Title
Earnings Adjustment Frictions: Evidence from the Social Security Earnings Test

Authors
Gelber, Alexander M.
Jones, Damon
Sacks, Daniel W.

Publication Date
2016-08-01

IRLE
Institute for Research on Labor and Employment

IRLE WORKING PAPER
#117-16
August 2016

Earnings Adjustment Frictions: Evidence from the Social Security Earnings Test

Alexander M. Gelber, Damon Jones, and Daniel W. Sacks

Cite as: Alexander M. Gelber, Damon Jones, and Daniel W. Sacks. (2016). "Earnings Adjustment Frictions: Evidence from the Social Security Earnings Test". IRLE Working Paper No. 117-16.
http://irle.berkeley.edu/workingpapers/117-16.pdf

Berkeley
UNIVERSITY OF CALIFORNIA

Earnings Adjustment Frictions: Evidence from the Social Security Earnings Test

Alexander M. Gelber, Damon Jones, and Daniel W. Sacks*

August 2016

Abstract

We study frictions in adjusting earnings in response to changes in the Social Security Annual Earnings Test (AET), using a one percent sample of earnings histories from Social Security Administration microdata from 1983 to 1999. We introduce a novel method for documenting adjustment frictions: individuals continue to "bunch" at the convex kink the AET creates even when they are no longer subject to the AET. We develop a framework for estimating an earnings elasticity and an adjustment cost using information on the amount of bunching at kinks before and after policy changes in earnings incentives around the kinks. We apply this method in settings in which individuals face changes in the AET benefit reduction rate, and we estimate in a baseline case that the earnings elasticity with respect to the implicit net-of-tax share is 0.35, and the fixed cost of adjustment is around $280. Our results demonstrate that the short-run impact of changes in the effective marginal tax rate can be substantially attenuated.

*Gelber: UC Berkeley Goldman School of Public Policy and NBER, agelber@berkeley.edu; Jones: University of Chicago, Harris School of Public Policy and NBER, damonjones@uchicago.edu; Sacks: Indiana University, Kelley School of Business, dansacks@indiana.edu. We thank Jérôme Adda (the Editor), Raj Chetty, Jim Cole, Jim Davis, Mark Duggan, Jonathan Fisher, Richard Freeman, John Friedman, Bill Gale, Hilary Hoynes, Adam Isen, Henrik Kleven, Olivia Mitchell, Emmanuel Saez, Chris Walters, numerous seminar participants, and four anonymous referees for helpful comments. We are extremely grateful to David Pattison for generously running the code on the data. We acknowledge financial support from the Wharton Center for Human Resources and the Wharton Risk and Decision Processes Center, from NIH grant #1R03 AG043039-01, from a National Science Foundation Graduate Research Fellowship, and from support from the U.S. Social Security Administration through grant #5RRC08098400-05-00 to the National Bureau of Economic Research (NBER) as part of the SSA Retirement Research Consortium. The findings and conclusions expressed are solely those of the authors and do not represent the views of SSA, any agency of the Federal Government, or the NBER. The research uses data from the Census Bureau's Longitudinal Employer Household Dynamics Program, which was partially supported by the following National Science Foundation Grants: SES-9978093, SES-0339191 and ITR-0427889; National Institute on Aging Grant AG018854; and grants from the Alfred P. Sloan Foundation. All results have been reviewed to ensure that no confidential information is disclosed. All errors are our own.

1 Introduction

In a traditional model of workers' earnings or labor supply choices, individuals optimize their behavior frictionlessly. Recently, several papers have suggested that individuals face frictions in adjusting behavior to policy (Chetty, Looney, and Kroft, 2009; Chetty, Friedman, Olsen, and Pistaferri, 2011; Chetty, Guren, Manoli, and Weber, 2012; Chetty, Friedman, and Saez, 2012; Chetty, 2012; Kleven and Waseem, 2013). Adjustment frictions, which we interpret broadly to encompass factors preventing individuals from adjusting their earnings, may reflect a variety of elements including lack of knowledge of a tax regime, the cost of negotiating a new contract with an employer, or the time and financial cost of job search. These frictions could impede immediate or long-term adjustments to tax policy changes. Such adjustment frictions can also affect the welfare consequences of taxation. For example, if taxes are not fully salient, this must be measured to calculate the welfare costs of taxation (Chetty *et al.,* 2009; Farhi and Gabaix, 2015). Adjustment frictions also help to explain heterogeneity across contexts in the observed elasticity of earnings with respect to the net-of-tax rate (Chetty *et al.,* 2011, 2012b; Chetty, 2012).[1] Frictions in adjusting earnings may underlie other patterns in the data, such as the slow rise in retirement at age 62 subsequent to the introduction of the Social Security Early Retirement Age (Gruber, 2013), or the lack of "bunching" at many kink points in budget sets (Chetty, Friedman, Olsen, and Pistaferri, 2011).

