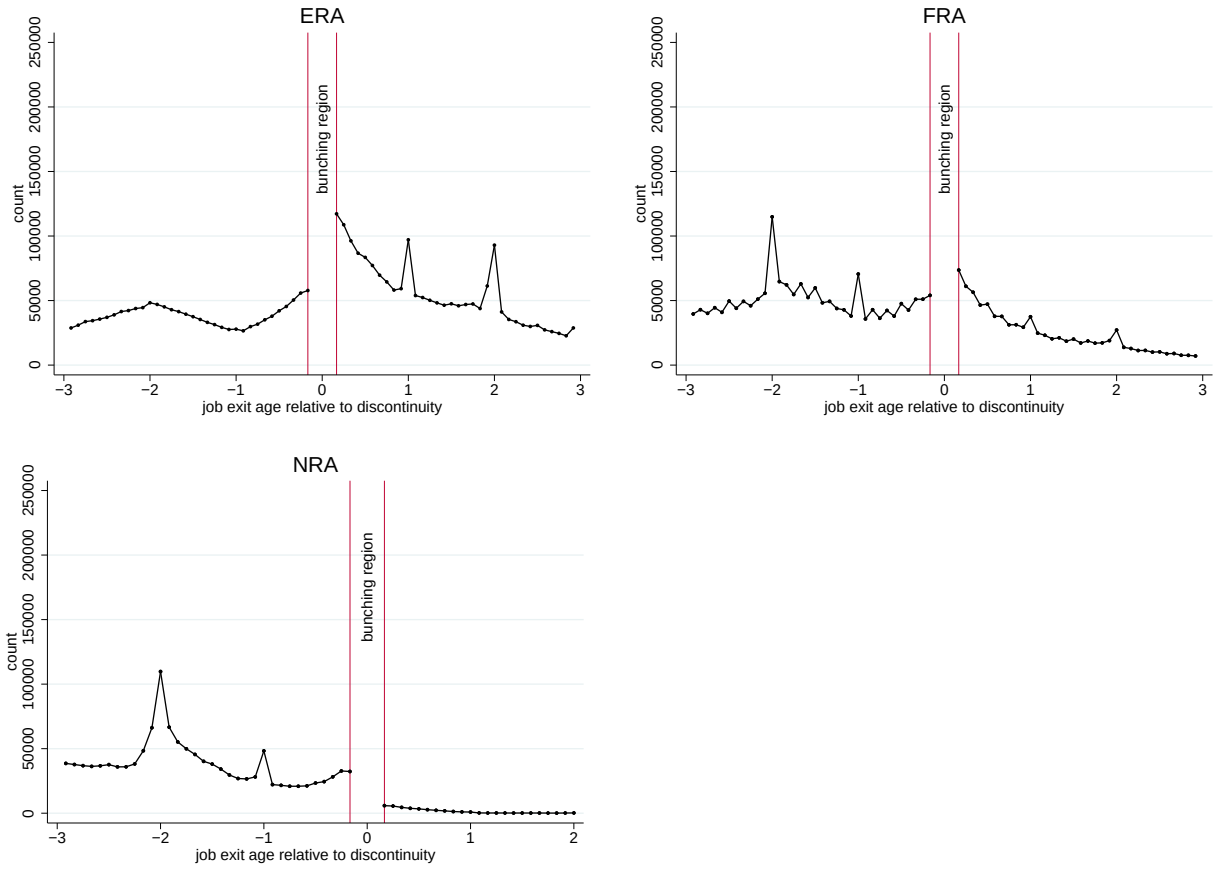
Note: This figure shows bunching at round retirement ages. Panel A shows the full distribution of retirement ages for workers born between 1933 and 1948, corresponding to chapter 1, figure 1.1. Note that the large spikes at ages 60, 63 and 65 correspond to the main statutory retirement ages. Panels B to D zoom into the distribution around other round ages, namely 61, 62 and 64. The connected black dots show counts of job exit ages in monthly bins. In panels B to D, the red curve is the counterfactual distribution estimated as a 7th-order polynomial. Vertical red lines indicate the location of the round age where bunching is estimated. b is the excess mass and λ_l is the implied parameter capturing the strength of reference dependence in labor supply estimated according to equation (2.23). Bootstrapped standard errors are in parantheses.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Figure 2.7: Structural Parameter Estimates from Statutory Retirement Ages



Note: This figure shows a range of estimated combinations of the parameters λ_c^s and λ_l^s for each statutory age type. The solid line in each panel shows the full range of possible combinations obtained from a simulation moving the share of bunching from the left at each discontinuity from 0 to its maximum in one-percentage point steps as described in appendix 2.A.5.2. The labeled dots mark parameter combinations corresponding to the estimated bunching shares based on the different methods and estimation windows shown in appendix table 2.A1.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

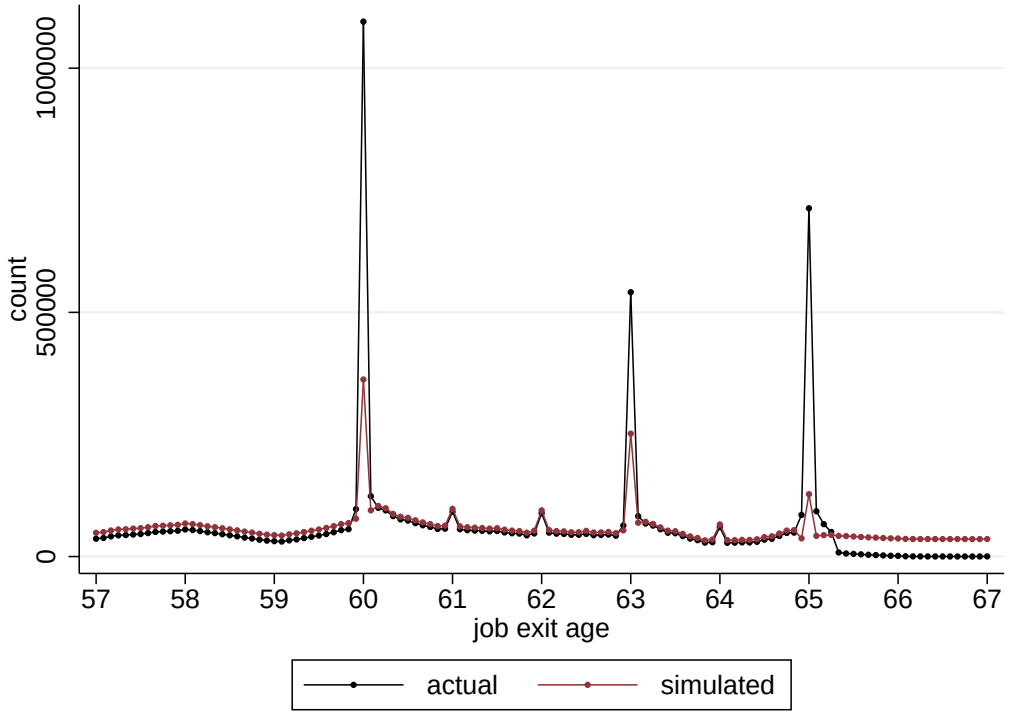
Figure 2.8: Empirical Density around Statutory Retirement Ages



Note: The connected black dots show the pooled density around all Early Retirement Ages (panel A), Full Retirement Ages (panel B), and Normal Retirement Ages (panel C), each excluding the threshold +/- 1 month.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Figure 2.9: Counterfactual Simulations

Panel A: Workers only respond to financial incentives

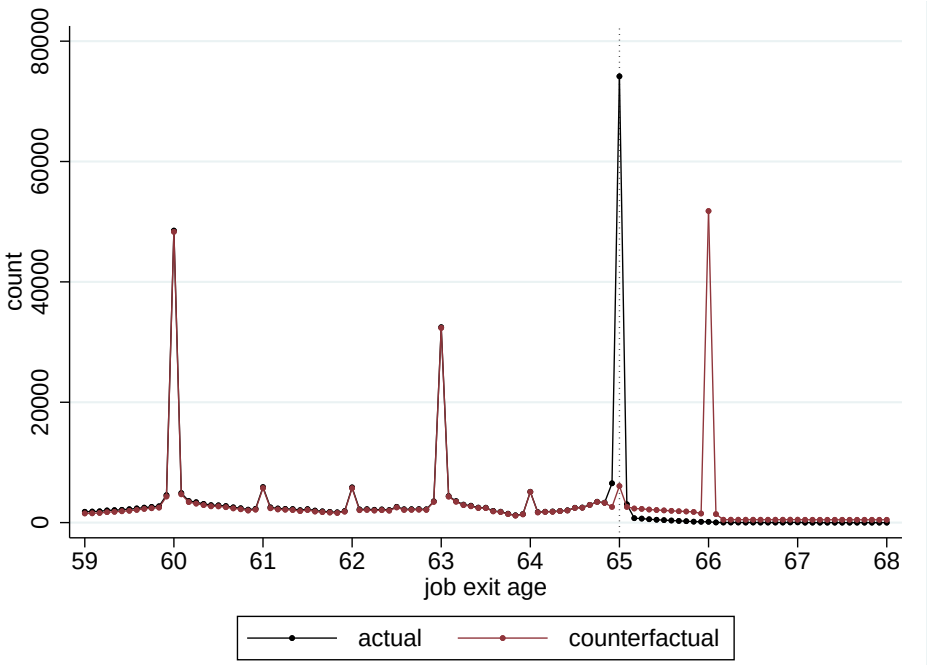


Note: The black connected dots show the actual distribution of job exit ages for all workers born between 1933 and 1948. The red connected dots show the distribution of job exits among the same workers, simulated under a counterfactual scenario with no reference point effects ($\lambda_c^s = \lambda_l^s = 0 \, \forall s$) and the central elasticity estimate $\varepsilon = 0.15$. Bunching at each discontinuity is simulated based on equation (2.25) and the remaining bunching mass is distributed over the remaining density according to the estimated bunching shares from the two sides.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

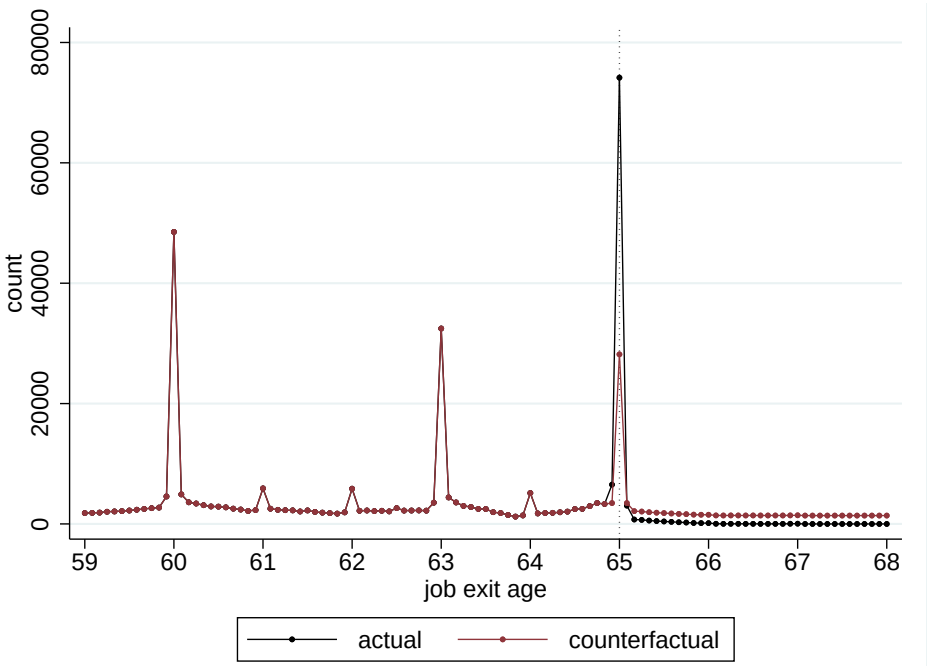
Figure 2.8: Counterfactual Simulations (continued)

Panel B: Policy counterfactuals

Policy 1: NRA increase to 66



Policy 2: Increase in late retirement rewards



Note: The black connected dots show the actual distribution of job exit ages for all workers born in 1946. The dotted vertical line indicates the actual NRA of 65. The red connected dots show the distribution of job exits among the same workers, simulated under a counterfactual scenario with an increase in the NRA from 65 to 66 (panel A), and an increase in financial rewards for late retirement (panel B).
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 2.1: Parameter Estimates at Round Retirement Ages

	(1) 61 years	(2) 62 years	(3) 64 years
b	0.61***	0.89***	1.12***
	(0.01)	(0.01)	(0.01)
λ_l	0.053***	0.069***	0.067***
	(0.0005)	(0.001)	(0.001)
Equivalent $\frac{\Delta\tau}{1-\tau}$	0.11	0.15	0.17
Observations	1,472,040	1,049,232	890,454

Note: This table shows estimates from round-number bunching at retirement ages 61, 62 and 64, corresponding to the bunching graphs in figure 2.6. The first row shows the excess mass b at each round age. In the second row, λ_l is the implied kink in disutility from work estimated according to equation (2.23). The third row shows the size of the budget kink that would induce the same response as the estimated reference point. The latter relies on the equivalence between a kink in disutility from work and a budget set kink discussed in section 2.2.3.2. Bootstrapped standard errors are in parentheses. "Observations" refers to the number of individuals included in the bunching estimation sample at each threshold.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 2.2: Parameter Estimates at Statutory Ages and Some Implications

Panel A: Parameter estimates			
Elasticity ε	0.15*** [0.13,0.17]		
	(1)	(2)	(3)
	ERA	FRA	NRA
λ_c	3.15***	4.48***	0.48***
	[1.87,6.34]	[2.27,10.00]	[0.12,1.00]
λ_l	0.083***	0.077***	0.38***
	[0.049,0.12]	[0.0035,0.079]	[0.38,0.45]
Panel B: Implied bunching responses			
	(1)	(2)	(3)
	ERA	FRA	NRA
Left bunching share α	0.64	0.47	0.058
ΔR^+ (months)	-4.74	-10.94	-29.46
ΔR^- (months)	9.02	9.67	2.32
ΔC^+ (Euros)	-4,932	-10,983	-36,662
$\frac{\Delta C^+}{C}$	-0.53%	-1.04%	-3.58%
ΔC^- (Euros)	13,916	12,468	1,879
$\frac{\Delta C^-}{C}$	1.31%	1.25%	0.17%
Discontinuities	117	257	93

Note: Panel A of this table shows parameter estimates resulting from a non-linear least squares estimation of the structural equation (2.25), using the bunching sample. The estimates shown here are from the preferred specification using gradient-corrected density shares. Appendix table 2.A2 presents full results from the structural estimation. Panel B shows metrics related to bunching responses, based on results from the structural estimation. Left bunching share is the estimated fraction of bunching originating from the left. ΔR^+ and ΔR^- is the change in the retirement of the right marginal buncher and the left marginal buncher, respectively. ΔC^+ and ΔC^- is the change in consumption implied by the right marginal bunching response and the left marginal bunching response, respectively, calculated in 2012 Euros. $\frac{\Delta C^+}{C}$ and $\frac{\Delta C^-}{C}$ is the change in consumption implied by the right marginal bunching response and the left marginal bunching response, as a fraction of estimated total lifetime consumption. All figures calculated at each statutory age discontinuity, the table shows weighted averages across discontinuities.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 2.3: Counterfactual Simulations

Panel A: Workers respond only to incentives				
	(1)	(2)	(3)	(4)
	actual	counterfactuals		
		$\varepsilon = 0.15$	$\varepsilon = 0.09$	$\bar{\varepsilon}_g = 0.29$
Percentage of job exits at statutory ages	29.0	6.04	4.92	10.1
% explained (of actual)		20.8%	17.0%	34.8%
Average excess mass				
at all discontinuities	19.2	0.99	0.48	2.00
% explained (of actual)		5.16%	2.50%	10.43%
at all statutory age kinks	21.8	0.65	0.27	1.81
% explained (of actual)		2.98%	1.24%	8.29%
Average job exit age	60.84	60.77		
change (months)		-0.9		

Table 2.3 continued

Panel B: Policy counterfactuals

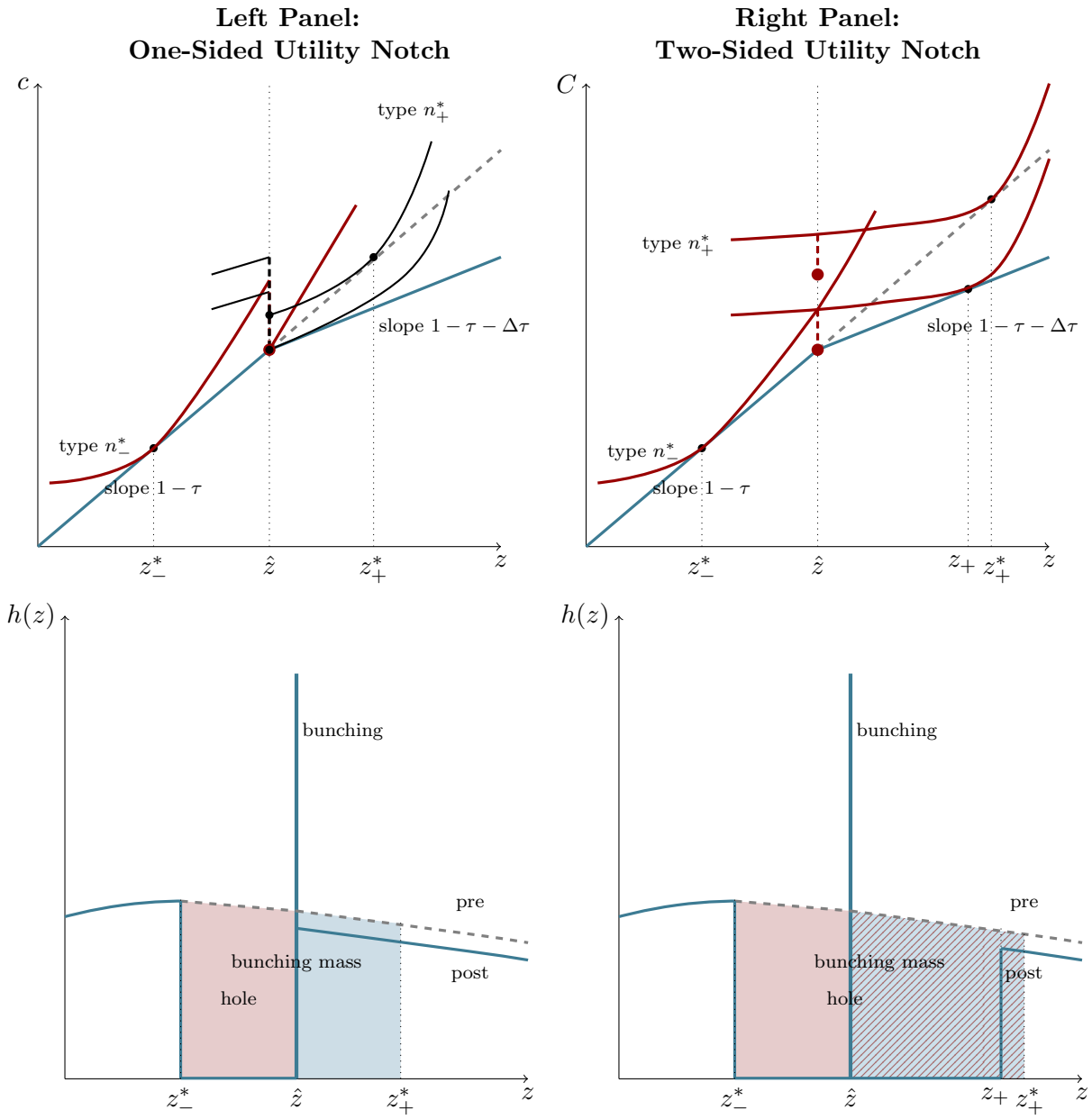
	(1) actual	(2) counterfactuals	(3) counterfactuals
Policy		NRA increase from 65 to 66	increase in rewards for late retirement from 6% to 10.5%
Average job exit age (65 and above)	65.0	65.9	65.9
change		+11	+11
Average job exit age (60 and above)	62.8	63.1	63.0
change (months)		+4	+3
Excess mass at NRA	28.5	23.1	12.4
change		-5.4	-16.1
Net fiscal effect (NPV)		+€731m	-€206m
contributions collected		+€301m	+€301m
benefits paid		-€433m	+€504m
Worker consumption (NPV)		+€630m	+€1567m
Worker welfare (NPV equivalent)			
$\omega = 1$		+€66m	+€804m
$\omega = 0$		+€62m	+€1262m
Total welfare (NPV equivalent)			
$\omega = 1$		+€797m	+€598m
$\omega = 0$		+€793m	+€1056m

Note: This table shows results from a simulation of bunching at budget constraint discontinuities in the absence of statutory age effects. Panel A assumes that the only source of bunching at statutory ages are financial incentives for workers. Column (1) shows the actually observed percentage of job exits at statutory ages and average excess mass. Columns (2) and (3) show figures from simulating bunching, using the elasticity estimates from columns (2) and (3) of table (1.7), respectively. Panel B shows results from a simulation of two counterfactual policies: an increase in the NRA from 65 to 66 (column 1) and an increase in rewards for late retirement (column 2). The size of rewards in column (2) is calibrated to match the effect on the average job exit age in column (1). All excess mass statistics weighted by group size.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

2.A Appendix

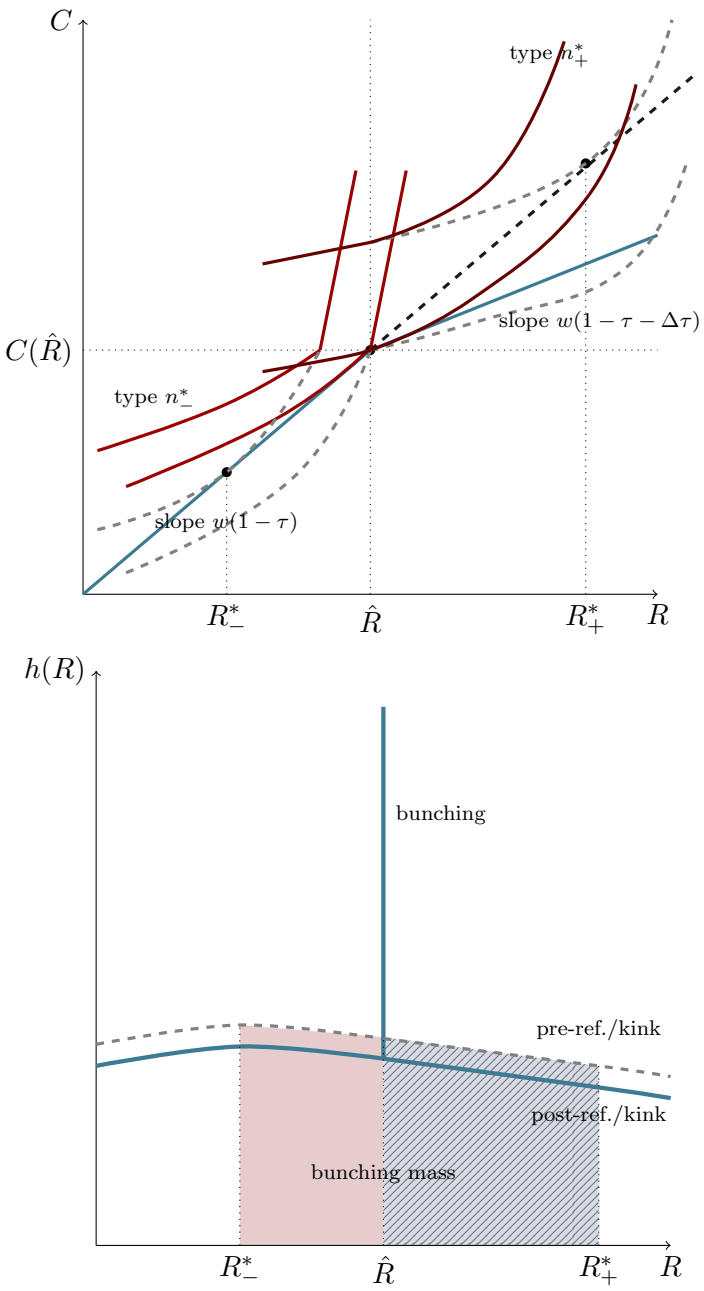
2.A.1 Appendix Figures and Tables

Figure 2.A1: Bunching when Utility Notches Coincide with Economic Incentives



Note: This figure shows bunching at a one-sided utility notch (left panel) and at a two-sided utility notch (right panel) in combination with a budget set kink. In both top diagrams, the blue line is the kinked budget set, whereas the dashed grey line is the initial budget set. The solid red curves are the pre-notch indifference curves of the lower marginal buncher with ability n_-^* and the upper marginal buncher with ability n_+^*. Post-notch, the red dots become parts of the indifference sets to which they are connected with the dashed lines. The lower marginal buncher is indifferent between z_-^* and $\hat{z}$ with the reference point and the budget set kink. The upper marginal buncher is tangent at z_+^* in the absence of the budget set kink, and tangent at z_+ with the budget set kink, where she is indifferent between this point and $\hat{z}$ due to the reference point effect. In the top right diagram, the solid red curves are the pre-notch indifference curves of the lower marginal buncher with ability n_-^* and the upper marginal buncher with ability n_+^*. Post-notch, the red dots become parts of the indifference sets to which they are connected with the dashed lines. The lower marginal buncher is indifferent between z_-^* and $\hat{z}$ with the reference point and the budget set kink. The upper marginal buncher is tangent at z_+^* in the absence of the budget set kink, and tangent at z_+ with the budget set kink, where she is indifferent between this point and $\hat{z}$ due to the reference point effect. In the bottom diagrams, the solid blue line denotes the pre density, whereas the dotted line denotes the post density. In the bottom left, the red shaded is the initial mass of workers bunching from below, and the blue area is the initial mass bunching from above. In the bottom right, the red shaded is the origin of bunching from below, and the red and blue area is the origin of bunching from above.