This paper develops evidence on the existence and size of earnings adjustment frictions. The U.S. Social Security Annual Earnings Test (AET) represents a promising environment for studying these questions. This setting provides an illustration of issues—including the development and application of a methodology for documenting adjustment frictions, and for estimating elasticities and adjustment costs simultaneously—that may be applicable to studying adjustment frictions more broadly. The AET reduces Social Security Old Age and Survivors Insurance (OASI) benefits in a given year as a proportion of an OASI claimant's earnings above an exempt amount in that year. For example, for OASI claimants aged 62-65 in 2016, current OASI benefits are reduced by 50 cents for every dollar earned above $15,720.

[1] The net-of-tax rate is defined as one minus the marginal tax rate (MTR). Labor economics literature examines hours constraints in the context of labor supply (*e.g.* Cogan, 1981; Altonji and Paxson, 1990, 1992; Dickens and Lundberg, 1993).

The AET may lead to large effective benefit reduction rates (BRRs) on earnings above the exempt amount, creating a strong incentive for many individuals to bunch at the convex kink in the budget constraint located at the exempt amount (Burtless and Moffitt, 1985; Friedberg, 1998, 2000; Song and Manchester, 2007; Engelhardt and Kumar, 2014).

The AET is an appealing context for studying earnings adjustment for at least three reasons. First, bunching at the AET kink is easily visible on a graph, allowing credible documentation of behavioral responses.[2] Second, the AET represents one of the few known kinks at which bunching occurs in the U.S.; indeed, our paper represents the first study to find robust evidence of sharp bunching at the intensive margin among the non-self-employed at any kink in the U.S.[3] Third, the AET is important to policy-makers in its own right, as it is a significant factor affecting the earnings of the elderly in the U.S.

We make two main contributions to understanding adjustment frictions in the earnings context. First, we develop a new method for documenting earnings adjustment frictions and show that such frictions exist in the U.S. We focus particularly on cases in which a kink in the effective tax schedule disappears.[4] In the absence of adjustment frictions, the removal of a convex kink in the effective tax schedule should result in the immediate dissolution of bunching at the former kink; thus, any observed delay in reaching zero bunching should reflect adjustment frictions. We observe clear evidence of delays: individuals continue to bunch around the location of a former kink. Nonetheless, the vast majority of individuals' adjustment occurs within at most three years. Thus, we interpret the frictions we observe as evidence of barriers to making an *immediate* adjustment in response to changes in incentives. These findings are similar in spirit to those of Best and Kleven (2014) in the housing market context, who find evidence of delays in adjustment but also find that adjustment occurs quickly, on the order of 3-4 months. We provide suggestive evidence consistent with the hypothesis that these frictions are driven by the costs of finding a new employment arrangement, rather than the informational costs of learning about a new tax regime.

Second, we specify a model of earnings adjustment that allows us to estimate a fixed

[2] Other papers have examined bunching in the earnings schedule, including Blundell and Hoynes (2004) and Saez (2010).

[3] Chetty *et al.* (2012) find evidence of diffuse earnings responses to the Earned Income Tax Credit among the non-self-employed in the U.S.

[4] For consistency with the previous literature on kink points that has focused on the effect of taxation, we sometimes use "tax" as shorthand for "tax-and-transfer," while recognizing that the AET reduces Social Security benefits and is not administered through the tax system. The "effective" marginal tax rate is potentially affected by the AET BRR, among other factors.

adjustment cost and the elasticity of earnings with respect to the effective net-of-tax rate. Recent work demonstrating the importance of earnings adjustment frictions has raised the question of how to estimate both the elasticity and adjustment cost simultaneously. Adding adjustment frictions to the model of Saez (2010), we develop tractable methods that allow estimation of elasticities and adjustment costs with kinked budget sets. Our method complements Kleven and Waseem (2013), who innovate a method to estimate elasticities and the share of the population that is inert in the presence of a notch in the budget set (but do not estimate adjustment costs). To our knowledge, our method is the first to allow estimation of both elasticities and adjustment costs using bunching in earnings. The elasticity and adjustment cost are both necessary for welfare calculations in many applications (Chetty *et al.*, 2009). Our method is also applicable in a different context than that of Kleven and Waseem (2013): ours relies on bunching at kinks, rather than notches, to perform the estimates. Finally, we present a dynamic version of our model, which extends current bunching techniques beyond the typical static approach and allows us to address gradual adjustment in bunching over time.