Figure 2.A2: Theoretical Bunching at a Statutory Retirement Age



Note: This figure shows bunching responses to a threshold combining a budget set kink, a kink in utility from consumption and a kink in disutility from work in an indifference curve diagram (upper panel) and a density diagram (lower panel). In the upper panel, the blue line is the kinked budget set, whereas the dashed grey line is the initial budget set. The dashed grey curves to the left of $\hat{R}$ are the initial indifference curves of the lower marginal buncher with ability n_-^* without the reference point, whereas the solid red curves are her indifference curves with the reference point. The lower marginal buncher is tangent at R_-^* in the absence of the reference point and the budget set kink, and tangent at $\hat{R}$ with the reference point and the budget set kink. The dashed grey curves to the right of $\hat{R}$ are the initial indifference curves of the upper marginal buncher with ability n_+^*, whereas the solid red curves are her indifference curves with the budget set kink and the reference point. The upper marginal buncher is tangent at R_+^* in the absence of the reference point and the budget set kink, and tangent at $\hat{R}$ with the reference point and the budget set kink. In the lower panel, the solid blue line denotes the post-ref./kink density, whereas the dotted line denotes the pre-ref./kink density. The red shaded area is the initial location of the mass of workers bunching from the left in response to the kink in utility from consumption, while the blue and red shaded area is the initial location of the mass of workers bunching from the right in response to the budget set kink and the kink in disutility from work.

Table 2.A1: Estimated Bunching Shares from the Left vs. Right at Statutory Ages

	(1)	(2)	(3)	(4)	(5)	(6)
	Basic Estimation			Gradient-Corrected Estimation		
Window	12 months	24 months	36 months	12 months	24 months	36 months
m_{-}^{ERA}	0.71	0.81	0.87	0.64	0.75	0.81
	(0.21)	(0.26)	(0.27)	(0.34)	(0.34)	(0.32)
m_{-}^{FRA}	0.38	0.33	0.32	0.47	0.45	0.44
	(0.25)	(0.34)	(0.37)	(0.41)	(0.44)	(0.46)
m_{-}^{NRA}	0.20	0.054	0.0085	0.058	0.024	0.015
	(0.12)	(0.078)	(0.036)	(0.16)	(0.13)	(0.10)
Discont.	386	386	386	386	386	386

Note: This table shows bunching shares from the left vs. the right of statutory age thresholds based estimated based on the relative density on both sides. m_{-}^{stat} denotes the share of missing density from the left out of total bunching at statutory type *stat*. Missing density is computed as described in appendix 2.A.5.2. Each column shows averages across all discontinuities of the respective type, with standard deviations in parantheses. All statistics weighted by group size.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 2.A2: Full Set of Structural Estimates from Statutory Ages

parameter	(1) upper bounds on λ^c	(2) upper bounds on λ^l	(3) ad-hoc comb. of upper bounds	(4) basic density shares	(5) gradient-corrected density shares
ε	0.15***	0.15***	0.15***	0.15***	0.15***
	[0.14,0.17]	[0.14,0.17]	[0.14,0.17]	[0.13,0.17]	[0.13,0.17]
λ^c_{ERA}	4.75***		4.79***	4.36***	3.15***
	[3.18,7.49]		[3.19,7.73]	[2.82,7.69]	[1.87,6.34]
λ^c_{FRA}	57.16***		41.72***	2.28***	4.48***
	[18.99,216.58]		[12.47,125.76]	[1.56,3.54]	[2.27,10.00]
λ^c_{NRA}	437,346***			3.28***	0.48***
	[7649,7.76e+09]			[1.94,5.76]	[0.12,1.00]
λ^l_{ERA}		0.15***		0.099***	0.083***
		[0.14,0.16]		[0.062,0.13]	[0.049,0.12]
λ^l_{FRA}		0.077***		0.068*	0.077***
		[0.0035,0.079]		[-0.0036,0.073]	[0.0035,0.079]
λ^l_{NRA}		0.38***	0.38***	0.38***	0.38***
		[0.38,0.45]	[0.38,0.45]	[0.38,0.45]	[0.38,0.45]
Obs. (discontinuities)	644	644	644	644	644

Note: This table shows results from a non-linear least squares estimation of the structural equation (2.25). The estimation proceeds in two steps: First, estimation is run on the subsample of pure financial incentive discontinuities, thus obtaining an estimate of ε. Second, the remaining coefficients are estimated using the full sample. Appendix 2.A.5 provides further details on estimation equations. The specification in column (1) estimates upper bounds on λ^s_c by assuming that all $\lambda^s_l = 0$. Conversely, column (2) estimates upper bounds on λ^s_l by assuming that $\lambda^s_c = 0$. Columns (3) estimates an ad-hoc combination by assuming that $\lambda^{NRA}_c = 0$ and $\lambda^{ERA}_l = \lambda^{FRA}_l = 0$. Estimates in columns (4) and (5) are based on observed density shares on both sides of the threshold as described in appendix 2.A.5.2. Bootstrapped 95% confidence intervals are shown in square brackets. *** p<0.01, ** p<0.05, * p<0.1.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

Table 2.A3: Counterfactual Bunching Simulations: Robustness

Allowing for some mandatory retirement

	(1) actual	(2) counterfactual $\lambda_l^{NRA} = \lambda_l^{FRA}$	(3) counterfactual all NRA due to mandatory
Percentage of job exits at statutory ages	29.0	7.41	16.6
% explained (of actual)		25.6%	57.2%
Average excess mass			
at all discontinuities	19.2	1.73	12.7
% explained (of actual)		9.01%	66.1%
at all statutory age kinks	21.8	1.51	14.3
% explained (of actual)		6.93%	65.6%
at convex statutory age kinks	14.7	1.83	2.24
% explained (of actual)		12.4%	15.2%

Note: This table shows results from an analogous exercise to panel A of table 2.3, assuming that the only sources of bunching at statutory ages are financial incentives for workers and firms using the NRA to lay off some workers. Column (2) is based on a restricted specification where $\lambda_l^{NRA} = \lambda_l^{FRA}$ and the remaining effect of the NRA is attributed to firm responses. In column (3), the entire effect of the NRA is attributed to firm responses. All statistics weighted by group size.
Data source: FDZ-RV - Themenfile SUFRTZN1992-2014XVSBB_Seibold

2.A.2 Reference Dependence and Economic Incentives: The Case of Utility Notches

2.A.2.1 One-Sided Utility Notch

Suppose that utility is reference-dependent according to equation (2.7). The left panels of appendix figure 2.A1 illustrates the joint effect of the utility notch and the budget set kink. The individual initially located to the left of the threshold at z_-^* is now indifferent between z_-^* and $\hat{z}$. At the same time, the individual initially tangent at z_+^* to the right of the threshold is now tangent at $\hat{z}$. These two individuals are the lower and the upper marginal buncher, respectively, and all workers initially located between z_-^* and z_+^* now bunch. Individuals initially to the left of z_-^* do not alter their choice, but individuals initially to the right of z_+^* decrease their labor supply in response to the flatter budget line. Hence, bunching at the threshold occurs from both sides, but there is a density hole only to the left of the threshold.

Analogously to equation (2.12), the indifference condition for the lower marginal buncher $\hat{U}_- = U_{I-}$ implies

$$\frac{1}{1+\varepsilon}\frac{z_-^*}{\hat{z}} + \frac{\varepsilon}{1+\varepsilon}\left(\frac{z_-^*}{\hat{z}}\right)^{-\frac{1}{\varepsilon}} = 1 + \frac{\delta}{c(\hat{z})}$$

For the upper marginal buncher, bunching is analogous to equation (2.3). The two tangency conditions $z_+^* = n_+^*(1-\tau)^\varepsilon$ and $\hat{z} = n_+^*(1-\tau-\Delta\tau)^\varepsilon$ imply

$$\frac{z_+^*}{\hat{z}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^\varepsilon$$

Hence, bunching has two additive components in this case. All bunching due to the reference point character of the threshold is from below, occurs independently of the size of the budget set kink, and follows the standard utility notch bunching formula (2.12). On the other hand, all bunching due to the budget set kink occurs from above, is independent of the strength of the reference point, and follows the standard budget set kink bunching formula (2.3).

2.A.2.2 Two-Sided Utility Notch

Suppose finally that utility is reference-dependent according to equation (2.8). The right panels of appendix figure 2.A1 illustrates the joint effect of the utility notch and the budget set kink. The individual initially located to the left of the threshold at z_-^* is now indifferent between z_-^* and $\hat{z}$. This individual is the lower marginal buncher. To the right of the threshold, the individual whose indifference curve is initially tangent to the budget set with slope $1-\tau$ at z_+^* is now indifferent between the point of tangency to the new budget set with slope $1-\tau-\Delta\tau$ at z_+ and the threshold $\hat{z}$. This worker is the upper marginal buncher. Thus, bunching occurs from both sides, with a hole in the density between z_-^* and z_+.

Analogously to equation (2.12), the indifference condition for the lower marginal buncher $\hat{U}_- = U_{I-}$ implies

$$\frac{1}{1+\varepsilon}\frac{z_-^*}{\hat{z}} + \frac{\varepsilon}{1+\varepsilon}\left(\frac{z_-^*}{\hat{z}}\right)^{-\frac{1}{\varepsilon}} = 1 + \frac{\delta_2}{c(\hat{z})}$$

The upper marginal buncher's utility at $z_+ = n_+^*(1-\tau-\Delta\tau)^\varepsilon$ can be written as

$$U_{I+} = \Delta\tau z + \frac{1}{1+\varepsilon}n^*(1-\tau-\Delta\tau)^{1+\varepsilon}$$

The indifference condition $\hat{U}_+ = U_{I+}$ implies

$$\frac{1}{1+\varepsilon}\frac{z_+}{\hat{z}} + \frac{\varepsilon}{1+\varepsilon}\left(\frac{z_+}{\hat{z}}\right)^{-\frac{1}{\varepsilon}} = 1 + \frac{\delta_2}{c(\hat{z})}\frac{1-\tau}{1-\tau-\Delta\tau} \tag{2.29}$$

However, $z_+/\hat{z}$ relates to the additional density shift due to the utility notch, given that there is a budget set kink. In order to capture the entire density shift due to the joint effect, the tangency condition $z_+^* = n_+^*(1-\tau)^\varepsilon$ can be combined with the tangency condition at z_+ to yield

$$\frac{z_+^*}{z_+} = \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^\varepsilon$$

and plugging this into the above indifference condition yields

$$\frac{1}{1+\varepsilon}\frac{z_+^*}{\hat{z}}\left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^{-\varepsilon} + \frac{\varepsilon}{1+\varepsilon}\left(\frac{z_+^*}{\hat{z}}\left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^{-\varepsilon}\right)^{-\frac{1}{\varepsilon}} = 1 + \frac{\delta_2}{c(\hat{z})}\frac{1-\tau}{1-\tau-\Delta\tau} \tag{2.30}$$

Hence, while bunching from the left occurs only due to the utility notch and is independent of the budget set kink, bunching from the right is due to a combined effect of the two. Compared to the standard utility notch bunching result from equation (2.12), the budget set kink has two additional effects on bunching in equation (2.30). First, the dominated region to the right of the reference point is extended by the budget set kink:

$$\lim_{\varepsilon\to 0+}\frac{z_+^*}{\hat{z}} = 1 + \frac{\delta_2}{c(\hat{z})}\frac{1-\tau}{1-\tau-\Delta\tau}$$

Intuitively, the lower net-of-tax rate to the right makes the reference point more attractive to workers initially located to the right of it, since the consumption gain from any deviation to the right is now smaller. Second, the budget set kink causes a density shift to the left, moving additional workers into the region where they bunch due to the utility notch. This can be seen in equation (2.30) where any occurrence of $z_+^*/\hat{z}$ on the left-hand side is scaled down by the inverse kink size, implying that $z_+^*/\hat{z}$ must have a larger value to satisfy the equation.

Algebraically, there is no straightforward relationship between an upper marginal buncher described by equation (2.3) and the one in (2.30), so quantifying the additional bunching due to the two-sided utility notch given an existing budget set kink is less straightforward than in the previous cases. However, the following argument yields an approximation of the additional bunching: First, denote by z^* the marginal buncher at a budget constraint kink of size $(1-\tau)/(1-\tau-\Delta\tau)$ described by equation (2.3). For this marginal buncher, $\hat{U} = U_I$ and for all other bunchers $\hat{U} > U_I$. Define $h_k(z)$ the post-kink density. Now introduce a two-sided utility notch at the location of the kink. After introducing the notch, all workers bunching previously due to the kink still bunch since $\hat{U} \geq U_I$ implies $\hat{U} + \delta_2 \geq U_I$. Suppose the density shift due to the kink has a negligible effect on the local density to the right of $\hat{z}$, i.e. $h_k(z) \approx h_0(\hat{z})$. Then the additional bunching from above due to the utility notch is the same as bunching from above with a utility notch and a global implicit net-of-tax rate $1-\tau-\Delta\tau$. This is the case because i) the location of the upper marginal buncher in the left panel of figure 2.4 only depends on the net-of-tax rate to the right of the kink, and ii) this implies the same amount of bunching when $h_k(z) \approx h_0(\hat{z})$ on $[\hat{z}, z_+^*]$.

Hence, total bunching is bunching from below plus bunching from above, where bunching from above is approximately additive in bunching due to the budget set kink and bunching due to the utility notch if interior responses due to the kink have a small effect on the local density.

2.A.3 Approximate Bunching Quantities

2.A.3.1 Budget Set Kink

Bunching can be written as

$$B = h_0(\hat{z})\Delta z^*$$

where $\Delta z^* = z^* - \hat{z}$. Rearranging equation (2.3)

$$\frac{b}{\hat{z}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^{\varepsilon} - 1$$

where $b = B/h_0(\hat{z})$ is the excess mass. When $\Delta\tau$ is small and hence Δz^* is small, such that $\log(z^*/\hat{z}) \approx \Delta z^*/\hat{z}$, and $\log(1-\tau-\Delta\tau)/(1-\tau)) \approx -\Delta\tau/(1-\tau)$. Take logs on equation (2.3)

$$\log\frac{z^*}{\hat{z}} = -\varepsilon\,\log\left(1 - \frac{\Delta\tau}{1-\tau}\right)$$

which implies

$$\frac{b}{\hat{z}} \approx \varepsilon\frac{\Delta\tau}{1-\tau} \tag{2.31}$$

This corresponds to the bunching formula derived by Saez (2010) that holds regardless of the functional form of utility.

2.A.3.2 Combining a Budget Set Kink and a Reference Point

2.A.3.2.1 Kink in Utility from Consumption

Bunching can be written as

$$B = h_0(\hat{z})(\Delta z_+^* + \Delta z_-^*)$$

where $\Delta z_+^* = z_+^* - \hat{z}$ and $\Delta z_-^* = \hat{z} - z_+^*$. Combining equations (2.3) and (2.9)

$$\frac{b}{\hat{z}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^{\varepsilon} - \left(\frac{1}{1+\lambda_c}\right)^{\varepsilon}$$

When $\Delta\tau$ is small and hence Δz_+^* is small,

$$\frac{b}{\hat{z}} \approx \varepsilon\frac{\Delta\tau}{1-\tau} + 1 - \left(\frac{1}{1+\lambda_c}\right)^{\varepsilon} \tag{2.32}$$

When in addition λ_c is small, $\log(1+\lambda_c)) \approx \lambda_c$, and

$$\frac{b}{\hat{z}} \approx \varepsilon\frac{\Delta\tau}{1-\tau} + \varepsilon\lambda_c \tag{2.33}$$

2.A.3.2.2 Kink in Disutility from Work

Bunching can be written as

$$B = h_0(\hat{z})\Delta z^*$$

or

$$\frac{b}{\hat{z}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau-\lambda_l}\right)^{\varepsilon} - 1$$

To make bunching (approximately) additively separable in the two components, consider a situation where $\Delta\tau+\lambda_l$ is small, such that $\log(z^*/\hat{z}) \approx \Delta z^*/\hat{z}$ where $\Delta z^* = z^* - \hat{z}$, and $\log(1-(\Delta\tau+\lambda_l)/(1-\tau)) \approx -(\Delta\tau+\lambda_l)/(1-\tau)$. Then

$$\frac{b}{\hat{z}} \approx \varepsilon\frac{\Delta\tau}{1-\tau} + \varepsilon\frac{\lambda_l}{1-\tau} \tag{2.34}$$

2.A.3.2.3 One-Sided Utility Notch

Bunching is

$$B = h_0(\hat{z})(\Delta z_+^* + \Delta z_-^*)$$

Denote by $f(\varepsilon, \frac{\delta}{c(\hat{z})})$ the solution for $\frac{\Delta z_-^*}{\hat{z}}$ implicitly defined by equation (2.12). Combining with equation (2.3), total bunching can be written as

$$\frac{b}{\hat{z}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^{\varepsilon} - 1 + f(\varepsilon, \frac{\delta}{c(\hat{z})}) \tag{2.35}$$

and when $\Delta\tau$ is small,

$$\frac{b}{\hat{z}} \approx \varepsilon\frac{\Delta\tau}{1-\tau} + f(\varepsilon, \frac{\delta}{c(\hat{z})}) \tag{2.36}$$

2.A.3.2.4 Two-Sided Utility Notch

Bunching is

$$B = h_0(\hat{z})(\Delta z_+^* + \Delta z_-^*)$$

Denote by $g(\varepsilon, \frac{\delta_2}{c(\hat{z})}, \frac{\Delta\tau}{1-\tau})$ the solution for $\frac{\Delta z_+^*}{\hat{z}}$ implicitly defined by equation (2.30). Combining with equation (2.12), total bunching can be written as

$$\frac{b}{\hat{z}} = f(\varepsilon, \frac{\delta_2}{c(\hat{z})}) + g(\varepsilon, \frac{\delta_2}{c(\hat{z})}, \frac{\Delta\tau}{1-\tau}) \tag{2.37}$$

Following the approximation argument made in section 3.2.2, when the density shift due to the budget set kink is small,

$$g(\varepsilon, \frac{\delta_2}{C(\hat{z})}, \frac{\Delta\tau}{1-\tau}) \approx \frac{z_+^* - z_+}{z_+} + \frac{z_+ - \hat{z}}{\hat{z}}$$

Denote $h(\varepsilon, \frac{\delta_2}{c(\hat{z})} \cdot \frac{\Delta\tau}{1-\tau})$ the implicit solution for $\frac{z_+ - \hat{z}}{\hat{z}}$ given by equation (2.29). Bunching is approximately

$$\frac{b}{\hat{z}} \approx \left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^{\varepsilon} - 1 + f(\varepsilon, \frac{\delta_2}{c(\hat{z})}) + h(\varepsilon, \frac{\delta_2}{c(\hat{z})} \cdot \frac{\Delta\tau}{1-\tau})$$

When $\Delta\tau$ is small,

$$\frac{b}{\hat{z}} \approx \varepsilon\frac{\Delta\tau}{1-\tau} + f(\varepsilon, \frac{\delta_2}{c(\hat{z})}) + h(\varepsilon, \frac{\delta_2}{c(\hat{z})} \cdot \frac{\Delta\tau}{1-\tau}) \tag{2.38}$$

2.A.4 Dynamic vs. Static Models of Retirement

2.A.4.1 A Life-Cycle Model of Retirement

Consider a life-cycle model of consumption for an individual with a fixed life span T who makes an extensive labor supply choice selecting a retirement age R. Assume that period utility is separable in consumption and leisure and that working at age t causes disutility α_t. Then lifetime utility at age zero[17] from retiring at R is

$$U_0(R) = \sum_{t=0}^{R-1} \beta^t \left(u(c_t) - \alpha_t\right) - \sum_{t=R}^{T} \beta^t u(c_t)$$

where β is the discount factor. The individual's lifetime budget constraint requires that lifetime consumption equals lifetime earnings, $C = Y(R)$ or

$$\sum_{t=0}^{T} \left(\frac{1}{1+r}\right)^t c_t = \sum_{t=0}^{R-1} \left(\frac{1}{1+r}\right)^t w_t + \sum_{t=R}^{T} \left(\frac{1}{1+r}\right)^t B(R)$$

where r is the interest rate, w_t is the wage at age t that reflects earnings capacity at that age and $B(R)$ is the pension benefit per period paid for retiring at age R.

2.A.4.2 Solution of the Dynamic Model

ASSUMPTION 1.1. Dynamic uncertainty in earnings capacity. *The worker is subject to a shock to earnings capacity w_t at every age t.*

This captures unexpected age-specific shocks such as to health or labor market opportunities and could for example be generated by a Markov process $w_{t+1} = \rho w_t + \varepsilon_{t+1}$. Note that disutility from work is assumed to follow a deterministic process throughout, i.e. all α_t are known based on α_0.[18]

Dynamic uncertainty forces the worker to re-evaluate the choice whether to retire at every age based on the new information arriving. Following Stock and Wise (1990) and Manoli and Weber (2016*a*), this problem can be solved by comparing the values of working and retiring at every age. The relevant lifetime utility is now utility at age t from retiring at R

$$U_t(R) = \sum_{s=t}^{T} \beta^{s-t} u(c_t) + \sum_{s=t}^{R-1} \beta^{s-t} \alpha_t$$

Making the decision whether to retire at age t, the value of retirement is

$$V^R(t, B(t)) = u(c_t^R(t)) + \beta V^R(t+1, B(t))$$

and the value of employment is

$$V^W(\Omega_t) = u(c_t^W) - \alpha_t + \beta E_t \left[V(\Omega_{t+1})\right]$$

[17] The starting age can be interpreted as the beginning of "old age" where retirement plays a role.

[18] The same retirement patterns could be generated by dynamic uncertainty in disutility from work and deterministc earnings capacity.

where $\Omega_t = \{t, B(t), w_t, \alpha_0\}$ is the set of state variables at age t and
$V(\Omega_{t+1}) = \max\{V^R(t+1, B(t+1)), V^W(\Omega_{t+1})\}$ is the value of next period's decision.

The worker's optimal choice follows a reservation value rule, retiring if her earnings capacity drops below a certain age-specific threshold $\bar{w}_t(\Omega_t)$, which is implicitly defined by

$$V^R(t, B(t)) = V^W(t, B(t), \bar{w}_t, \alpha_0)$$

or

$$u(c_t^W) - u(c_t^R(t)) + \beta OV_t = \alpha_t$$

where $OV_t = E_t\left[V(\Omega_{t+1})\right] - V^R(t+1, B(t))$ is the *option value* from working one more period. Hence, at the critical value $\bar{w}_t(\Omega_t)$ the benefits from working one more period, namely the gain in current consumption plus the option value equal the cost of postponing retirement in terms of disutility from work.

Notice that no assumption has been made so far about saving and borrowing behavior. At the one extreme, there can be full consumption smoothing so that there is no drop in consumption at retirement (other than an intended one due to the arrival of new information). At the other extreme, consumption could follow a hand-to-mouth pattern without saving or borrowing such that $c_t^W = w_t$ and $c_t^R(t) = B(R)$. Either case, including intermediate cases, can be accomodated by the dynamic model.

2.A.4.3 Derivation of the Static Model

ASSUMPTION 1.2. No dynamic uncertainty. *The time path of earnings capacity w_t is deterministic given the initial realization w_0.*

ASSUMPTION 2. Full consumption smoothing. *The worker is able to borrow and lend freely to maximize lifetime utility.*

Under assumption 1.2, the retirement decision can be made in period 0 as no additional information becomes available later on. Moreover, under assumption 2 consumption at each age t can be written as a function of lifetime wealth only. In particular, when $\beta = 1/(1+r)$, the individual wishes to consume the same amount at each age and

$$c_t = \frac{Y(R)}{\sum_{t=0}^{T}\left(\frac{1}{1+r}\right)^t} = \frac{C}{\sum_{t=0}^{T}\left(\frac{1}{1+r}\right)^t} \quad \forall t$$

Thus, the relevant lifetime utility at age 0 from retiring at R is

$$U_0(R) = u(c_t)\sum_{t=0}^{T}\beta^t - \sum_{t=0}^{R-1}\beta^t\alpha_t = U(C) - v(R)$$

where $U(C) := u(c_t)\sum_{t=0}^{T}\beta^t$ and $v(R) := \sum_{t=0}^{R-1}\beta^t\alpha_t$ are reduced-form utility from lifetime consumption and disutility from working until age R, respectively. $U(C)$ is increasing and concave in C if period utility $u(c_t)$ is increasing and concave in c_t. The properties of $v(R)$ depend on the α_t's. For example, a process $\alpha_{t+1} = \phi\alpha_t$ with $\phi > 1$ yields increasing and convex disutility $v(R)$.

The nonstochastic lifetime budget constraint is

$$C = \sum_{t=0}^{R-1} \left(\frac{1}{1+r}\right)^t w_t + \sum_{t=R}^{T} \left(\frac{1}{1+r}\right)^t B(R)$$

For further simplification, suppose the interest rate r is zero and the worker earns a constant period wage w. Then the constraint becomes

$$C = wR + (T-R)B(R)$$

The model derived in this section corresponds to the so-called "lifetime budget constraint" model of retirement suggested by Burtless (1986). While being based on the two strong assumptions specified above, its significant advantage is that retirement decisions can be treated in a way analogous to hours of work decisions in a standard labor supply model. In particular, the optimal date of retirement is characterized by the first-order condition[19]

$$\frac{v'(R)}{U'(C)} = \frac{dC}{dR}$$

where dC/dR is the marginal gain in lifetime consumption from postponing retirement given by the budget constraint.

2.A.5 Derivation of Estimation Equations

2.A.5.1 Structural Equation and Upper Bounds

The theoretical framework predicts that bunching at a pure budget set kink is given by equation (2.19), while bunching at a combined threshold is given by equation (2.24). Allowing for the possibility of reference dependence in utility from consumption and disutility from work at all statutory age types $stat \in (ERA, FRA, NRA)$, bunching at any discontinuity can be expressed as

$$\frac{b}{\hat{R}} = \left[\left(\frac{1-\tau}{1-\tau-\Delta\tau-\sum_s \lambda_l^s D^s}\right)^{\varepsilon} - 1\right] + \left[1 - \left(\frac{1}{1+\sum_s \lambda_c^s D^s}\right)^{\varepsilon}\right]$$

The equation can be used to structurally estimate ε along with the set of λ_c^s or λ_l^s across discontinuities.
To obtain upper bounds on λ_l^s, set all $\lambda_c^s = 0$ in the above equation, resulting in

$$\frac{b}{\hat{R}} = \left[\left(\frac{1-\tau}{1-\tau-\Delta\tau}\right)^{\varepsilon} - 1\right] + \left[1 - \left(\frac{1}{1+\sum_s \lambda_c^s D^s}\right)^{\varepsilon}\right]$$

Conversely, setting $\lambda_l^s = 0$ can yield upper bounds on λ_l^s:

$$\frac{b}{\hat{R}} = \left[\left(\frac{1-\tau}{1-\tau-\Delta\tau-\sum_s \lambda_l^s D^s}\right)^{\varepsilon} - 1\right]$$

[19] The FOC together with the corresponding second-order condition characterizes the optimum if the budget set is convex. However, certain features of pension schedules $B(R)$ may cause the budget set to be nonconvex.

2.A.5.2 Bunching from the Left vs. Right

Bunching from the right at any discontinuity is

$$\frac{b^+}{\hat{R}} = \left[\left(\frac{1-\tau}{1-\tau-\Delta\tau-\sum_s \lambda_l^s D^s}\right)^{\varepsilon} - 1\right]$$

and bunching from the left is

$$\frac{b^-}{\hat{R}} = \left[1 - \left(\frac{1}{1+\sum_s \lambda_c^s D^s}\right)^{\varepsilon}\right]$$

where $b = b_i^+ + b_i^-$. Hence, when the shares of bunching from the left and from the right are known, λ_c and λ_l can be separately identified. This idea is also implicitly used above, where upper bounds on λ_l^s is obtained by setting the bunching share from the left to zero, and the upper bound on λ_c^s is obtained by setting the bunching share from the right to its maximum. This maximum is given the amount of bunching from the right that would be implied by the budget set kink only, and is strictly less than 1 if there is a convex budget set kink at the threshold. The full range of possible parameter combinations can then be estimated by varying bunching shares from the right α_i between 0 and $\hat{\alpha}_i$ in $B_i^+ = \alpha_i B_i$, $B_i^- = (1-\alpha_i)B_i$, where $\hat{\alpha}_i$ is the maximum right bunching share at i. $\hat{\alpha}_i$ can be be calculated as

$$\hat{\alpha}_i = \frac{\left(\frac{1-\tau_i}{1-\tau_i-\Delta\tau_i}\right)^{\varepsilon} - 1}{b_i} \tag{2.39}$$

2.A.5.3 Measuring Bunching Shares from both Sides

2.A.5.3.1 Basic Approach

Bunching at the threshold must equal the total missing density from both sides:

$$B = \int_{R_{min}}^{\hat{R}} \left(h_0(R) - h(R)\right) dR + \int_{\hat{R}}^{R_{max}} \left(h_0(R) - h(R)\right) dR$$

where R_{min} and R_{max} bound the support of the density.

Measuring the true density shift over the full support is impossible in practice for two reasons. First, the shift $h_0(R) - h(R)$ may vary across R in an unknown way so that $h_0(R)$ cannot be measured for all R based on the observed density. Second, the full support of the counterfactual density may not be observed. Even if the full support of the actual density could observed, this does not necessarily correspond to the counterfactual support since some counterfactual density is predicted to "disappear" at the bounds because all individuals shift out a certain range.[20]

One solution to this problem is to approximate the true density shift by a constant shift over a certain range on each side. Denote by h_+ and h_- the observed density immediately to the right and left, respectively, of the threshold $\hat{R}$. Furthermore, denote by h_+^0 and h_-^0 the corresponding counterfactual density in the absence of the threshold. The approximation is

$$B \approx \left(h_-^0 - h_-\right)\left(\hat{R} - R^-\right) + \left(h_+^0 - h_+\right)\left(R^+ - \hat{R}\right)$$

where a constant density shift observed immediately to the left of the threshold over a range $[R^-, \hat{R}]$

[20] Besides, although theory predicts individuals responding to the threshold along the entire density, it is unclear in practice whether those far from the threshold perceive it in the same way as those closer.

approximates for the true shift on the left and a constant shift observed immediately to the right of $\hat{R}$ over $[\hat{R}, R^+]$ approximates for the shift on the right.

Assume also that the counterfactual density is continuous at $\hat{R}$ such that $h_+^0 = h_-^0 = h_0$. Then h_0 can be recovered as

$$h_0 \approx \frac{B + (\hat{R} - R^-)h_- + (R^+ - \hat{R})h_+}{R^+ - R^-}$$

From this, the implied bunching shares from both sides can be computed as $B^- = (h_0 - h_-)(\hat{R} - R^-)$ and $B^+ = (h_0 - h_+)(R^+ - \hat{R})$ since bunching from either side must be equal to the total density shift from that side. Finally, the structural parameters ε, λ_c and λ_l can be estimated by plugging the implied bunching shares into the equations in section 2.A.5.2.

2.A.5.3.2 Correcting for the Density Gradient

Approximating the true density shift by an observed vertical shift extrapolated over a given range comes with problems. To see this, consider a situation where the density is decreasing towards the threshold. In this case, the underlying density shift towards the threshold translates into an *upward* shift just beside the threshold combined with a large "disappearing" mass in the tail of the counterfactual density. More generally, the gradient of the density on each side determines to what extent an underlying horizontal density shift translates into an observed vertical density shift around the threshold.

To correct for this, relative bunching from both sides can be taken as the relative horizontal shift implied by the combination of the observed vertical shift and the gradient of the density beside the threshold. In particular, denoting by h'_- and h'_+ the slope of density to the left and to the right of threshold, the horizontal shift is given by the vertical shift multiplied with the inverse of this gradient. The true density shift is then approximated as

$$B \approx \frac{h'_- - h'_+}{2}\left[\frac{1}{h'_-}\left(h_-^0 - h_-\right)(\hat{R} - R^-) + \frac{1}{-h'_+}\left(h_+^0 - h_+\right)(R^+ - \hat{R})\right]$$

where $(h'_- - h'_+)/2$ is a rescaling factor used to transform the total horizontal shift back into the vertical dimension related to bunching. From this, the counterfactual density h_0 can be recovered as

$$h_0 \approx \frac{B + \frac{h'_- - h'_+}{2h'_-}h_-(\hat{R} - R^-) + \frac{h'_- - h'_+}{-2h'_+}h_+(R^+ - \hat{R})}{\frac{h'_- - h'_+}{2h'_-}(\hat{R} - R^-) + \frac{h'_- - h'_+}{-2h'_+}(R^+ - \hat{R})}$$

The implied bunching shares from both sides can then be computed as $B^- = \frac{h'_- - h'_+}{2h'_-}(h_0 - h_-)(\hat{R} - R^-)$ and $B^+ = \frac{h'_- - h'_+}{-2h'_+}(h_0 - h_+)(R^+ - \hat{R})$, and parameters can be estimated from the equations in section 2.A.5.2.

2.A.5.4 Linking Structural and Reduced-Form Estimation Equations

Consider now a situation where $\Delta\tau$ and λ_l^s are small, such that $\log(R_+^*/\hat{R}) \approx \Delta R_+^*/\hat{R}$ where $\Delta R_+^* = R_+^* - \hat{R}$, and $\log(1 - (\Delta\tau + \sum_s \lambda_l^s D^s)/(1-\tau)) \approx -(\Delta\tau + \sum_s \lambda_l^s D^s)/(1-\tau)$. Then

$$\frac{\Delta R_+^*}{\hat{R}} \approx \varepsilon\frac{\Delta\tau}{1-\tau} + \frac{\varepsilon}{1-\tau}\sum_s \lambda_l^s D^s$$

and

$$\frac{b}{\hat{R}} \approx \varepsilon \frac{\Delta\tau}{1-\tau} + \frac{\varepsilon}{1-\tau} \sum_s \lambda_l^s D^s + \left[1 - \left(\frac{1}{1+\sum_s \lambda_c^s D^s}\right)^{\varepsilon}\right]$$

Thus, ε can be estimated as the coefficient on kink size in the linear reduced-form specification according to equation (1.5), and the coefficients β^s yield a reduced-form estimate of the combined additional reference point bunching from both sides. More concretely,

$$\beta^s \approx \frac{\varepsilon}{1-\tau}\lambda_l^s + \left[1 - \left(\frac{1}{1+\lambda_c^s}\right)^{\varepsilon}\right]$$

If in addition λ_c^s are small, $\log(1+\sum_s \lambda_c^s D^s)) \approx \sum_s \lambda_c^s D^s$, and

$$\frac{b}{\hat{R}} \approx \varepsilon \frac{\Delta\tau}{1-\tau} + \frac{\varepsilon}{1-\tau} \sum_s \lambda_l^s D^s + \varepsilon \sum_s \lambda_c^s D^s$$

In this case, the statutory age coefficients simplify to

$$\beta^s \approx \frac{\varepsilon}{1-\tau}\lambda_l^s + \varepsilon\lambda_c^s$$

Chapter 3

Dual Tax Systems and Firms: Evidence from Brazil

3.1 Introduction

Presumptive tax regimes are prevalent in many countries (Bird et al. 2003). In such schemes, "regular" taxes are replaced by a simplified regime that somewhat approximates the regular tax liability. A common form of presumptive taxation are tax regimes tailored to small and medium enterprises (SME) due to compliance costs, and administrative constraints that make it costly for tax authorities to observe the tax base for enforcement (Slemrod and Gillitzer 2013). As a result, modern systems of firm taxation – characterized by some combination of payroll, valued-added, and corporate income taxes whose statutory incidence falls on firms – often exist alongside special tax regimes that rely on presumptive tax bases (e.g. a single turnover tax). This chapter uses novel administrative data on inter-firm trade linked to labor inputs from São Paulo, Brazil, to shed light on implications of such dual systems for firm growth, market competition and production decisions.

The precise way in which taxpayers are partitioned into different systems – turnover thresholds for value-added taxes (e.g, Keen and Mintz 2004) or payroll tax exemptions (e.g. Hsieh and Olken 2014) – also varies across countries (Kanbur and Keen 2014). In developing countries, particularly in Latin America and West Africa, firms below a turnover thresholds can often opt into a simplified regime where a single tax replaces a number of different taxes (Shome 2004). The case we focus on is a special tax regime in Brazil in which firms below a revenue threshold can choose between being part a presumptive turnover tax system (SIMPLES) or be taxed in the "regular" tax system.[1] They key difference between the SIMPLES system and the regular system is that turnover becomes a surrogate tax base for both the value-added tax (VAT) and payroll taxes for firms that opt into SIMPLES.

Conceptually, the effects of such dual tax systems on firms matter in at least four ways. First,

[1] Henceforth, we will refer to the non-presumptive system as the "regular" system. However, given the usual shape of firm size distributions with a large mass of small firms, firms in the "regular" systems are actually less numerous than firms in the presumptive system. See section 3.3 for descriptive statistics.

they could generate production inefficiencies through mis-allocation: a taxpayer may not choose the cheapest or best supplier as its choice may be affected by tax incentives. For instance, firms in the regular system can take VAT credit from suppliers in the regular system, but not from suppliers in the SIMPLES system. Thus, they may prefer to trade with firms that are also in the regular system. As a result, such a dual system may create partial segmentation of trade between firms in the regular tax regime and firms in the presumptive regime (De Paula and Scheinkman, 2010). Second, a size-based tax system generates notches that affect the firm size distribution (Dharmapala et al. 2011) by reducing incentives for firms to grow. Such distortions in taxpayers' behavior to ensure eligibility for the presumptive system may spill over to their trade partners (De Paula and Scheinkman, 2010). Third, the co-existence of taxpayers facing different tax burdens can also create an unleveled playing field for market competition. Finally, firms may change their input composition between intermediate inputs and labor inputs depending on whether inputs are deductible and whether payroll is taxed.

In this chapter, we systematically analyze the effects of dual tax systems on firms using anonymized administrative data from the tax authority of the state of São Paulo. The data includes yearly trade flows between firms from electronic invoices. One of the key advantages of the Brazilian electronic invoice data over other newly available datasets on firm-to-firm transactions is that all firms must use electronic invoicing irrespective of the tax regime they choose.[2] As a result, the data allows us to map the trade network of regular firms and SIMPLES firms in our sample, including when firms switch tax regimes. Moreover, the data is linked to administrative records containing the total number of employees and payroll. We exploit this data using a rich set of research designs, including a reform that changed the location of the revenue-based VAT threshold within the period of analysis.

We begin by showing a number of stylized facts on the firm size distribution, the composition of firms' inputs across the distribution, and the tax regime of firms top suppliers. We find clear bunching at the eligibility threshold for the presumptive tax system. Also, consistent with the tax incentives firms face, we find that firms in the presumptive tax system use relatively more labor input and source relatively more of their intermediate input from other firms in the presumptive tax system, which are not subject to the VAT. This leads to partial segmentation of the market between firms in the two systems.

Next, we use variation arising from firms' tax regime switches in order to show that there is a causal link between tax regimes and trade networks. This rules out an explanation of pure selection, where for example firms characteristics such as sector of activity that could determine both registration and trade networks. We implement an event study exploiting firms switching

[2]Usually, such data is available through VAT declarations where firms itemize inputs and outputs (e.g. in India and Uganda). Therefore, only transactions for which a VAT-registered firm is either a buyer or a supplier can be observed, and transactions between non-VAT firms are not covered by the data. In addition, because firms declare their inputs and outputs, the declaration of the same transaction is subject to mistakes and misreporting by the two parties. In the case of Brazil, the electronic invoicing covers all business-to-business transactions and each invoice has a unique key such that the same transaction cannot be reported differently by suppliers and buyers. The data thus mitigates both concerns of data censoring and misreporting.

into the SIMPLES regime and out of the SIMPLES regime to show how the VAT intensity of their inputs changes. This research design exploits the panel feature of our data, which allows us to control for fixed characteristics of firms through fixed effects. Our main finding is that firms start trading relatively more with firms in their new tax regime as soon as they switch regime.

In the second part of the chapter, we exploit a reform that increased the SIMPLES threshold from R$2.4M to R$3.6M. to study firm growth and market competition. First, we document how the firm size distribution shifts with the reform, including a clear movement of the location of bunching to the new threshold. Then, we provide evidence for the adverse effect of the coexistence of presumptive and regular tax regimes, and the preferential tax treatment given to SMEs (the option to choose the presumptive tax regime) on ineligible (larger) firms competing in the same market as eligible firms. We show that ineligible firms in sectors with more competition from SIMPLES firms before the reform grew relatively less after the reform compared to ineligible firms in sectors with less competition from SIMPLES firms.

Then we study the relationship between tax systems and production choices. To get at the causal effect from firms' tax regime to trade networks, we employ two research designs. First, we exploit the reform through a differences-in-differences strategy. Second, we study the impact of a supplier's change of tax regime on a firm's decision to purchase inputs from that supplier through an event analysis. For the second research design we restrict attention to firms that are a "small economy" to their supplier such that they do not influence the decision of their supplier to switch tax regimes. The results provide evidence that the respective tax regime of (potential) trade partners has a causal effect on their trade and thus on client's input choices. This leads to market segmentation between firms in regular and presumptive tax regimes.

Our results show that firms respond to the tax incentives they face: once labor inputs become taxed and intermediate inputs are deductible, firms change their production decisions away from inputs with higher effective marginal cost. Furthermore, our overall results are consistent with production distortions: we find (partial) segmentation in the network between regular and SIMPLES firms, i.e. firms trade relatively more with firms in the same tax regime. The degree of segmentation, however, is mitigated by the fact that firms are heavily dependent on key suppliers. In fact, most firms trade across tax regimes, and most firms have a regular firm among their main trade partners. Moreover, these distortions should be weighted against other key motivations for exemptions such as compliance costs that could be quite large for small firms below the threshold.