Our paper also complements Chetty (2012), who derives bounds on the "structural" elasticity as a function of the elasticity observed empirically, the size of the price change used for identification, and the degree of optimization frictions. Chetty (2012) uses these theoretical results, in combination with estimates of observed elasticities from prior empirical literature, to calculate bounds on the structural labor supply elasticity given an assumption about the utility cost of ignoring policy, but that paper does not document frictions or introduce methods for estimating elasticities and adjustment costs together. That exercise naturally leads to the question of how to estimate elasticities and adjustment costs together using data; our paper is the first in this literature to introduce a method that accomplishes this. Kleven's (2016) survey of bunching papers discusses the fact that the new method we develop here is applicable to estimating elasticities more broadly in contexts beyond the AET. Indeed, our method has been used by Gudgeon and Trenkle (2016) and He, Peng, and Wang (2016) to estimate elasticities in other contexts.

Our method relies on clear patterns in the data. In our model, the amount of bunching at a newly introduced kink increases with the elasticity (as a higher elasticity will induce

more individuals to locate at the kink) but decreases with the adjustment cost (as the adjustment cost prevents bunching among some individuals). This prevents estimation of both parameters using a single cross-section—since a small amount of bunching, for example, could be consistent with either a low elasticity or a high adjustment cost. However, with two or more cross-sections of individuals facing different tax rates in the region of the kink, we can specify two or more equations and find the values of two variables (the elasticity and the adjustment cost). All else equal, the amount of bunching in each cross-section is increasing in the elasticity, but the absolute value of the change in bunching is decreasing in the adjustment cost. Intuitively, these patterns help us to identify the adjustment cost, as well as the elasticity.

We apply our method to data spanning the decrease in the AET BRR from 50 percent to 33.33 percent in 1990 for those aged 66 to 69, as well as a setting in which the AET ceases to apply, when moving from age 69 to ages 70 and older in the 1990-1999 period. In a baseline specification examining the 1990 change, we estimate that the fixed adjustment cost is around $280 (in 2010 dollars)—if the gains exceed this level, then the individual adjusts earnings—and that the earnings elasticity with respect to the net-of-tax share is 0.35. This specification examines data on individuals in 1989 and 1990; thus, our estimated adjustment cost represents the cost of adjusting earnings in the year of the policy change. Other strategies—including an extension of our method that allows for dynamic adjustment over time—show results in the same range. Just following the reduction in the kink in 1990, if we constrain the adjustment cost to be zero and thus fail to account for excess bunching following the policy change due to inertia, we instead estimate a statistically significantly higher earnings elasticity of 0.58, due to delayed adjustment of excess bunching. Although the constrained and unconstrained elasticities are only moderately different in absolute value, the percentage difference between the elasticities is large, as the constrained estimate is 66 percent higher.

Our estimates suggest that while adjustment costs are modest in our setting, they have the potential to change earnings elasticity estimates significantly, illustrating that it can be important to incorporate adjustment costs when estimating earnings elasticities. Our estimates apply to the population bunching at kinks; it is particularly striking that we find

evidence of adjustment frictions even among those initially bunching at the kink, whose initial bunching may indicate flexibility. By demonstrating that earnings adjustment frictions exist and substantially change elasticity estimates even in this setting, our results suggest the importance of taking them into account in other settings. Our results show that adjustment frictions can substantially attenuate short-run earnings reactions even to large changes in the effective marginal tax rate, frustrating the goal of affecting short-run earnings as envisioned in many recent discussions of tax policy.