This chapter contributes to the literature by shedding new light on how tax systems interact with taxpayers sourcing decisions in the context of presumptive taxation in a developing country. Although the literature has emphasized the potential relevance of network effects (Liu et al. 2018; Pomeranz 2015; De Paula and Scheinkman 2010), there is little empirical evidence of chain effects using micro-data on firm trade flows.[3] The chapter also contributes to a broader literature on misallocation (e.g. Hsieh and Klenow 2009), size-based regulation and taxation (e.g. Garicano

[3]Concurrent to this chapter, there is some work in progress using data from India (Gadenne et al. 2018, Rios and Setharam 2018).

et al. 2016, Monteiro and Assunção 2012, Boonzaaier et al. 2017, Best et al. 2015), and a growing literature documenting firm responses to VAT thresholds through avoidance (e.g. Onji 2009), evasion (e.g. Asatryan and Peichl 2017), and real desincentives to grow (e.g. Harju et al. 2015).

The remainder of this chapter is organized as follows. Section 3.2 describes the institutional setting and data, section 3.3 presents stylized empirical facts, section 3.4 shows the effects of dual tax systems on growth, competition and input choice, and section 3.5 concludes.

3.2 Institutional Background and Data

3.2.1 Institutional Background

Brazil is a federal country and the state of São Paulo, the setting of our empirical analysis, is the largest of the 27 Brazilian states, both in terms of population and economic activity. It has about 42 million inhabitants and it accounts for about 34% of the Brazilian GDP.Firms in Brazil are subject to taxes from federal, state and municipal governments. In particular, they are subject to a federal-level payroll tax (INSS; the tax rate is about 20% for all firms), to a state-level value-added tax (ICMS; see below),[4] and several municipal- and federal-level turnover taxes (PIS, COFINS, CLSS, IRPJ, ISS; implying a tax rate of at least 6% for the sale of goods and 13.3% for services).[5] We denote this tax system as the "regular system" and firms subject to it as "regular firms."

Like many countries, Brazil has a special tax legislation for SME. Firms that report total revenue below a threshold are eligible to opt into a presumptive tax system called SIMPLES. Eligible firms must choose whether to opt for SIMPLES at the beginning of the year. SIMPLES firms are subject to taxes in a presumptive way where the tax base is turnover and collection is unified in a single tax return.[6] The tax applies to their gross revenue during the year at an average tax rate that increases from 4% to 12% in the tax base.[7]

Until 2011, firms with yearly gross revenue (across all their establishments) below R$ 2.4 million in the previous year were eligible for SIMPLES (the exchange rate was about R$2 $\simeq$ US$1 in 2012); in 2012, the threshold increased to R$ 3.6 million. As a result, most firms can opt for a presumptive tax regime in Brazil.[8] If their yearly revenue exceeds the threshold by the end of the year, they

[4]Manufacturing firms are subject to an additional federal tax on value added (IPI).

[5]These numbers are referring to firms on the system called "Lucro Presumido". For very large firms – only about 3% of formal Brazilian firms according to Receita Federal – the federal taxes becomes value added (PIS, COFINS, CLSS) or corporate income (IRPJ) taxes ("Lucro Real"). The municipal tax on services (ISS with a tax rate ranging from 2% to 5% depending on the type of services) remains a turnover tax for those firms. For confidentiality reasons, our data do not allow us to study these firms.

[6]This type of single tax that replaces many taxes - "impuesto unico" or "monotributo" - are common in Latin America (Shome 2004).

[7]The tax rate that applies to the turnover depends on their economic activity as it aims to replace the tax base in accordance to what firms would otherwise be liable for. i.e. firms in the service sector that would be liable for the service tax (ISS) face a different rate than firms in commerce that would not be liable for ISS but would be liable for the value added tax (ICMS).

[8]We present the average tax rate schedule for firms in the commerce sector in the appendix. Some firms are ineligible for SIMPLES, independently of their gross revenue. This is the case, for instance, for firms with foreign owners or foreign capital.

become ineligible in the following year.[9]

It is important to mention that there is also a sizable number of firms in the informal economy that are exempt *de facto* from all taxes on their economic activity.[10] Studying these firms is often challenging due to data constraints. Because informal firms are relatively small, they would likely be eligible for SIMPLES even if they were registered with the tax authority. For the purpose of our study, "SIMPLES firms" are likely to be a good approximation to these informal firms in terms of the sourcing incentives they face, since any tax on inputs cannot be recovered through deductions and payroll is *de facto* not taxed.

The set of firms we focus on are in the commerce sector, so the most relevant changes – besides modifying firms' overall tax and compliance burden – from the regular system to SIMPLES is the replacement of the tax on value added (ICMS) and the tax on payroll (INSS) by a turnover tax. This will likely affect the type of firms opting for the presumptive tax regime, as well as their production choices. The effective marginal cost of a unit of labor input is lower for SIMPLES firms, as they are not subject to the payroll tax. In contrast, the effective marginal cost of a unit of intermediate input may be higher for SIMPLES firms depending on the tax regime of their supplier. This is because SIMPLES firms cannot deduct any state-level VAT charged on their purchases.

Firms subject to the ICMS, which is a VAT based on the credit-invoice method, are required to charge a tax on all their sales of taxable goods and services ("outputs"). The most common rate is 18% in São Paulo; imports are taxed; exports are exempt; and a reduced rate of 7% or a higher rate of 25% applies in some cases.[11] Firms can obtain a tax credit for the VAT charged on their purchases (inputs) as long as they are themselves subject to the ICMS.[12] As a result, the marginal cost of a unit of intermediate input purchased from a regular firm is higher for a SIMPLES firm than for another regular firm. This may lead to some market segmentation as SIMPLES firms have a greater incentive to buy their intermediate inputs from other SIMPLES firms (De Paula and Scheinkman, 2010). The inability to deduct taxes charged on intermediate inputs also incentivizes SIMPLES firms to rely to a greater extent on own production, i.e. labor input.

Finally, and importantly for the data used in this project, all firms are required to issue electronic invoices for business-to-business transactions in São Paulo, including SIMPLES firms. This innovation is part of a nationwide effort called SPED (Sistema Público de Escrituração Digital)

[9] If a firm's revenue exceeds the threshold within the calendar year, they become ineligible for SIMPLES in the following year. The tax rate applied to the amount that exceeds the threshold is 20% higher than the rate applied to the revenue up to the threshold. If the revenue exceeds 20% of the threshold within the calendar year, the firm has to leave the SIMPLES regime in the following month.

[10] The informal sector is relatively large in São Paulo: an estimated 2.1 million firms are informal (SEBRAE, 2003); over 90% of them are self-employed workers with no employees (SEBRAE, 2003).

[11] ICMS is often used as an industrial policy instrument. For instance, states create targeted exemptions for specific firms or sectors.

[12] The reporting period is monthly; every month, firms must calculate their actual tax liabilities as the difference between the total VAT charged on their output invoices and on their input invoices. Note that Brazilian states often use ICMS withholding (substituição tributària). In that case, one taxpayer withholds the ICMS for all other taxpayers along the supply chain, but firms must file their taxes even if all payments have already been made on their behalf. In principle, if there are differences between what was withheld and what is due by the firm, there could be additional payments or reimbursement.

to modernize and integrate tax systems in the country. SPED was created in 2003 and many innovations around tax filing and electronic invoicing were proposed since. Since 2011, electronic invoices (NF-e) cover all business-to-business transactions in the São Paulo.[13] Firms are required to use specific invoicing machines that issue standardized invoices and that are hard to temper with. These machines prevent firms from issuing different invoices for a given transaction for the selling firm (e.g. an invoice with a low value) and the buying firm (e.g. an invoice with a high value), a common mechanism of VAT fraud in other countries. They also send the information for each invoice to the tax authority in real time through the mobile phone network. The tax authority thus possesses complete records of all business-to-business transactions carried out in accordance with the legal framework. Although Brazil was an early adopter of the technology, other countries have mandated the use of comparable electronic invoicing machines in recent years (Steenbergen, 2017).

3.2.2 Data

Our empirical analysis relies on anonymized administrative data made available for research purposes by the tax authority of the state of São Paulo (SEFAZ/SP).

3.2.2.1 Electronic Invoices as a Source of Inter-Firm Trade Data

A key dataset used in this chapter provides inter-firm trade data based on the universe of electronic invoices for business-to-business transactions in São Paulo. These electronic invoices present unique advantages for our purpose compared to other sources of transaction data. First, they capture all legal transactions between formal firms, including between SIMPLES firms. In contrast, other countries that require firms to report their transactions with each trade partner to the tax authority usually apply this requirement to VAT-registered firms only, as in Uganda or India (Almunia et al. 2017; Gadenne et al. 2018). Administrative transaction data based on tax withholding schemes also only record transactions involving withholding agents, such as government institutions (Brockmeyer and Hernandez 2018).

Second, these electronic invoices capture all legal transactions irrespective of the payment mode or the transaction value. Transaction data based on tax withholding schemes involving credit-card companies only record transactions paid by credit card (Brockmeyer and Hernandez 2018). In countries where VAT-registered firms are required to report their transactions with each trade partner to the tax authority, they are often exempt from this requirement if the total transaction value falls below some specific amount, as in Costa Rica or India (Brockmeyer and Hernandez 2018; Gadenne et al. 2018). Transaction data from firm surveys typically ask only about transactions with a few top partners.

Third, because of the electronic invoicing, there is a unique key for each transaction such that the two sides of the transaction are unable to report different amounts. In countries that require

[13] In 2015, electronic billing machines for final consumer sales (NF-c) were also introduced in the state of São Paulo.

firms to report transactions with each trade partner to the tax authority without such a technology, transaction values reported separately by the supplier and the client have been shown to be plagued by very large discrepancies (see Almunia et al. 2017 for the case of Uganda and Brockmeyer and Hernandez 2018 for the case of Costa Rica). Such discrepancies can arise from errors due to the manual reporting of the transactions or from the strategic *unilateral* misreporting of these transactions (i.e. to minimize tax liabilities). The electronic invoices in Brazil avoid both issues.

Transaction data based on these electronic invoices thus present useful features in order to measure, e.g. a firm's trade flow with another specific firm or its total trade flow with firms in a specific tax regime. Yet, they also come with some limitations. First, some information on these invoices is not yet fully harmonized, such as product codes and units. Therefore, they are not yet suitable to measure prices and study incidence questions. Second, by construction, these invoices do not record any transaction with informal firms. Third, they do not record any transaction that formal firms agree to carry illegally, i.e. without using the electronic invoicing machine. Even though the data avoid the issue of unilateral misreporting, this type of *collusive* misreporting may still affect the trade actually reported by formal firms in the data. We discuss this issue in our analysis below, as compliance incentives may affect our measurement of the degree of market segmentation between SIMPLES firms and regular firms in the São Paulo economy.

3.2.2.2 Inter-firm trade Data Made Available

Due to logistical constraints given by the sheer size of the data and in order to protect the confidentiality of firms' information, the specific transaction data made available for this research were constructed as follows.

Origin Sample. Our starting point are all tax-registered firms in the state of São Paulo with at least one wholesale establishment in the state. Wholesalers are interesting to study the effect of tax regimes on input choice decisions given their central position in supply chains. Hereafter, these firms are denoted as "origin" firms; all their establishments located in São Paulo (including non-wholesale establishments) are denoted the "origin" establishments; and together they form the "origin" sample.

All the electronic invoices involving origin establishments, as supplier or client, have been extracted from the universe of the electronic invoices involving at least one São Paulo establishment. Because of logistical constraints, the information is aggregated by pair of supplier and client at the yearly level. Next, for confidentiality reasons, the data are anonymized by replacing firms' tax ID by a unique scrambled ID number. Firms that could potentially be identifiable despite the anonymization (e.g. very large firms) were aggregated into one observation without individualized information.

Figure 3.1 illustrates the detailed transaction data available for this project using a random sample of observations in one year. The overall dataset covers the years from 2011 to 2017. It includes 121,202 unique origin establishments belonging to 88,359 unique origin firms; 17,609,284 input transaction-years, in which origin establishments make purchases from 631,450 unique suppli-

ers (567,355 unique firms); and 96,689,168 output transaction-years, in which origin establishments make sales to 4,393,038 unique clients (4,062,897 unique firms).[14] For each supplier-client pair in each year, we know the (scrambled) firm and establishment ID of the supplier and the client, the tax regime of the supplier (i.e. SIMPLES or regular), the total value of all invoices, the value subject to VAT, the VAT charged, and an indicator for whether the trade partner is located in São Paulo. For international transactions, we observe the total value of exports, the total value of imports, and the VAT charged on imports.

Full Sample. A second dataset has been constructed based on the electronic invoices from 2011 to 2017 to provide aggregate trade information for all establishments in São Paulo belonging to supplier and client firms of origin establishments that are taxpayers in the state. The resulting sample includes 1,674,363 unique firms (including those from the origin sample). Hereafter, we refer to this sample as the "full" sample (it includes many retailers making purchases from origin establishments). For each establishment, we know the total value of their input transactions and the total VAT charged in each year, broken down by type of supplier: SIMPLES firms, regular firms, and imports. Firms that could potentially be identifiable despite the anonymization were again aggregated as in the origin sample.

3.2.2.3 Other Data

Additional information on all establishments in the full sample was made available for this research. From the state tax registry, we have firm-level information on the year of registration and the tax regime in each year from 2008 to 2017. From the state-level VAT declarations and tax declarations for the SIMPLES regime, we have establishment-level information on the yearly value of exports and the yearly revenue in each year from 2008 to 2017. This allows us to capture sales to final consumers. Moreover, we have a variable capturing the anonymized 5-digit sector of the establishment. For confidentiality reasons, the sector of activity was scrambled such that we only observe the overall position in the supply chain (retail, wholesale, manufacturing) and whether two establishments are in the same 5-digit sector, but not which sector they belong to. Finally, from the Brazilian matched employee-employer data (RAIS), we have establishment-level information on the number of employees and the total payroll in each month from 2011 to 2016.

3.2.2.4 Data Construction

In our analysis below, we aggregate all the establishment-level information at the firm-year level; we mostly use the origin sample given the more detailed information available for firms in this sample; and we mostly focus on the period 2011-2016 for which we have the most complete set of variables available.

One point to make before delving into the analysis is that the data is best suited to study input

[14] This number is quite large because it includes trade partners that are outside São Paulo. Although we do observe a scrambled ID for these firms, they are not part of the "full sample" we describe below because they are outside São Paulo.

transactions as opposed to output transactions. For input transactions, we know the tax regime of the trade partner (the supplier) from the transaction data directly. In contrast, for output transactions, we know the tax regime of the trade partner (the client) by matching the partner's ID in the transaction and the registry data. As a result, we only know the tax regime of a client if it is registered as a taxpayer in the state of São Paulo. Consequently, we focus on input transactions when studying segmentation and the effect of tax systems on inter-firm trade.

3.3 Stylized Facts

As discussed in the introduction, the coexistence of presumptive and regular tax systems can have important implications for firm growth, market competition, and production decisions. In this section, we present descriptive statistics and key stylized facts consistent with two of these effects. We show that there is clear bunching below the SIMPLES threshold, indicating that some firms reduce their revenue to remain eligible for the presumptive tax system. We also show that firms in the presumptive tax system use relatively more labor input and source relatively more of their intermediate input from other firms in the presumptive tax system, leading to partial segmentation of the market between firms in the two systems.

3.3.1 Bunching and Choice of Tax Regime

Figure 3.2 shows that there is large bunching in the revenue distribution below the SIMPLES threshold, but that many firms eligible for SIMPLES do not choose this presumptive tax regime. These patterns are presented for origin firms between 2012 and 2016, when the SIMPLES threshold was set at R$ 3.6 million. Panel (a) displays the overall revenue distribution for all firm-year observations with positive revenue over the period (censored at R$5 million). It shows that most firms fall below the SIMPLES threshold (the vertical line) and are thus eligible for the presumptive tax regime. This is the case for 83.1% of firms in the origin sample; this number is even larger in the full sample (95.8%), as it includes many (smaller) retailers. Panel (b) displays the same revenue distribution but we zoom in around the SIMPLES threshold. It shows that there is clear bunching below the threshold, indicating that many firms value the option to choose the presumptive tax regime in the following year, which distorts the revenue distribution.

Panel (c) displays the share of firm-year observations in the regular system by firms' revenue levels in the previous year ($t-1$) and in the concurrent year (t), separately. The first graph shows that the SIMPLES threshold is binding: almost all firms above the threshold in $t-1$ are in the regular system in the following year. It also shows that many firms below the threshold, and thus eligible for SIMPLES, do not choose the presumptive tax regime in the following year.[15] Overall, this is the case for 28.8% of firms in the origin sample overall and for more than 60% in the vicinity of the threshold. Together with panel (a), the graph shows that most firms in São Paulo are in

[15] Liu et al. (2018) make a related point that the combination of bunching and voluntary registration in the context of the VAT in the UK is consistent with market segmentation between VAT-registered and Non-VAT firms.

the presumptive tax regime, namely 58.3% of the firms in the origin sample (this figure reaches 88.8% of firms in the full sample). The second graph in panel (c) shows that the share of firms in the regular system is smaller to the right of the threshold if we consider the revenue distribution in the concurrent year, after a firm chooses its tax regime. This is because firms crossing the threshold only become ineligible for SIMPLES in the following year. The two graphs in panel (c) are almost identical to the left of the threshold, as most firms choose the same tax regime from year to year. In fact, the aggregate patterns presented below are similar whether we use the revenue distribution in the year before or after a firm chooses its tax regime, so we present them using the year-t distribution only.

A striking pattern in panel (c) is that the share of firms in the regular system, which is generally increasing in revenue below the threshold, drops just below the threshold before increasing discontinuously at the threshold. This pattern suggests that the bunching in panel (b) is mostly driven by SIMPLES firms. This is confirmed in panel (d), which displays the revenue distribution around the SIMPLES threshold by tax regime. The bunching is entirely driven by SIMPLES firms, showing that they strongly value the option to remain in SIMPLES in the following year. In contrast, the revenue distribution is completely smooth at the threshold among regular firms. As the option value of choosing tax regime in the following year is likely positive (and likely strictly positive for some firms), this suggests that switching regimes involves non-negligible fixed costs, as often argued in the literature (Harju et al., 2015).

3.3.2 Descriptive Statistics and Input Choices of Firms in the Two Tax Regimes

Table 3.1 presents descriptive statistics comparing firms in the presumptive and regular tax regimes. Column (1) considers all firm-year observations with positive revenue in the origin sample from 2012 to 2016 (the same observations as in figure 3.2). Column (2)-(4) consider firms with revenue levels above the SIMPLES threshold, regular firms with revenue levels below the threshold, and SIMPLES firms with revenue levels below the threshold, respectively.[16] As discussed above, almost all firms above the threshold are regular firms, but most firms are below the threshold and are SIMPLES firms, although a sizable share of firms eligible for the presumptive tax regime opt for the regular tax regime. SIMPLES firms account for most of the firm-year observations, but they only account for 6% of the total revenue across all firms in the sample, as regular firms have higher revenue levels (even below the threshold). For instance, firms above the threshold account for only 16.9% of the firm-year observations but for 89.9% of the total revenue.

Regular firms tend to have higher input levels – the sum of intermediate input, i.e. input purchased from other firms, and labor input, i.e. total gross wages paid to employees – than SIMPLES firms, even taking into account their higher revenue levels. This is also shown graphically in panel (a) of figure 3.3. It displays the mean log(input) by revenue bins for regular and SIMPLES firms, separately. Input levels are higher for regular firms than for SIMPLES firms at all revenue

[16] The descriptive statistics are similar if we use revenue levels in the previous year to categorize firms in the four groups.

levels.

Regular and SIMPLES firms differ in the composition of their input as well. Table 3.1 and figure 3.3 show that regular firms use relatively more intermediate input and thus relatively less labor input, than SIMPLES firms (see panel b). Almost all firms use some intermediate input but the share of intermediate input out of all inputs is higher among regular firms (between 80% and 95%) than among SIMPLES firms (between 75% and 80%) at all revenue levels. Table 3.1 and figure 3.3 also show that regular firms source relatively more of their intermediate input from suppliers subject to the VAT, namely other regular firms and foreign firms; while SIMPLES firms source relatively more of their intermediate input from other SIMPLES firms (not subject to the VAT; see panel c).[17] The share of intermediate input subject to the VAT is higher among regular firms (between 80% and 95%) than among SIMPLES firms (between 70% and 80%) at all revenue levels.

This provides some first evidence of market segmentation between firms in the two tax regimes. Moreover, we find similar patterns when we focus on top suppliers. This is important as inter-firm trade is typically very concentrated, with a few trade partners playing a critical role. Panels (e) and (f) of figure 3.3 display the average share of origin firms' intermediate inputs by the rank of their top 10 suppliers (for firms with at least 10 suppliers) for regular firms and SIMPLES firms below the SIMPLES threshold. The top-1 supplier accounts for about 35% of firms' intermediate inputs in the two groups, but that share drops to 15% for the second most important supplier and continues to fall for the following suppliers, showing the concentration of inter-firm trade. Yet, panels (e) and (f) also show that the same degree of market segmentation between regular firms and SIMPLES firms appear among top suppliers. The top-1 supplier of regular firms is a regular firm in more than 90% of the observations, but that figure drops by about 11pp among SIMPLES firms, and the difference persists across all supplier ranks.

Finally, table 3.1 shows that that the share of firms with positive labor input is higher among SIMPLES firms than among regular firms below the threshold. Moreover, conditional on having some positive labor input, panel (d) of figure 3.3 shows that SIMPLES firm use more labor input than regular firms at most revenue levels (above the very bottom of the firm size distribution), and not only as a share of all inputs.

The differences in the input composition between regular firms and SIMPLES firms are consistent with the differential tax treatment of inputs in the two regimes. The marginal cost of a unit of labor input is lower for SIMPLES firms, while the marginal cost of a unit of intermediate input from a supplier subject to the VAT is higher for SIMPLES firms, as they cannot deduct the VAT paid on these purchases. Interestingly, there is no evidence that input choices differ for bunching firms, as the patterns in figure 3.3 are not systematically different just to the left of the threshold among SIMPLES firms.

Table 3.2 presents some multivariate regressions reiterating these correlations between tax

[17] We classified trades with firms that were aggregated due to confidentiality concerns as being VAT since the exclusion criterion of identifiability imples that the vast majority are too large to be eligible for SIMPLES.

regime and input choices. The sample is composed of all firm-year observations in the origin sample with positive revenue, positive intermediate input, and revenue levels below the SIMPLES threshold such that they are eligible for SIMPLES. In column (1), we simply regress an indicator for opting for the regular tax system on log(revenue), controlling for year fixed effects. Larger firms are more likely to choose the regular tax system, which is consistent with the increasing average tax rate on revenue in the presumptive tax regime. In column (2), we add controls for log(input), the share of intermediate input (out of all inputs) and the share of intermediate input subject to VAT. Firms with a higher input level, a higher share of intermediate input, and a higher share of intermediate input subject to VAT are more likely to be part of the regular tax regime even when we consider all these variables simultaneously. The correlations are similar if we include sector fixed effects (column 3) and if we restrict attention to the balanced panel of firms observed in all years (column 4). Finally, column (5) shows that the sizes of the coefficients drop but they remain significant even with firm fixed effects, thus exploiting firms' change of tax regime over time.

3.3.3 Tax Regime and Input Choices: Event Analysis around Firms' Tax Regime Switches

The previous subsection documented systematic relationships between firms' tax regime and input choices. These relationships could be driven by causal links in both directions. On the one hand, firms differing in their production functions may opt for different tax regimes. For instance, firms with labor-intensive production functions have an incentive to opt for the presumptive tax regime in order to avoid the payroll tax. Firms with many potential trade partners in the regular tax regime may opt for the regular tax regime in order to be able to deduct the VAT charged on the intermediate input purchased from these firms. On the other hand, tax regimes may have direct effects on input choices. Firms opting for the presumptive tax regime have an incentive to rely more heavily on own production using labor input. Moreover, conditional on purchasing intermediate inputs, they have an incentive to seek suppliers in the presumptive tax regime in order to avoid paying for the VAT.

It is important to note that some of these relationships could be driven by reporting effects, namely by differential compliance incentives in the two regime and the difference between actual input choices (e.g. actual input transactions) and reported input choices (e.g. input transactions observed in the data). For instance, firms in the regular tax regime have an incentive to report input transactions with other firms in the regular tax regime accurately, such that they can deduct the VAT charged on their purchases; SIMPLES firms do not have a similar incentive. At the same time, the observed degree of market segmentation could be an underestimate: SIMPLES firms have limited incentives to report any input transaction accurately, while they have more incentives to source from SIMPLES (or even informal) suppliers who may in turn want to under-report their sales.

In this subsection, we provide a more careful analysis of tax regime changes and consider transitions in and out of the presumptive tax regime separately. This allows us to document

patterns that are not easily explained by reporting effects and that cannot be simply inferred from the results in column (5) of table 3.2. In particular, we conduct an event analysis to study how input choices compare before and after a tax regime switch. Because we use variation within firms over time, correlations between tax regimes and trade patterns in this analysis cannot be due to fixed firm characteristics.

We find that firms change their input choices as soon as they switch tax regime. There are three candidate explanations for this result. First, there could be time-variant omitted variables unrelated to a firm's potential input choices (e.g. to its potential trade network) driving both tax regime switches and changes in input choices. The most obvious concern is that firms that switch into (out of) the regular tax regime are growing more (less) in previous years than the average firm. Yet, we show that controlling for pre-trends in revenue and input does not affect our results. Although this does not remove all potential sources of omitted variable bias unrelated to a firm's (potential) input choices, it indicates that such a bias does not seem to drive much of our results. In that case, there remains two explanations that imply causal links between tax regimes and potential input choices. On the one hand, firms may change tax regime because they experience or expect changes in input choices, such as changes in the composition of their trade network. On the other hand, firms may also change their input choices because they change tax regime. We interpret our findings here as first suggestive evidence for these causal links, in any one of the two directions. We provide evidence consistent with each of these directions separately in a later section.

Empirical strategy. We begin by selecting origin firms that switch tax regime between 2013 and 2015, such that we can follow them for two years before and after the tax regime switch (the year of the switch is our first "post-event" year as the tax regime choice is made on January 1st), over the period for which we have all the variables available (2011-2016). We then keep a balanced panel of firms with positive revenue and intermediate input for the four years around the event. Additionally, we restrict attention to firms that are in a same tax regime in the two years before the switch and in a same tax regime in the two years after the switch. We thus focus on the clearer "persistent" changes in tax regime, but results are similar if we allow firms to change tax regime again after the first switch. This sample constitutes our "treatment" sample. We also construct a "control" sample to create a counterfactual and to show that our results are robust to controlling for pre-trends in revenue and intermediate input. We randomly assign a "placebo" event year between 2013 and 2015 to all firms observed with positive revenue and intermediate input for a 4-year window around that year *in the same tax regime*. We end up with four groups of firms, two control groups and two treatment groups: (i) firms that are always in the regular tax regime ("Stayers regular regime"; 14,270 firms), (ii) firms that are always in the presumptive tax regime ("Stayers presumptive regime"; 18,329 firms), (iii) firms that switch out of the presumptive tax regime into the regular tax regime ("Presumptive to Regular"; 842 firms), and (iv) firms that switch out of the regular tax regime into the presumptive tax regime ("Regular to Presumptive";

1042 firms).[18]

We present raw empirical patterns in relevant variables for these four groups graphically. We then quantify changes in an outcome y for our treatment groups, by comparing firms in each treatment group to the control group of firms that remains in their initial tax regime using difference-in-differences specifications:

$$y_{i,k,t} = \alpha_i + \beta_k + \gamma_t + \delta_k \cdot Treat_i + \psi_k \cdot \Delta^{pre} Revenue_i + \phi_k \cdot \Delta^{pre} Input_i + \varepsilon_{i,k,t} \quad (3.1)$$

$$y_{i,k,t} = \alpha_i + \beta_k + \gamma_t + \delta \cdot Treat_i \cdot After_k + \psi_k \cdot \Delta^{pre} Revenue_i + \phi_k \cdot \Delta^{pre} Input_i + \varepsilon_{i,k,t} \quad (3.2)$$

where α_i, β_k, and γ_t are fixed effects for each firm i, each event year k, and each calendar year $t = 2011, ..., 2016$. Event years are normalized such that we have $k = -2, ..., 1$ with $k = 0$ as the year of the tax regime switch. The coefficients ψ_k and ϕ_k control for possible changes in the outcome in each event year related to pre-trends (from $k = -2$ to $k = -1$) in revenue and intermediate input (in logs), respectively. The coefficients δ_k in equation (3.1) capture differential changes in the treatment vs. control groups in each year relative to base year $k = -1$; the coefficient δ in equation (3.2) summarizes the difference-in-differences by capturing the change in the treatment vs. control groups after vs. before the tax regime switch. Standard errors are clustered by firm. We run separate regressions for firms initially in the regular tax regime and initially in the presumptive tax regime.

Results. Figure 3.4 displays raw patterns in the four groups of firms around the tax regime switch (we only take out calendar-year effects from the raw data). Panel (a) shows that switchers from the presumptive to the regular tax regime were growing relatively more, and switchers from the regular to the presumptive tax regime were growing relatively less prior to the tax regime switch compared to the two control groups. This is the differential pre-trend mentioned above, which we control for in the difference-in-differences specifications. After the tax regime switch, firms in both treatment groups seem to grow relatively more compared to firms in the control groups. Panel (b) shows raw patterns for one variable capturing input choices, namely the share of intermediate input subject to VAT. It increases discontinuously for switchers into the regular regime, but it decreases discontinuously for switchers into the presumptive regime. Differences in pre-trends are also less severe. This is evidence that the systematic correlations between tax regime and input choices, specifically the composition of the trade network, are not due to fixed firm characteristics.

Figure 3.5 displays the estimated $\widehat{\delta_k}$ with their 95% confidence intervals from running separate regressions for firms initially in the regular tax regime and firms initially in the presumptive tax regime using the specification in equation (3.1). The estimated $\widehat{\delta}$ from using the specification in equation (3.2) are reported in the notes below each graph.

Panel (a) shows that, once we control for pre-trends, revenue levels increase in both treatment groups after the tax regime switch compared to the control groups. This is consistent with the raw patterns in figure 3.4a. Panel (b) shows that input levels increase even more than revenue levels

[18] This sampling strategy implies that no firm belongs to more than one group.

for switchers into the regular regime. In contrast, revenue levels increase without an increase in input levels for switchers into the presumptive tax regime. The increase in revenue levels without an increase in input levels for switchers into the presumptive tax regime is consistent with the differential incentive to rely more on own production or informal inputs for firms in the presumptive tax regime. Panel (c) shows that the share of intermediate input out of all inputs increases discontinuously after a firm switches to the regular tax regime but decreases discontinuously for tax regime switch in the other direction. The size of the difference-in-differences estimates are consistent with the estimates in column (5) of table 3.2. Together, panels (d), (h), and (i) show that firms switching to the regular tax regime increase their intermediate input and decrease their labor input, while firms switching to the presumptive tax regime increase their labor input without decreasing their intermediate input.

Finally, panel (e) shows that there is no differential pre-trend in the share of intermediate input subject to VAT once we control for pre-trends in revenue and input. Yet, the level of this variable still increases discontinuously after a firm switches to the regular tax regime but decreases discontinuously for tax regime switch in the other direction. This pattern shows that the market segmentation between firms in the regular and presumptive tax regimes is not entirely due to fixed firm characteristics. Panels (f) and (g) further investigate the patterns in panel (e), by displaying estimates for intermediate input subject to VAT and intermediate input not subject to VAT, separately. The levels of both types of intermediate inputs increase for switchers into the regular tax regime, but the level of intermediate input subject to VAT increases relatively more. This is consistent with pure reporting effects due to the different compliance incentives when joining the regular tax regime, including participating in the VAT system. However, the pattern for switchers into the presumptive tax regime is not consistent with pure reporting effects. While the level of intermediate input subject to VAT decreases for these firms, the level of intermediate input not subject to VAT actually increases. The change in compliance incentives when firms switch into the presumptive tax regime would not predict a relative increase in reported inputs from SIMPLES firms. Therefore, even though our measure of market segmentation may still be biased by compliance incentives, we interpret the segmentation that we observe in the data as capturing real market segmentation between firms in the regular and presumptive tax regimes.

3.4 Coexistence of presumptive and regular tax regimes: effects on firm growth, market competition, and production decisions

We now provide evidence on the effects of the coexistence of presumptive and regular tax regimes on firm growth, market competition, and production decisions.

3.4.1 Firm growth

We begin with the first of these three effects. The clear bunching of firms below the SIMPLES threshold in figure 3.2 already provides strong evidence that the option of the presumptive tax

regime distorts firm size, at least around the threshold. In this section, we provide additional evidence by exploiting the reform that increased the SIMPLES threshold from R$2.4M to R$3.6M in 2012. We show the impact of the reform on the firm size distribution in figure 3.6.

Panel (a) displays the firm size distribution of origin firms around the pre- and post-reform thresholds in 2010, 2012, and 2014. It shows clear bunching below the pre-reform threshold in 2010. This feature of the firm size distribution completely disappears by 2012, however, with an increase in the number of firms concentrated between the pre- and post-reform thresholds.[19] This is even more visible in panel (b). It displays the number of firms in 2012 and 2014 by revenue bins for the entire distribution (censored at R$5M) relative to the number of firms in the same revenue bins in 2010. The relative increase in the number of firms in 2012 and 2014 is disproportionately concentrated between the pre- and post-reform thresholds. Moreover, there is clear bunching below the post-reform threshold corresponding to a fourfold increase in the number of firms located within that that specific revenue bin by 2014.

Panels (c) and (d) restrict attention to a balanced panel of origin firms with positive revenue in all years from 2008 to 2016 and revenue levels below the pre-reform threshold in 2010. They display the median year-to-year growth rates for firms in three revenue bins (R$1.2M-R$1.6M, R$1.6M-R$2M, R$2M-R$2.4M) in 2010 for firms that were in the presumptive tax regime (panel c) and the regular tax regime (panel d) in 2010, separately (in the panels, we display the growth rate from $t-1$ to t in year t). The growth rate of SIMPLES firms that were just below the threshold in 2010 was relatively lower in 2010 and 2011, consistent with the bunching in 2010 in panel (a). However, these same firms experienced a relatively higher growth rate in 2012, which is consistent with the rapid disappearance of bunching below the pre-reform threshold after the reform. In contrast, firms in the regular tax regime in 2010 grew similarly across the three revenue bins. In sum, figure 3.6 shows that increasing the SIMPLES threshold had a substantial effect on the firm size distribution.

3.4.2 Market Competition

We now provide evidence for the adverse effect of the coexistence of presumptive and regular tax regimes, and the preferential tax treatment given to SMEs (the option to choose the presumptive tax regime) on ineligible firms competing in the same market as eligible firms. In particular, we exploit again the SIMPLES reform in 2012. The reform improved the tax treatment of SMEs in Brazil, by allowing SIMPLES firms to grow more without becoming ineligible for SIMPLES in the following year. It also reduced the average tax rate on turnover applying to SIMPLES firms (see appendix). We show that ineligible firms in sectors with more competition from SIMPLES firms before the reform grew relatively less after the reform compared to ineligible firms in sectors with less competition from SIMPLES firms before the reform.

Empirical strategy. We construct our sample of analysis as follows. We first restrict attention

[19] There is already some increase in the number of firms just above the SIMPLES threshold in 2011, as firms learned during the year that they could exceed the threshold without become ineligible for SIMPLES in the following year.

to the balanced panel of origin firms observed with positive revenue in all years between 2008 and 2014. We then select firms that were in the regular tax regime in 2010 and that had revenue levels well above the post-reform threshold at the time such that they were not directly affected by the reform (they were too far from the post-reform threshold to become eligible in later years). In practice, we select firms with revenue levels between R$4.6M (R$1M above the post-reform threshold) and R$20M (very large firms are unlikely to be competing with SMEs) in 2010. We then estimate regressions for $y_{i,s,t}$, the log(revenue) of firm i in sector s in year t, using specifications of the form:

$$y_{i,s,t} = \alpha_t + \beta \cdot SimplesMarketShare2010_s + \gamma \cdot SimplesMarketShare2010_s \cdot Post2012_t + \varepsilon_{i,s,t} \tag{3.3}$$

$$y_{i,s,t} = \alpha_t + \beta \cdot SimplesMarketShare2010_s + \gamma_t \cdot SimplesMarketShare2010_s + \varepsilon_{i,s,t} \tag{3.4}$$

where α_t are fixed effects for each calendar year t (2008,..., 2014). The variable $Post2012$ is an indicator for years after the reform. The variable $SimplesMarketShare2010_s$ is the market share of SIMPLES firms in 2010 for each sector, namely the total revenue of SIMPLES firms in the sector divided by the total revenue of all firms in the sector. It is meant to capture variation in the intensity of competition from SIMPLES firms across sectors: in sectors with high values of this variable, a larger share of the market benefited from the SIMPLES reform, likely intensifying competition faced by ineligible firms in these same sectors. The coefficient γ in equation (3.3) thus captures differential changes in the log(revenue) of ineligible firms after the 2012 SIMPLES reform by the intensity of competition from SIMPLES firms before the reform. The coefficients γ_t in equation (3.4) capture differential changes in each years compared to reference year 2010 in order to show that any differential effect actually occurred after the reform. Standard errors are clustered at the sector level, yielding 47 clusters.

Results. Table 3.3 presents the estimated $\widehat{\beta}$ and $\widehat{\gamma}$ from using the specification in equation (3.3). Column (1) first shows that the firms in our analysis sample were not directly affected by the reform. It uses an indicator for being in the regular tax regime in each year rather than log(revenue) as the outcome. This is to show that the market share of SIMPLES firms in 2010 is uncorrelated with tax regime choices both before and after the reform. The results confirm that we select firms that were too large prior to the reform in order to reduce revenue below the post-reform threshold after the reform.

Columns (2)-(5) then use log(revenue) as the outcome. Column (2) shows that the market share of SIMPLES firms in 2010 is negatively correlated with log(revenue) before the reform. This indicates that the average firm in sectors with a higher market share of SIMPLES firms in 2010 was smaller at baseline. Column (3) adds controls for firm size at baseline, namely ten indicators for deciles of the firm size distribution in 2010 as well as those indicators interacted with the post-2012 dummy. This is to separately identify a differential change post-reform related to the market share of SIMPLES firms in 2010 from a possible differential change post-reform related to firm size at

baseline. The correlation between log(revenue) and the market share of SIMPLES firms in 2010 is smaller and no longer significant in this case. In contrast, the change in this correlation after the reform remains almost unchanged with firm size controls and is significant at the 10% level. It implies that firms in sectors with a higher market share of SIMPLES firms in 2010 grew relatively less after the reform. This effect is an average over firms varying widely in their pre-reform size, and smaller firms are more likely to be competing with SMEs. Columns (4) and (5) thus presents results from separate regressions for firms below (smaller firms) and above (larger firms) the median firm size in the sample in 2010. The effect is concentrated among smaller firms, which reinforces our interpretation of the results as capturing a market competition effect.

Figure 3.7 further strengthens our results. First, panels (a)-(c) present binned scatterplots for the estimated $\widehat{\gamma}$ in the regressions presented in columns (3)-(5). It shows that the linearity assumption in our specifications is a good approximation of the relationship between firm size and the market share of SIMPLES firms post-reform, particularly for smaller firms. Second, panel (d) presents the estimated $\widehat{\gamma_t}$ from using the specification in equation (3.4) for all firms, for smaller firms, and for larger firms. It shows that the estimated coefficients moved around 0 prior to the reform and were very similar across the three groups. They became increasingly negative after the reform, particularly for smaller firms. Panel (d) suggests that some of the effect may have started in 2011 already, as firms learned during the year about the upcoming reform.[20] The coefficients presented in table 3.3 may thus even underestimate the adverse effect on ineligible firms.

Finally, column (6) in table 3.3 presents the estimated $\widehat{\beta}$ and $\widehat{\gamma}$ from using the specification in equation (3.3) at the sector-year level. The outcome is total revenue of the sector and the size fixed effects are four indicators for quartiles of the sector size distribution in 2010. Column (6) shows that sectors with a higher market share of SIMPLES firms in 2010 grew relatively more after the reform, despite the adverse effect on ineligible firms. This is consistent with the positive impact on the larger share of firms directly benefiting from the reform in these sectors.

3.4.3 Tax Regime and Input Choices

In this last subsection, we provide evidence supporting causal links between tax regime and input choices.

3.4.3.1 Effect of Input Choices on Tax Regime Choices

We first provide evidence supporting a causal effect of firms' heterogeneity in production functions on their choice of tax regime, i.e. of input choices on tax regime choices. We exploit again the reform that increased the SIMPLES threshold in 2012. We restrict attention to firms that were in the regular tax regime in both 2010 and 2011 (to avoid firms switching tax regime frequently) and that had revenue levels above the pre-reform threshold but below the post-reform threshold (the "reform region", between R$2.4M and R$3.6M)) in 2010. These firms became newly eligible for

[20] For instance, firms learned during the year that they could exceed the pre-reform threshold in 2011 without becoming ineligible for SIMPLES in the following year.

the presumptive tax regime in 2012. We then correlate their choice of tax regime in 2012 with their input choices in 2011 when all of these firms were in the regular tax regime. For this analysis, we use all firms in the full sample, thus including suppliers and clients of origin firms, given the small number of firms in the reform region in the origin sample (see figure 3.6a).

Panel (a) of figure 3.8 first shows that few of these firms switched to the presumptive tax regime in 2012. The figure uses a balanced panel of firms with positive revenue in all years from 2009 to 2014 and displays the share of firms in the regular tax regime in each year. By construction, it is equal to one in 2010 and 2011. The share dropped in 2012 to 94% and continued to drop in following years, although it remained above 92% in 2014.

Table 3.4 displays results from regressing an indicator for being in the regular tax regime in 2012 on the same firm characteristics as in table 3.2, but using the 2011 values of these variables when these firms were all in the regular tax regime (thus before they became eligible for SIMPLES). Column (1) only includes the log(revenue) in 2011. Columns (2) and (3) add log(input) in 2011, the share of intermediate input out of all input in 2011, and the share of intermediate input subject to VAT in 2011; column (3) adds sector fixed effects. Panels (b)-(d) of figure 3.8 also display binscatter plots of the estimated coefficients on log(revenue), the share of intermediate input out of all input, and the share of intermediate input subject to VAT using the specification in column (3). Larger firms were more likely to stay in the regular tax regime, which is consistent with the fact that they are closer to the new threshold (and so more likely to grow above it) and pay higher taxes in a turnover-based tax regime. The other results suggests that firms' production functions affect their choice of tax regime: firms with a higher share of intermediate input (and thus a lower share of labor input) and a higher share of their intermediate input subject to VAT in 2011 were more likely to stay in the regular tax regime in 2012.

3.4.3.2 Effect of Tax Regimes on Input Choices

Finally we provide evidence of a causal effect of tax regimes on input choices. Panel (a) of figure 3.8 suggests a first empirical design to study this question. Indeed, it present similar pattern for firms that were also in the regular tax regime in both 2010 and 2011 but that had revenue levels above the post-reform threshold in 2010 (between R$3.6M and R$4.8M). Some of these firms, which must have decreased their size in later years, also opt for the presumptive tax regime after the reform. However, more than 96% of them are still in the regular tax regime by 2014. The difference in the share of firms staying in the regular tax regime after the reform could form the "first stage" of a difference-in-differences analysis comparing input choices over time for the two groups of firms. We show in the notes below the graph in panel (a) that this first stage is statistically significant, but that it only reaches 4pp. We are unlikely to have the statistical power to detect change in input choices with such a small first stage in a relatively small sample to begin with. We are currently investigating ways to identify subgroups of the data with a larger first stage to carry out such an analysis.

To make some progress, we thus rely on an alternative empirical strategy. In particular, we

study the impact of a supplier's change of tax regime on a firm's decision to purchase intermediate inputs from that supplier through an event analysis. Importantly, we exploit heterogeneity in the tax regime of the client firms. When a supplier changes tax regime, it changes the marginal cost of a unit of intermediate input from that supplier for firms in the presumptive tax regime, as the VAT potentially charged by the supplier acts as an input tax for these firms. In contrast, a supplier's change of tax regime does not necessarily have any implications for the marginal cost of a unit of intermediate input from that supplier for firms in the regular tax regime, as they can deduct any VAT potentially charged by the supplier.

Empirical strategy. We select all pairs of firms in the transaction data in which the supplier switches tax regime at some point between 2013 and 2015, such that we can follow each pair for two years before and after the tax regime switch, over the period for which all our variables are available (the client firms are always origin firms). We then keep only pairs in which both the supplier and the client are observed with positive revenue for the four years around the event; we also impose that intermediate input should be positive in those years for the client firms. We index again event years by $k = -2, ...1$ with $k = 0$ as the year of the tax regime switch. Finally, we restrict attention to pairs in which the supplier changes tax regime only once over the 4-year period and the client is observed in a same tax regime in the two years before the event. Intuitively, the idea is to compare trade outcomes before and after the event for pairs in which the supplier switches into vs. out of the tax regime of the client. Our causal effect of interest implies that trade would increase in the first group relative to the second group for client firms in the presumptive tax regime. There are several issues to address before pursuing such an empirical strategy:

Relevance. A firm's supplier in a given year may not be a meaningful supplier in following years, irrespective of changes in tax regime. In that case, we would not expect any effect of the partners' respective tax regimes on their trade. In an attempt to focus on more relevant suppliers, we thus restrict attention to pairs in which the supplier was among the top-10 suppliers of the client in the year prior to the event ($k = -1$) and to pairs with positive trade in the two years before the event (in both $k = -1$ and $k = -2$).

Reverse causality. The supplier's choice of tax regime may in theory be affected by the client's decisions, in particular by its (expected) input decisions (e.g. whether to purchase inputs from that supplier) and tax regime choice (e.g. whether the client also changed tax regime around the same time). To limit such concerns of reverse causality, we restrict attention to clients that remain in the same tax regime for the four years around the event and we apply a "small economy argument", restricting attention to pairs in which the client is "small" for the supplier. Specifically, we eliminate pairs in which the client accounts for more than 5% of the supplier's sales in the year prior to the event ($k = -1$).

Omitted variable bias. Another concern is that, in theory, common shocks could affect both the supplier's tax regime choice and the client's input decisions. For instance, there is a general trend towards the presumptive tax regime in the years after 2012, both in number of firms and volume of trade. What drives this trend may cause both the supplier to switch towards the presumptive tax

regime and the client to buy more intermediate inputs from SIMPLES firms. To control for such trends in the data, we assign placebo switching years to comparable control pairs of trade partners in which neither the supplier nor the client switched tax regime during the four years around the placebo event.[21]

In sum, our analysis compares trade outcomes of four types of treatment pairs in which the supplier changes tax regime in year $k = 0$ to control pairs in which the supplier and the client are in the same tax regimes at baseline as the treatment pairs, but in which the supplier does not change tax regime. Table 3.5 defines our four types of treatment pairs (T1 to T4) and their associated control pairs (C1-C4) and displays samples sizes for each groups. One limitation is that samples sizes are relatively small for treatment pairs.

We present raw empirical patterns of relevant variables for all these groups (T1-T4 and C1-C4), separately. We then quantify changes in an outcome y using the following difference-in-differences specifications for each type of pair separately (e.g. T1 and C1 in one regression, and T2 and C2 in a different regression):

$$y_{i,k,t} = \alpha_i + \beta_k + \gamma_t + \delta_k \cdot Treatment_i + \varepsilon_{i,k,t} \tag{3.5}$$

$$y_{i,k,t} = \alpha_i + \beta_k + \gamma_t + \delta \cdot Treatment_i \cdot After_k + \varepsilon_{i,k,t} \tag{3.6}$$

where α_i, β_k, and γ_t are fixed effects for each firm i, each event year k, and each calendar year $t = 2011, ..., 2016$. The variable $Treatment_i$ is an indicator equal to one for treatment pairs. The coefficients δ_k in equation (3.5) thus capture differential changes for treatment pairs over time compared to a reference year. The coefficient δ in equation (3.6) summarizes the difference-in-differences by capturing the average change in the two years after vs. before the event. Standard errors are clustered by client firms.

Results. Figure 3.9 displays raw patterns in the data and figure 3.10 presents our estimated $\widehat{\delta_k}$ coefficients using the specification in equation (3.5). The left panels (right panels) in both figures consider client firms that are in the presumptive tax regime (in the regular tax regime). The first outcome is an indicator that the pair is still trading some positive intermediate input value in the year, capturing possible extensive margin responses. Given the sample selection criteria, this variable is equal to one in the two years before the event. There is thus some mean reversion in the two years after the event: the share of pairs still trading decreases over time in all treatment and control pairs. However, we find that client firms in the presumptive tax regime are more likely

[21] Control pairs are selected following the same criteria as for treatment pairs: the placebo events take place between 2013 and 2015; the client and the supplier have positive revenue in the four years around the placebo event; the client has positive intermediate input in these same four years; we restrict attention to pairs in which neither the supplier nor the client changes tax regime over the 4-year period; we restrict attention to pairs in which the supplier was among the top-10 suppliers of the client in the year prior to the placebo event and to pairs with positive trade in the two years before the placebo event; we eliminate pairs in which the client accounts for more than 5% of the supplier's sales in the year prior to the placebo event. Additionally, we restrict attention to pairs in which the supplier and the client are in similar sectors as suppliers and clients in treatment pairs and in which the supplier is eligible for the presumptive tax regime, such that it could potentially switch tax regime.

to continue trading with their supplier if the supplier switches to their own regime, such that the supplier does not charge them VAT anymore (see figure 3.9a). Compared to pairs in which the supplier remains in the regular tax regime, the effect amounts to a 10pp increase in the likelihood that the supplier continue to purchase intermediate inputs from that supplier (see figure 3.10a). In contrast, we don't find that client firms in the presumptive tax regime are less likely to continue to trade with their supplier if the supplier switches to the regular tax regime, such that the supplier starts charging them VAT (see figure 3.9a).[22] Moreoever, as expected, we find no differential effect for client firms in the regular tax regime.

The second outcome is the input value (log) bought from that supplier. In that case, we focus on the subset of pairs trading some positive amount in all years. This variable thus captures possible intensive margin responses. There is again clear mean reversion in the two years after the event. Moreover, we find that client firms in the presumptive tax regime are trading relatively more with their supplier if the supplier switches to their own regime, such that the supplier does not charge them VAT anymore (see figure 3.9c). Compared to pairs in which the supplier remains in the regular tax regime, the effect amounts to a 20% increase in the amount of intermediate input purchased from that supplier (see figure 3.10c). In contrast, we don't find that client firms in the presumptive tax regime trade less with their supplier if the supplier switches to the regular tax regime, such that the supplier starts charging them VAT (see figure 3.9c). Furthermore, we find no differential effect for client firms in the regular tax regime as expected.

In sum, the above results provide some evidence that the tax regime of (potential) trade partners has a causal effect on client's input choices and on market segmentation between firms in regular and presumptive tax regimes. Yet, the effects remain moderate in magnitude, such that overall trade of a firm is not affected much by the suppliers' change of tax regime. As a result, we find that a client's share of intermediate input subject to VAT increases discontinuously (decreases discontinuously) when a supplier switches to the regular tax regime (the presumptive tax regime). This is shown in panels (e) and (f) of figures 3.9 and 3.10.[23]

3.5 Conclusion

Dual tax systems allow for the co-existence of firms subject to regular modern taxes and firms that are taxed on a presumptive basis. This has implications for firm growth, input choices, and market competition. This chapter uses novel administrative data on inter-firm trade linked to labor inputs from São Paulo, Brazil, to document a number of empirical patterns on the relationship between tax systems and firms' sourcing choices.

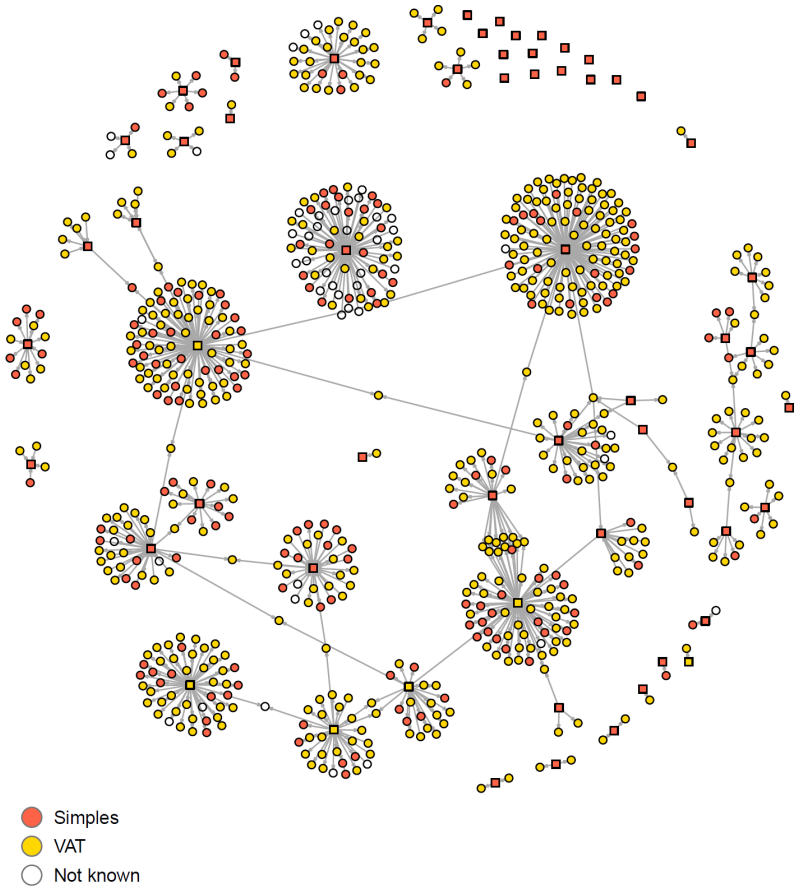
In particular, we show that a tax system in which the VAT and payroll taxes are replaced by a presumptive tax on revenue affects firms in three important ways. First, it distorts the firm size distribution and their incentive to grow. Second, it affects market competition, hurting

[22] We are still trying to understand the reason behind this asymmetry.

[23] We find no effect on other outcomes. We are currently increasing sample sizes by using the 2017 data and adding events taking place in 2016.

firms that are ineligible for the presumptive tax. Third, it affects the composition of their inputs since going from a presumptive system to the regular system means that labor becomes taxed and intermediate inputs become deductible. This has implications for the supply chain, as we find evidence of (partial) segmentation between regular firms and firms in the presumptive system. We show that heterogeneity in firm production choices drive part of these correlations, but that tax regimes also causally affect input choices and market segmentation.

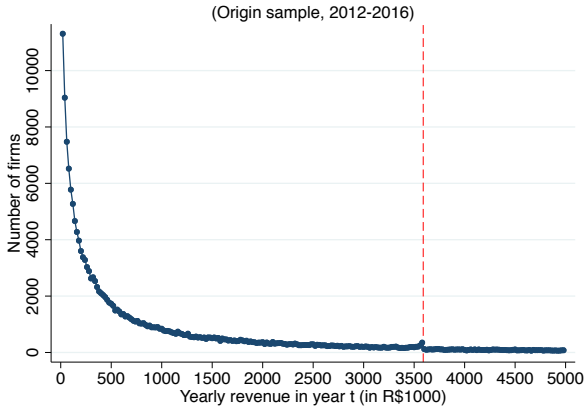
Figure 3.1: Illustration of the Inter-Firm Trade Data



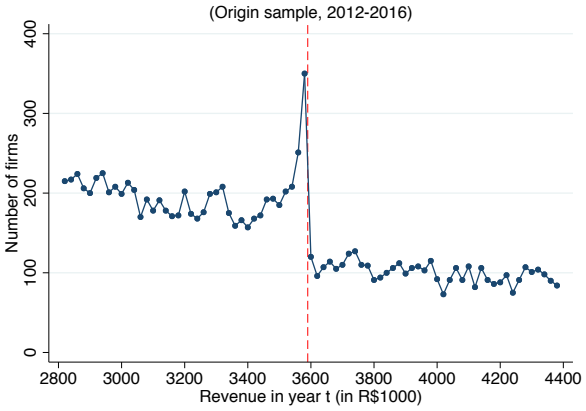
Note: The figure shows an illustration of the detailed transaction data available for this project using a small random sample of the data in one year. It does not show all connections across the firms that are included in the picture, but it helps illustrate the data structure. All the squares are firms in the origin sample. Circles are firms that are either supplier or clients of origin firms. Yellow circles or squares are VAT-registered (regular) firms, while orange circles or squares are SIMPLES firms. The white circles are firms that are not registered as taxpayers in the state of São Paulo (e.g. firms that provide services or that are located in other states).

Figure 3.2: Bunching and Choice of Tax Regime

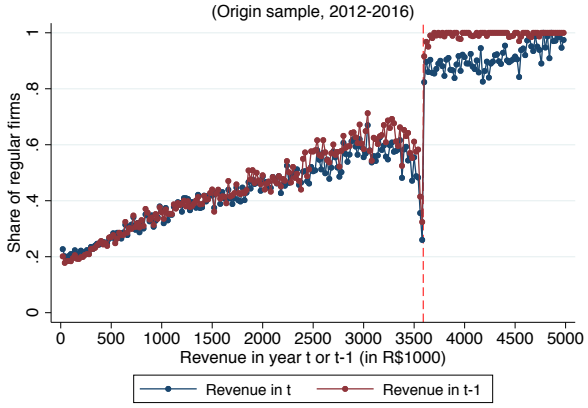
(a) Overall revenue distribution (origin sample)



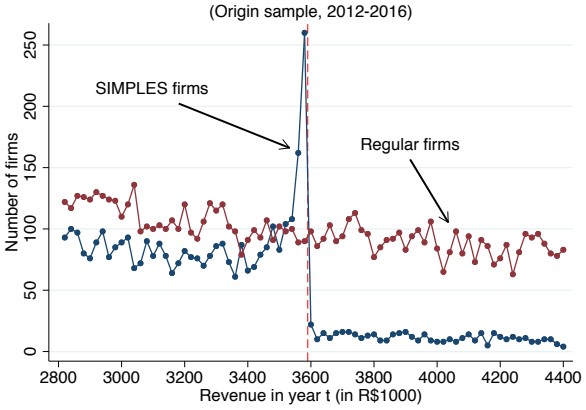
(b) Revenue distribution around the VAT threshold (origin sample)



(c) Share VAT registered by revenue level (origin sample)



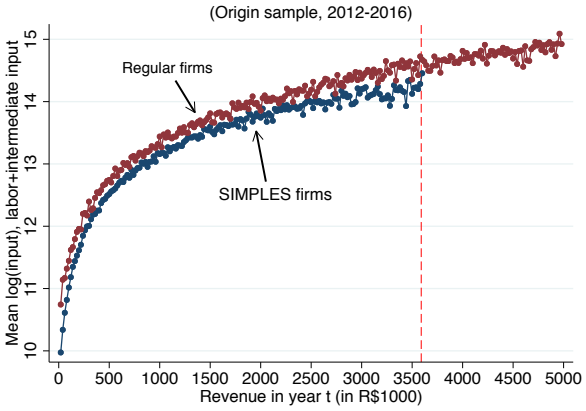
(d) Revenue distribution around the VAT threshold by tax regime (origin sample)



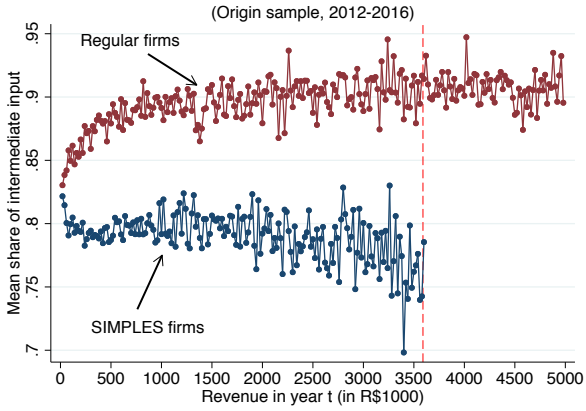
Note: The vertical line indicates the location of the SIMPLES threshold: R$ 3.6 million). The graphs use revenue bins of R$20,000. The sample is composed of all firm-year observations with positive revenue in the origin sample between 2012 and 2016.

Figure 3.3: Input Choices and Tax Regimes

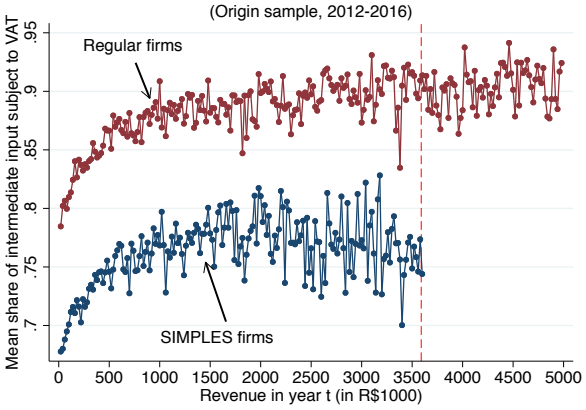
(a) Mean log(input), including intermediate and labor input



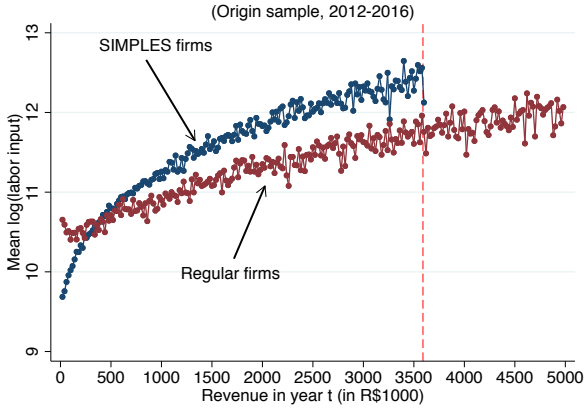
(b) Mean share of intermediate input (out of all inputs)



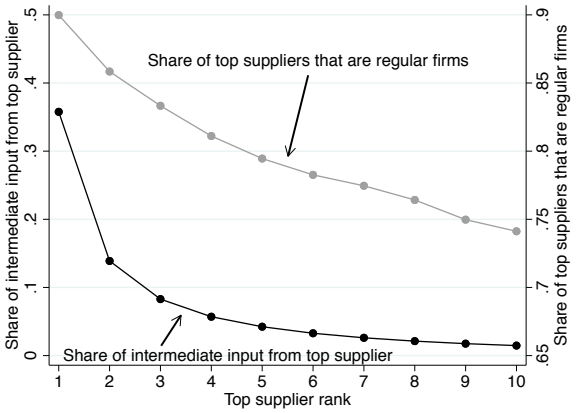
(c) Mean share of intermediate input subject to VAT (out of all intermediate inputs)



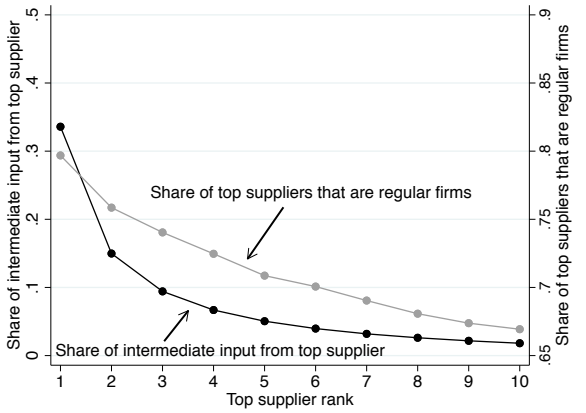
(d) Mean log(labor input)



(e) Top 10 suppliers of regular firms below the threshold



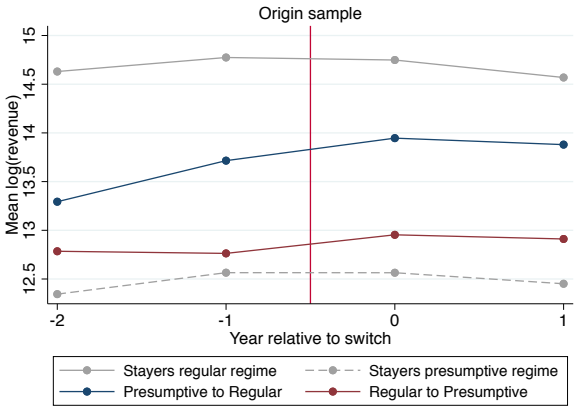
(f) Top 10 suppliers of SIMPLES firms below the threshold



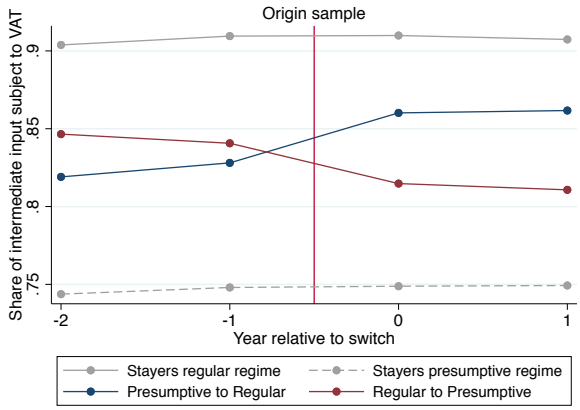
Note: The sample in panel is composed of all firm-year observations with positive revenue in the origin sample between 2012 and 2016, conditional of having any input (panels a and b), any intermediate input (panel c), and any labor input (panel d). The value of intermediate input is the pre-tax value of input transactions during the year. The value of labor input is the total payroll over the year (gross wages paid to employees). Panels (e) and (f) use samples restricted to firm-year observations with at least 10 suppliers.

Figure 3.4: Event Analysis around a Firms' Tax Regime Switches (Raw Data)

(a) Log(revenue)



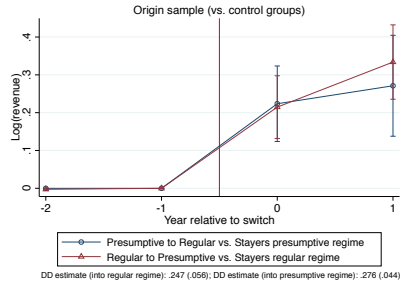
(b) Share of intermediate input subject to VAT



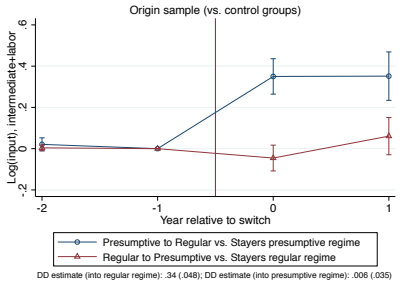
Note: The vertical line indicates the timing of the tax regime switch. The samples consists of origin firms with positive revenue and intermediate input in the two years before and after a (placebo) tax regime switch taking place between 2013 and 2015.

Figure 3.5: Event Analysis around a Firms' Tax Regime Switches (DD Estimates)

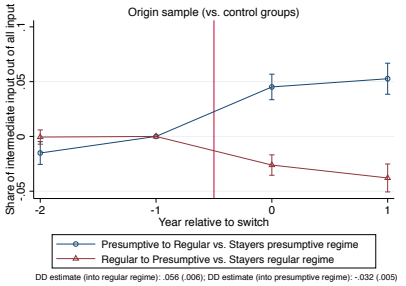
(a) Log(revenue)



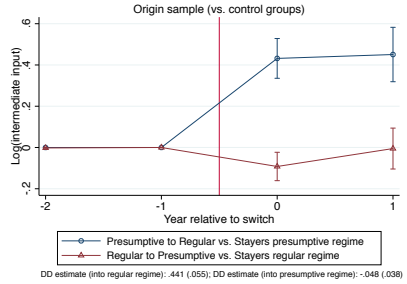
(b) Log(input)



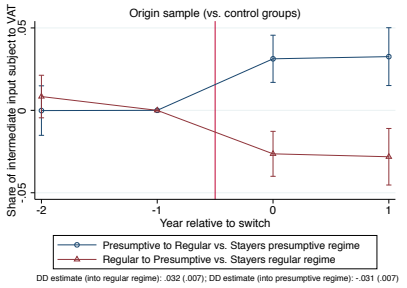
(c) Share of intermediate input out of all inputs



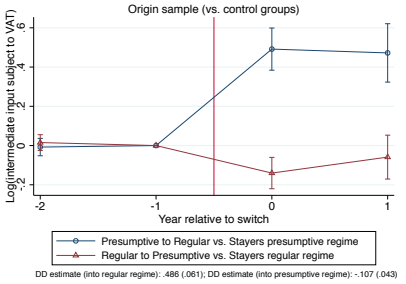
(d) Log(intermediate input)



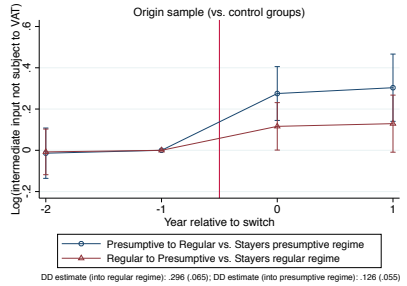
(e) Share of intermediate input subject to VAT



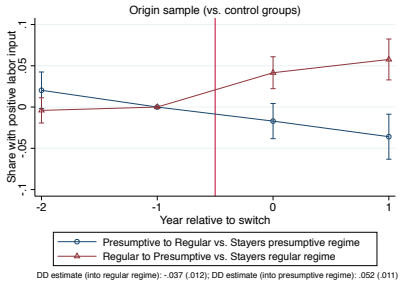
(f) Log(intermediate input subject to VAT)



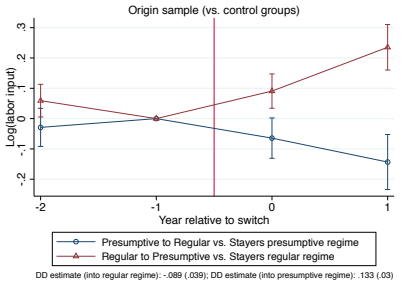
(g) Log(intermediate input not subject to VAT)



(h) Any labor input



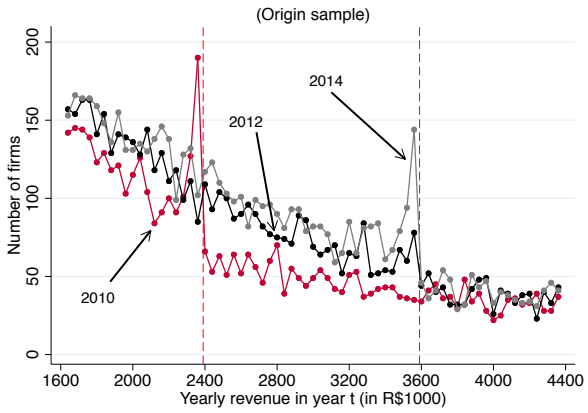
(i) Log(labor input)



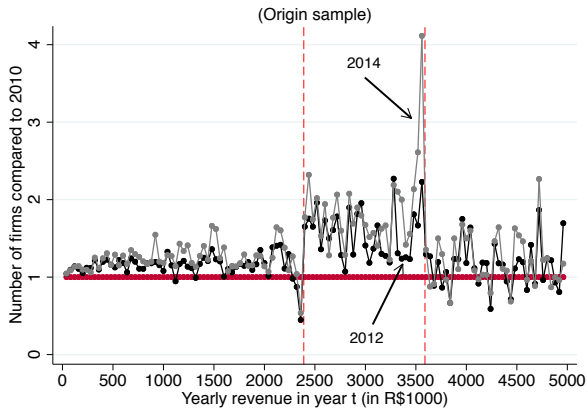
Note: The vertical line indicates the timing of the tax regime switch. The samples consists of origin firms with positive revenue and intermediate input in the two years before and after a (placebo) tax regime switch taking place between 2013 and 2015. The patterns for log(intermediate input subject to VAT), log(intermediate input not subject to VAT), and log(labor input) are similar if we restrict attention to firms with non-missing values for these variables in all years.

Figure 3.6: The Effect of Increasing the SIMPLES Threshold on the Firm Size Distribution

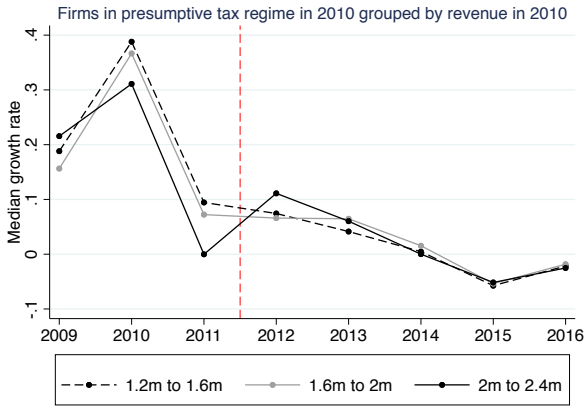
(a) Number of firms by revenue level around the SIMPLES thresholds



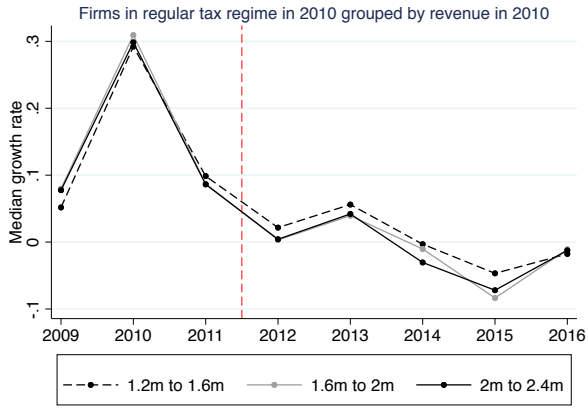
(b) Number of firms by revenue level compared to 2010



(c) Median growth rate around SIMPLES extension (firms grouped by revenue level in 2010) – SIMPLES firms in 2010



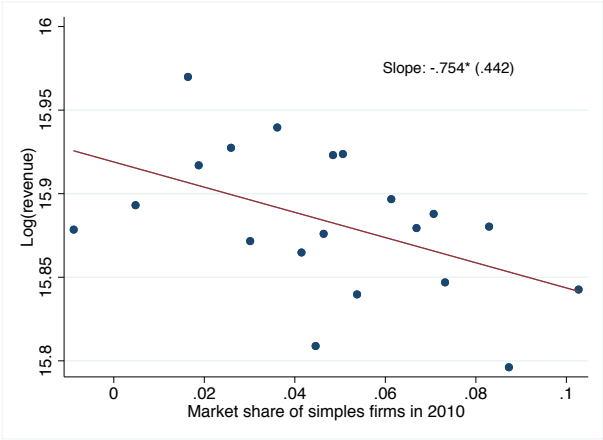
(d) Median growth rate around SIMPLES extension (firms grouped by revenue level in 2010) – SIMPLES firms in 2010



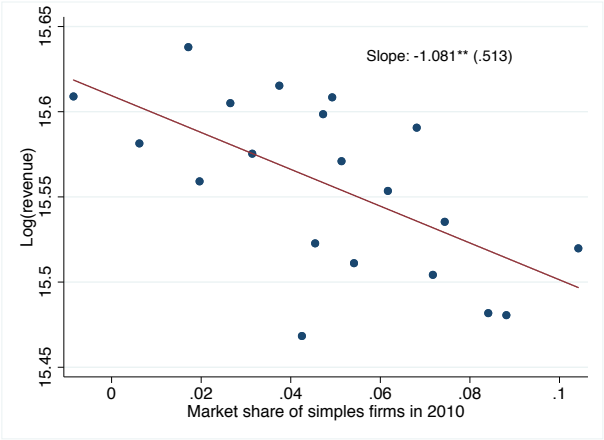
Note: The samples consist of origin firms only. The vertical lines in panels (a) and (b) indicate the location of the SIMPLES threshold before (R$ 2.4M) and after (R$ 3.6M) the 2012 reform. The graphs use revenue bins of R$20,000. The sample is composed of all firm-year observations with positive revenue in the origin sample in 2010, 2012, and 2014. The samples in panels (c) and (d) are restricted to a balanced panel of firms with positive revenue in all years from 2008 to 2016. Panel (c) considers firms in the SIMPLES regime in 2010 with revenue levels between R$1.2M and R$2.4M in 2010. Panel (d) considers firms in the regular tax regime in 2010 with revenue levels between R$1.2M and R$2.4M in 2010. The vertical line indicates the timing of the reform location of the SIMPLES threshold: R$ 3.6 million). The graphs use revenue bins of R$20,000. The sample is composed of all firm-year observations with positive revenue in the origin sample between 2012 and 2016.

Figure 3.7: The Effect of the 2012 SIMPLES Reform on Ineligible Firms

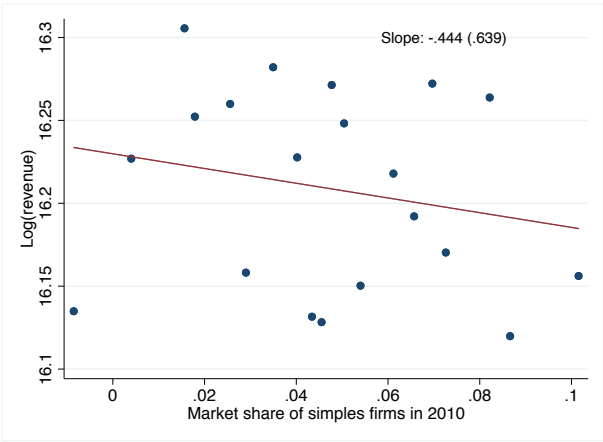
(a) Differential impact on ineligible firms, all firms



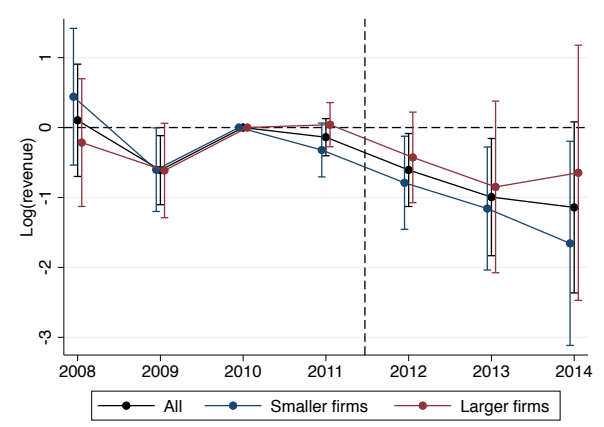
(b) Differential impact on ineligible firms, smaller firms



(c) Differential impact on ineligible firms, larger firms



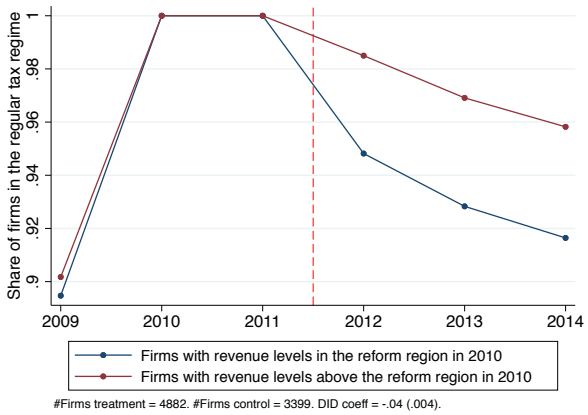
(d) Differential impact on ineligible firms over time



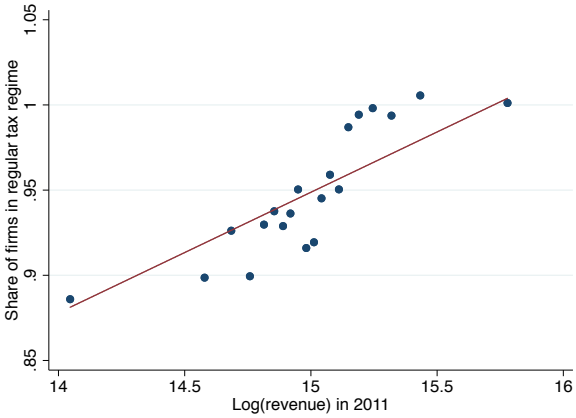
Note: Panels (a)-(c) present binscatter plots for the estimated $\widehat{\gamma}$ in the regressions presented in columns (3)-(5) of table 3.3, respectively. Panel (d) presents the estimated $\widehat{\gamma}_t$ from using the specification in equation (3.4) for all firms, for smaller firms, and for larger firms, separately.

Figure 3.8: Tax Regime Choice among firms Newly Eligible for SIMPLES

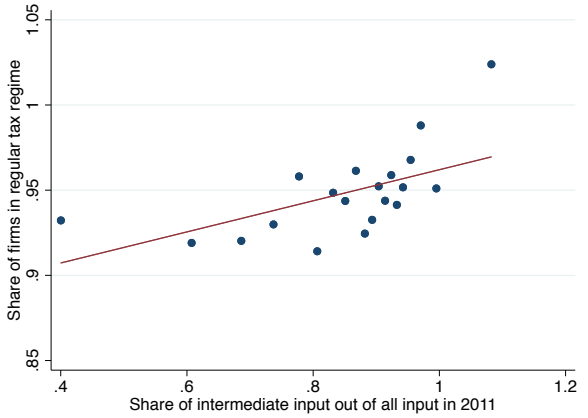
(a) Share of firms in the regular tax regime over time



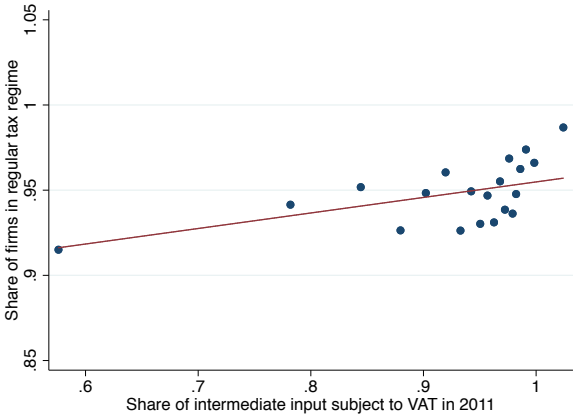
(b) Binscatter of tax regime choice in 2012 by 2011 log(revenue)



(c) Binscatter of tax regime choice in 2012 by 2011 share of intermediate input out of all input



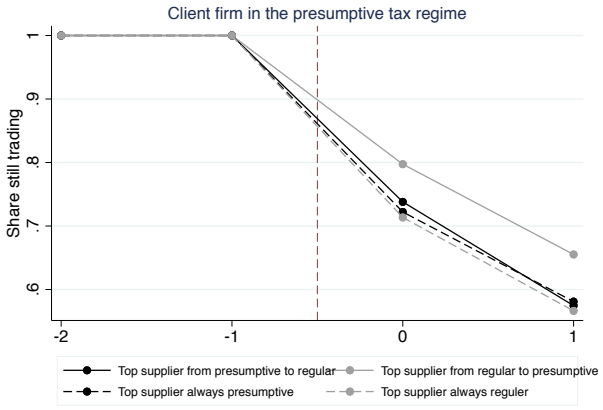
(d) Binscatter of tax regime choice in 2012 by 2011 share of intermediate input subject to VAT



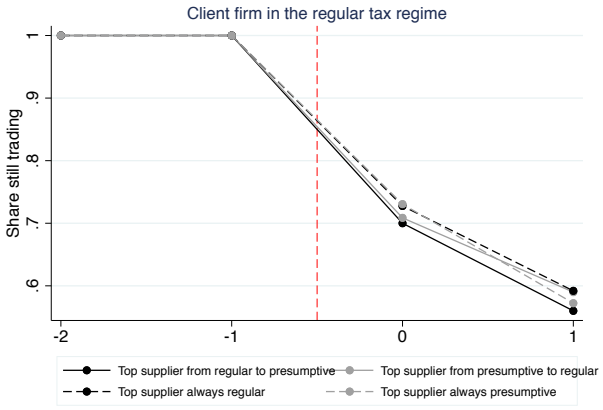
Note: The samples consist of a balanced panel of firms in the full sample with positive revenue in all years from 2009 to 2014, restricting attention to firms that were in the regular tax regime in both 2010 and 2011 (to avoid firms switching tax regime frequently) and that had revenue levels above the pre-reform threshold but below the post-reform threshold (the "reform region", between R$2.4M and R$3.6M)) in 2010. These firms became newly eligible for the presumptive tax regime in 2012. For comparison purposes, panel (a) also uses a sample of firms selected similarly but that had revenue levels above the post-reform threshold in 2010 (between R$3.6M and R$4.8M). Panel (a) display the share of firms in the regular tax regime in each year. Panels (b)-(d) displays binscatter plots for the correlations in table 3.4.

Figure 3.9: Event Analysis around a Supplier's Change of Tax Regime (Raw Data)

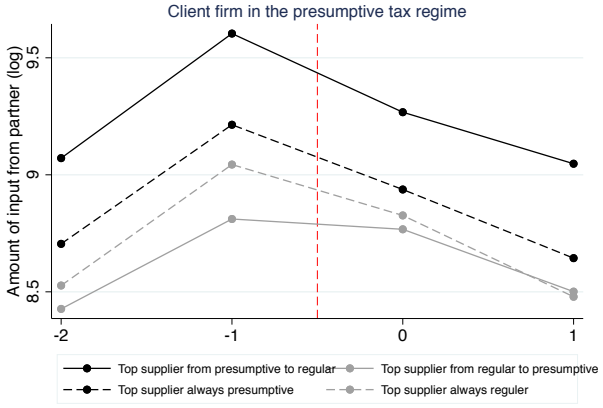
(a) Pair still trading (client presumptive regime)



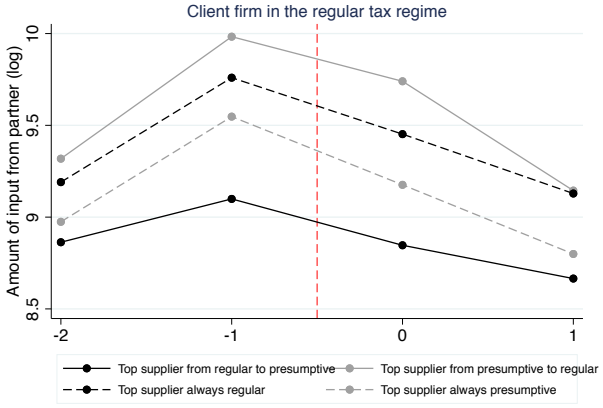
(b) Pair still trading (client regular regime)



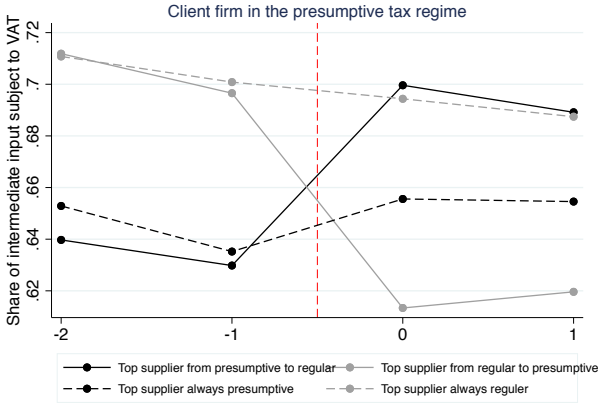
(c) Log(trade volume) if trading in all years (client presumptive regime)



(d) Log(trade volume) if trading in all years (client regular regime)



(e) Share of intermediate input subject to VAT (client presumptive regime)



(f) Share of intermediate input subject to VAT (client regular regime)

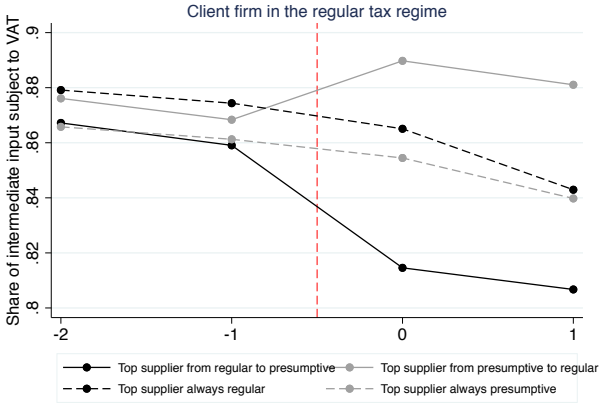
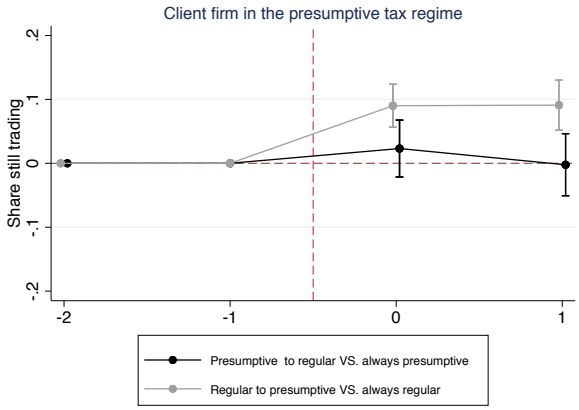
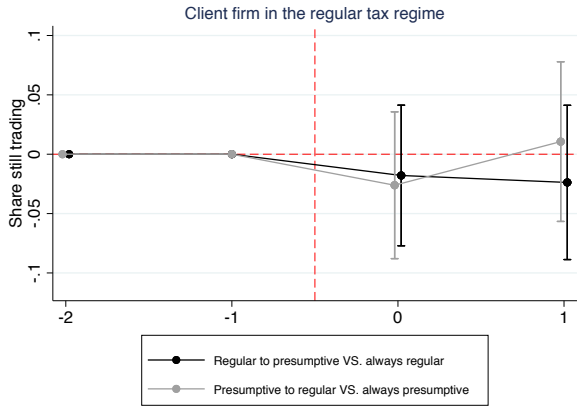


Figure 3.10: Event Analysis around a Supplier's Change of Tax Regime (DD estimates)

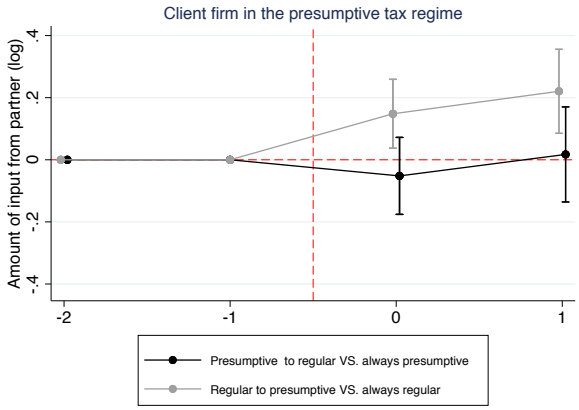
(a) Pair still trading (client presumptive regime)



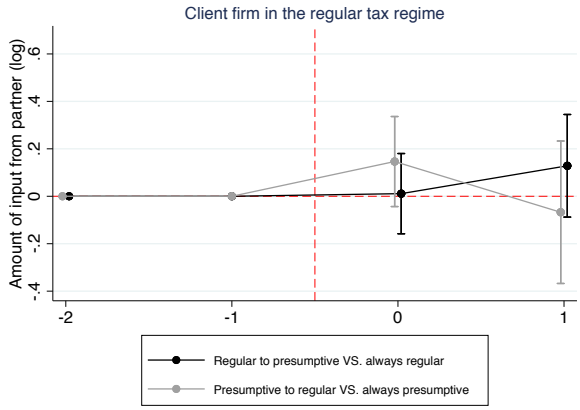
(b) Pair still trading (client regular regime)



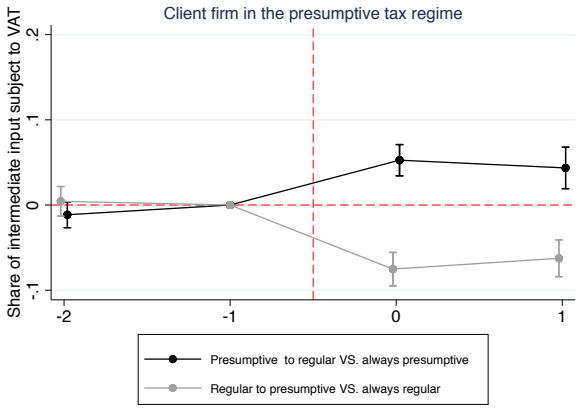
(c) Log(trade volume) if trading in all years (client presumptive regime)



(d) Log(trade volume) if trading in all years (client regular regime)



(e) Share of intermediate input subject to VAT (client presumptive regime)



(f) Share of intermediate input subject to VAT (client regular regime)

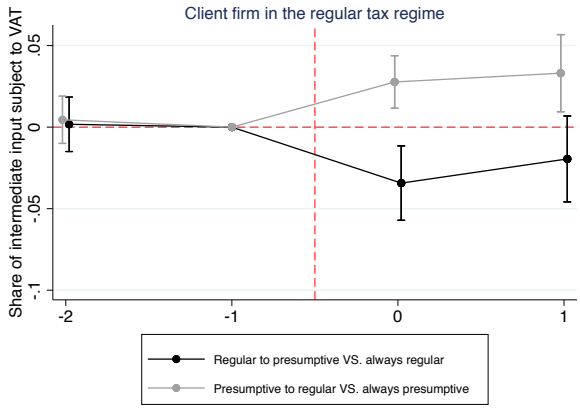


Table 3.1: Descriptive Statistics

Table 1: Descriptive statistics and comparison of firms in the presumptive and regular tax regimes (origin sample, 2012-2016)

	All firms	Firms above SIMPLES threshold	Below SIMPLES threshold	
			Regular firms	SIMPLES firms
	(1)	(2)	(3)	(4)
Group-level share of the sample	1	0.169	0.240	0.591
Share of SIMPLES firms	0.595	0.020	0	1
Mean log(revenue)	13.045	16.504	12.741	12.182
Group-level share of total revenue across all firms in the year	1	0.899	0.041	0.060
Share of firms with any (intermediate or labor) input	0.972	0.999	0.981	0.960
Mean log(input)	12.721	16.009	12.662	11.770
Share of firms with any intermediate input if any input	0.989	1.000	0.995	0.983
Share of firms with any labor input if any input	0.707	0.925	0.620	0.678
Firm-level share of intermediate input out of all inputs *[1- share of labor input]*	0.839	0.915	0.873	0.802
Mean log(intermediate input)	12.473	15.916	12.459	11.439
Group-level share of total intermediate input across all firms in the year	1	0.890	0.061	0.049
Number of suppliers (p10-p50-p90)	2-15-88	2-71-255	2-13-57	2-11-50
Firm-level share of intermediate input subject to VAT out all intermediate input	0.791	0.923	0.852	0.725
Mean log(labor Input)	11.318	13.012	10.993	10.755
Mean number of employees	12.036	51.670	4.298	3.875
Mean number of establishments	1.093	1.508	1.033	1.000
Number of firm-year observations	230490	38866	55316	136308
Number of firms	66456	11502	21778	42644

Note: The sample is composed of all firm-year observations with positive revenue in the origin sample between 2012 and 2016. The value of intermediate input is the pre-tax value of input transactions during the year. The value of labor input is the total payroll over the year (gross wages paid to employees).

Table 3.2: Correlates of Tax Regime Choice

Table 2: Correlates of tax regime choice (firms below the SIMPLES threshold, 2012-2016)

	Regular tax regime (1)	Regular tax regime (2)	Regular tax regime (3)	Regular tax regime (4)	Regular tax regime (5)
Log(revenue)	0.0331***	-0.0124***	-0.0118***	-0.00117	-0.0232***
	(0.000990)	(0.00129)	(0.00126)	(0.00296)	(0.00214)
Log(input)		0.0533***	0.0547***	0.0389***	0.0132***
		(0.00120)	(0.00119)	(0.00275)	(0.00180)
Share of intermediate input out of all inputs		0.224***	0.214***	0.248***	0.0582***
		(0.00616)	(0.00630)	(0.0102)	(0.00813)
Share of intermediate inputs subject to VAT		0.211***	0.182***	0.183***	0.0233***
		(0.00611)	(0.00590)	(0.00774)	(0.00489)
Year fixed effects	X	X	X	X	X
Sector fixed effects			X	X	
Balanced panel				X	X
Firm fixed effects					X
R-squared	0.019	0.085	0.124	0.115	0.893
Number of observations	182,619	182,619	182,619	98,040	98,040
Number of clusters	58018	58018	58018	19608	19608

Note: The sample is composed of all firm-year observations in the origin sample with positive revenue, positive intermediate input, and revenue levels below the SIMPLES threshold.

Table 3.3: Impact of the 2012 SIMPLES Reform on Ineligible Firms

Table 3: Impact of 2012 SIMPLES reform on ineligible firms

		All firms		Smaller firms	Larger firms	Sector-level analysis
	Regular tax regime	Log(revenue)	Log(revenue)	Log(revenue)	Log(revenue)	Log(revenue)
	(1)	(2)	(3)	(4)	(5)	(6)
Market share of SIMPLES firms in 2010	-0.0284	-0.418*	-0.172	-0.163	-0.170	-0.883
	(0.0332)	(0.243)	(0.126)	(0.157)	(0.157)	(0.859)
Market share of SIMPLES firms in 2010 x Post2012	-0.0160	-0.713*	-0.754*	-1.081**	-0.444	0.567**
	(0.0419)	(0.423)	(0.442)	(0.513)	(0.639)	(0.259)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Size fixed effects interacted with Post2012			Yes	Yes	Yes	Yes
R-squared	0.022	0.069	0.277	0.149	0.139	0.876
Number of observations	16,604	16,604	16,604	8,302	8,302	336
Number of sectors (clusters)	47	47	47	46	45	48

Note: The sample of analysis in columns (1)-(5) is composed of a balanced panel of origin firms with positive revenue in all years from 2008 to 2014 that were in the regular tax regime in 2010 and had revenue levels between R$4.6M and R$20M in 2010. The table displays coefficient estimates from using the specification in equation (3.3). The outcome in column (1) is an indicator for being in the regular tax regime. The outcome in columns (2)-(5) is the log(revenue). Columns (4) and (5) restrict attention to smaller and larger firms in the sample, respectively, namely those below and above the median revenue level in 2010. Column (6) displays coefficient estimates from using the same specification but with sector-year observations; the outcome is the total revenue of the sectors in each year in this case. Standard errors are clustered at the sector level in all columns.

Table 3.4: Tax Regime Choice in 2012 and 2011 Firm Characteristics

Table 4: Tax regime choice in 2012 and 2011 firm characteristics

	Regular tax regime in 2012 (1)	Regular tax regime in 2012 (2)	Regular tax regime in 2012 (3)
Log(revenue) in 2011	0.0723***	0.0708***	0.0742***
	(0.00914)	(0.0103)	(0.00969)
Log(input) in 2011		-0.00521	0.00131
		(0.00324)	(0.00352)
Share of intermediate input out of all inputs in 2011		0.0913***	0.123***
		(0.0220)	(0.0253)
Share of intermediate inputs subject to VAT in 2011		0.0881**	0.0877**
		(0.0347)	(0.0374)
Sector fixed effects			X
R-squared	0.018	0.025	0.115
Number of observations	4,879	4,878	4,878
Number of clusters	4,879	4,878	4,878

Note: The sample is composed of a balanced panel of firms in the full sample with positive revenue in all years from 2009 to 2014, restricting attention to firms that were in the regular tax regime in both 2010 and 2011 (to avoid firms switching tax regime frequently) and that had revenue levels above the pre-reform threshold but below the post-reform threshold (the "reform region", between R$2.4M and R$3.6M)) in 2010. These firms became newly eligible for the presumptive tax regime in 2012. The columns display the results from regressing an indicator for being in the regular tax regime in 2012 on various firm characteristics in 2011 (prior to the reform, when all firms where in the regular tax regime).

Table 3.5: Treatment and Control Group for To Supplier Switching Event Analysis

Table 5: Treatment and control groups for event analysis around top suppliers' tax regime switch

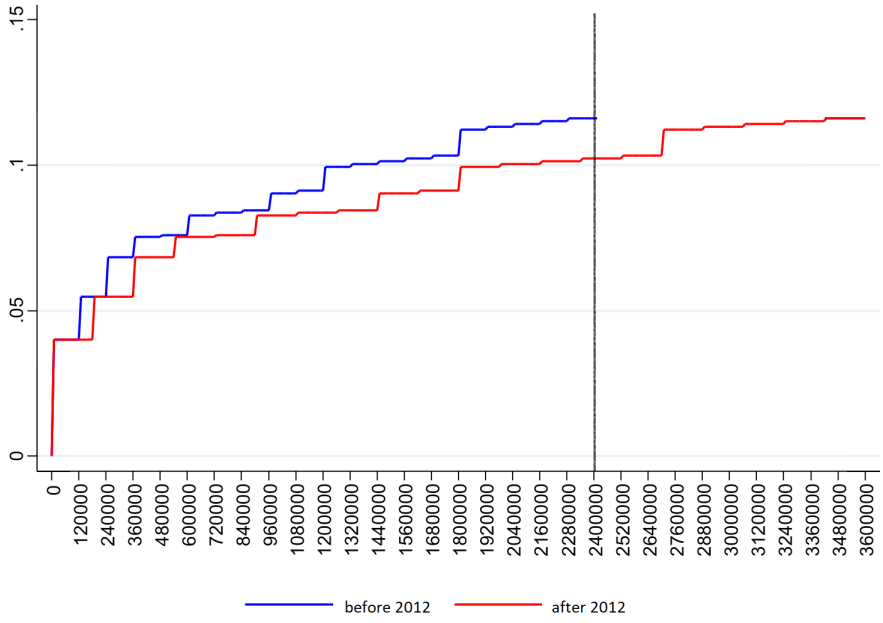
	Client tax regime	Supplier tax regime before the (placebo) event	Supplier tax regime after the (placebo) event	Number of pairs
	(1)	(2)	(3)	(3)
Treatment group 1 (T1)	Presumptive	Presumptive	Regular	710
Control group 1 (C1)	Presumptive	Presumptive	Presumptive	17,460
Treatment group 2 (T2)	Presumptive	Regular	Presumptive	612
Control group 2 (C2)	Presumptive	Regular	Regular	6,237
Treatment group 3 (T3)	Regular	Regular	Presumptive	250
Control group 3 (C3)	Regular	Regular	Regular	3,996
Treatment group 4 (T4)	Regular	Presumptive	Regular	278
Control group 4 (C4)	Regular	Presumptive	Presumptive	6,584

3.A Appendix

Average tax rate in SIMPLES regime

Figure 3.A1 displays the average tax rate on yearly gross revenue applying to firms opting for SIMPLES, the presumptive tax system in Brazil. Yearly gross revenue is both the tax base and the running variable in Figure 3.A1. The eligibility threshold was increased by 50% in 2012, from R$2.4 million to R$ 3.6 million. The tax rate schedule also changed at the time; the blue and red lines correspond to the pre-reform and the post-reform schedules, respectively.

Figure 3.A1: Average tax rate in SIMPLES Regime



Bibliography

Allen, E. J., Dechow, P. M., Pope, D. G. and Wu, G. (2016), 'Reference-Dependent Preferences: Evidence from Marathon Runners', *Management Science* **63**(6), 1657–1672.

Almunia, M., Gerard, F., Hjort, J., Knebelmann, J., Nakyambadde, D., Raisaro, C. and Tian, L. (2017), 'An Analysis of Discrepancies in Tax Declarations Submitted Under Value-Added Tax in Uganda'. IGC Report.

Angrist, J. D. (1998), 'Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants', *Econometrica* **66**(2), 249–288.

Asatryan, Z. and Peichl, A. (2017), 'Responses of Firms to Tax, Administrative and Accounting Rules: Evidence from Armenia'. CESifo Working Paper No. 6754.

Barberis, N. C. (2013), 'Thirty Years of Prospect Theory in Economics: A Review and Assessment', *Journal of Economic Perspectives* **27**(1), 173–196.

Barberis, N. and Xiong, W. (2009), 'What Drives the Disposition Effect? An Analysis of a Long-Standing Preference-Based Explanation', *Journal of Finance* **64**(2), 751–784.

Behaghel, L. and Blau, D. M. (2012), 'Framing Social Security Reform: Behavioral Responses to Changes in the Full Retirement Age', *American Economic Journal: Economic Policy* **4**(4), 41–67.

Best, M. C., Brockmeyer, A., Kleven, H. J., 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.

Best, M. and Kleven, H. (2018), 'Housing Market Responses to Transaction Taxes: Evidence from Notches and Stimulus in the UK', *Review of Economic Studies* **85**(1), 157–193.

Bird, R., Wallace, S. et al. (2003), 'Is it Really so Hard to Tax the Hard-to-Tax? The Context and Role of Presumptive Taxes'. International Tax Program Papers No. 0307.

Boonzaaier, W., Harju, J., Matikka, T. and Pirttilä, J. (2017), 'How Do Small Firms Respond to Tax Schedule Discontinuities? Evidence from South African Tax Registers'. VATT Working Paper No. 85.

Börsch-Supan, A. and Schnabel, R. (1999), Social security and retirement in germany, *in* J. Gruber and D. A. Wise, eds, 'Social Security and Retirement around the World', University of Chicago Press, pp. 135–180.

Börsch-Supan, A. and Wilke, C. B. (2004), 'The German Public Pension System: How It Was, How It Will Be'. NBER Working Paper No. 10525.

Brockmeyer, A. and Hernandez, M. (2018), 'Taxation, Information and Withholding: Evidence from Costa Rica'. World Bank Working Paper.

Brown, J. R., Kapteyn, A. and Mitchell, O. S. (2013), 'Framing and Claiming: How Information-Framing Affects Expected Social Security Framing Behavior', *The Journal of Risk and Insurance* **83**(1), 139–162.

Brown, K. M. (2013), 'The Link between Pensions and Retirement Timing: Lessons from California Teachers', *Journal of Public Economics* **98**(1–2), 1–14.

Burtless, G. (1986), 'Social Security, Unanticipated Benefit Increases, and the Timing of Retirement', *Review of Economic Studies* **53**(5), 781–805.

Camerer, C., Babcock, L., Loewenstein, G. and Thaler, R. (1997), 'Labor Supply of New York City Cab Drivers: One Day at a Time', *Quarterly Journal of Economics* **112**(2), 407–441.

Chetty, R. (2012), 'Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply', *Econometrica* **80**(3), 969–1018.

Chetty, R. (2015), 'Behavioral Economics and Public Policy: A Pragmatic Perspective', *American Economic Review: Papers & Proceedings* **105**(5), 1–33.

Chetty, R., Friedman, J. N., Olsen, T. and Pistaferri, L. (2011), 'Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records', *The Quarterly Journal of Economics* **126**(2), 749–804.

Cribb, J., Emmerson, C. and Tetlow, G. (2016), 'Signals Matter? Large Retirement Responses to Limited Financial Incentives', *Labour Economics* **42**, 203–212.

De Paula, A. and Scheinkman, J. A. (2010), 'Value-Added Taxes, Chain Effects, and Informality', *American Economic Journal: Macroeconomics* **2**(4), 195–221.

DellaVigna, S., Lindner, A., Reizer, B. and Schmieder, J. F. (2018), 'Reference-Dependent Job Search: Evidence from Hungary', *The Quarterly Journal of Economics* **132**(4), 1669–2018.

Dharmapala, D., Slemrod, J. and Wilson, J. D. (2011), 'Tax Policy and the Missing Middle: Optimal Tax Remittance with Firm-Level Administrative Costs', *Journal of Public Economics* **95**(9-10), 1036–1047.

Gadenne, L., Rathelot, R. and Nandi, T. (2018), 'Taxation and Supplier Networks: Evidence from India'. Working Paper.

Garicano, L., Lelarge, C. and Van Reenen, J. (2016), 'Firm Size Distortions and the Productivity Distribution: Evidence from France', *American Economic Review* **106**(11), 3439–79.

Gelber, A. M., Jones, D. and Sacks, D. W. (2017), 'Estimating Earnings Adjustment Frictions: Method & Evidence from the Earnings Test'. Working Paper.

Goda, G. S., Ramnath, S., Shoven, J. B. and Slavov, S. N. (2018), 'The Financial Feasibility of Delaying Social Security: Evidence from Administrative Tax Data', *forthcoming, Journal of Pension Economics & Finance* .

Harju, J., Matikka, T. and Rauhanen, T. (2015), 'The Effect of VAT Threshold on the Behavior of Small Businesses: Evidence and Implications'.

Heien, T., Kortmann, K. and Schatz, C. (2005), 'Altersvorsorge in Deutschland 2005'. Report by TNS Infratest Sozialforschung.

Hsieh, C.-T. and Klenow, P. J. (2009), 'Misallocation and Manufacturing TFP in China and India', *The Quarterly Journal of Economics* **124**(4), 1403–1448.

Hsieh, C.-T. and Olken, B. A. (2014), 'The Missing" Missing Middle"', *Journal of Economic Perspectives* **28**(3), 89–108.

Kahneman, D. and Tversky, A. (1979), 'Prospect Theory: An Analysis of Decision Making under Risk', *Econometrica* **47**(2), 263–291.

Kanbur, R. and Keen, M. (2014), 'Thresholds, Informality, and Partitions of Compliance', *International Tax and Public Finance* **21**(4), 536–559.

Keen, M. and Mintz, J. (2004), 'The Optimal Threshold for a Value-Added Tax', *Journal of Public Economics* **88**(3-4), 559–576.

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

Kleven, H. J. 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.

Lalive, R. and Staubli, S. (2015), 'How Does Raising Women's Full Retirement Age Affect Labor Supply, Income, and Mortality?'. NBER Retirement Research Center Paper No. NB 14-09.

Lalive, R., Staubli, S. and Magesan, A. (2017), 'Raising the Full Retirement Age: Defaults vs. Incentives'. Working Paper.

Liu, L., Lockwood, B. and Almunia, M. (2018), 'VAT Notches, Voluntary Registration, and Bunching: Theory and UK Evidence'.

Lumsdaine, R. L., Stock, J. H. and Wise, D. A. (1996), Why are retirement rates so high at age 65?, *in* D. A. Wise, ed., 'Advances in the Economics of Aging', NBER, pp. 61–82.

Manoli, D. and Weber, A. (2016*a*), 'Nonparametric Evidence on the Effects of Financial Incentives on Retirement Decisions', *American Economic Journal: Economic Policy* **8**(4), 160–182.

Manoli, D. and Weber, A. (2016*b*), 'The Effects of the Early Retirement Age on Retirement Decisions'. NBER Working Paper No. 22561.

Mastrobuoni, G. (2009), 'Labor Supply Effects of the Recent Social Security Benefit Cuts: Empirical Estimates Using Cohort Discontinuities', *Journal of Public Economics* **93**(11–12), 1224–1233.

Merkle, C., Schreiber, P. and Weber, M. (2017), 'The Willingness to Pay, Accept, and Retire', *Economic Policy* **32**(92), 757–809.

Monteiro, J. C. and Assunção, J. J. (2012), 'Coming Out of the Shadows? Estimating the Impact of Bureaucracy Simplification and Tax Cut on Formality in Brazilian Microenterprises', *Journal of Development Economics* **99**(1), 105–115.

OECD (2015), 'Pensions at a Glance 2015: Germany'. Unpublished Report.

Onji, K. (2009), 'The Response of Firms to Eligibility Thresholds: Evidence from the Japanese Value-Added Tax', *Journal of Public Economics* **93**(5), 766–775.

Pomeranz, D. (2015), 'No Taxation Without Information: Deterrence and Self-Enforcement in the Value Added Tax', *The American Economic Review* **105**(8), 2539–2569.

Rees-Jones, A. (2018), 'Quantifying Loss-Averse Tax Manipulation', *Review of Economic Studies* **85**(2), 1251–1278.

Rios, J. and Setharam, I. (2018), 'Propagating Formality via Value Added Tax Networks: Evidence from India'. Working Paper.

Saez, E. (2010), 'Do Taxpayers Bunch at Kink Points?', *American Economic Journal: Economic Policy* **2**(3), 180–212.

Shome, P. (2004), *Tax Administration and the Small Taxpayer*, International Monetary Fund.

Shoven, J. B., Slavov, S. N. and Wise, D. A. (2017), 'Social Security Claiming Decisions: Survey Evidence'. NBER Working Paper No. 23729.

Slemrod, J. and Gillitzer, C. (2013), *Tax Systems*, MIT Press.

Staubli, S. and Zweimüller, J. (2013), 'Does Raising the Early Retirement Age Increase Employment of Older Workers?', *Journal of Public Economics* **108**, 17–32.

Steenbergen, V. (2017), 'Reaping the Benefits of Electronic Billing Machines'. IGC Working Paper.

Stock, J. H. and Wise, D. A. (1990), 'Pensions, the Option Value of Work, and Retirement', *Econometrica* **58**(5), 1151–1180.