Our paper follows a large existing literature on adjustment costs in areas outside labor and public economics. For example, adjustment costs have been studied in inventory theory (Arrow *et al.*, 1951), macroeconomics (*e.g.* Baumol, 1952; Blinder, 1981; Caplin, 1985; Caplin and Spulber, 1987; Caplin and Leahy, 1991), firm investment (*e.g.* Abel and Eberly, 1994; Caballero and Engel, 1999; Attanasio, 2000; Cooper and Haltiwanger, 2006), durable good consumption (Grossman and Laroque, 1990), pricing and inflation (*e.g.* Sheshinkski and Weiss, 1977), and other settings. Relative to this literature, we make several contributions. First, we explore these issues in the context of earnings determination, and we exploit changes in policies creating effective tax rates.[5] Second, we offer new, transparent evidence for earnings adjustment frictions by showing that bunching persists after a kink has been removed. This method for documenting the presence of adjustment frictions could be more broadly applicable in other economic contexts in which budget set kinks are removed. Such non-linear budget set kinks occur not only in the context of earnings determination but have also been studied in other economic applications with non-linear pricing (*e.g.* Reiss and White, 2005). Third, we introduce methods for estimating adjustment costs and elasticities by exploiting bunching at non-linear budget set kinks.

The primary focus of the paper relates to developing methods for studying adjustment frictions. A secondary contribution of the paper is to provide new evidence on the effects of the AET in particular. We use SSA administrative data with a sample of 376,431 observations in our main period, building on previous studies of the AET that use survey data. Our study is the first to estimate bunching in the context of the AET through a method similar

[5] Within public finance, Marx (2015) also examines bunching in the reporting of revenue by charities within a dynamic context, while Werquin (2015) derives a continuous-time, "s-S" model of taxable earnings in the presence of adjustment costs. The former paper does not feature adjustment frictions, while the latter abstracts from kinks in the tax schedule and bunching in earnings.

to Saez (2010).

The remainder of the paper is structured as follows. Section 2 describes the policy environment. Section 3 describes our empirical strategy for quantifying bunching. Section 4 describes our data. Section 5 presents empirical evidence on the earnings response to changes in the AET. Section 6 specifies a tractable model of earnings adjustment. Section 7 estimates the fixed adjustment cost and elasticity simultaneously. Section 8 concludes.

2 Policy Environment

Figure 1 shows key features of the AET rules from 1961 to 2009. The AET became less stringent over this period. The dashed lines and right vertical axis show the BRR. From 1961 to 1989, an additional dollar of earnings above the exempt amount reduced OASI benefits by 50 cents (until OASI benefits reached zero). In 1990 and after, the BRR fell to 33.33 percent for beneficiaries at or older than the Normal Retirement Age (NRA); this change had been scheduled since the 1983 Social Security Amendments. The NRA, the age at which workers can claim their full OASI benefits, is 65 for those in the 1983 to 1999 period that we focus on. During this period, the AET applied to OASI beneficiaries aged 62-69.[6] The solid lines and left vertical axis show the real exempt amount. Starting in 1978, beneficiaries younger than NRA faced a lower exempt amount than those at NRA or above.

When current OASI benefits are lost to the AET, future scheduled benefits are increased in some circumstances, which is sometimes called "benefit enhancement." This can reduce the effective tax rate associated with the AET. For beneficiaries subject to the AET aged NRA and older, a one percent Delayed Retirement Credit (DRC) was introduced in 1972, meaning that each year of foregone benefits led to a one percent increase in future yearly benefits. The DRC was raised to three percent in 1982 and gradually rose to eight percent for cohorts reaching NRA from 1990 to 2008 (though the AET was eliminated in 2000 for those older than the NRA). An increase in future benefits between seven and eight percent is approximately actuarially fair on average, meaning that an individual with no liquidity constraints and average life expectancy should be indifferent between either claiming benefits now or delaying claiming and receiving higher benefits once she begins to collect OASI

[6] Prior to 1983, the AET applied to beneficiaries aged 62 to 71.

(Diamond and Gruber, 1999).

OASI claimants' future benefits are only raised due to the DRC when annual earnings are sufficiently high that the individual loses an entire month's worth of OASI benefits due to the reductions associated with the AET (Friedberg, 1998; Social Security Administration, 2012a). In particular, an entire month's benefits are lost—and benefit enhancement occurs—once the individual earns $z^* + (MB/\tau)$ or higher, where z^* is the annual exempt amount, MB is the monthly benefit, and τ is the AET BRR. With a typical monthly benefit of \$1,000 and a BRR of 33.33 percent, one month's benefit enhancement occurs when the individual's annual earnings are \$3,000 (=\$1,000/0.3333) above the exempt amount. As a result, benefit enhancement is only relevant to an individual considering earning substantially in excess of the exempt amount. Although the AET withholds benefits at the monthly level, the AET is generally applied based on *annual* earnings—the object we observe in our data. We model the AET as creating a positive implicit marginal tax rate for some individuals—reflecting the reduction in current benefits—consistent with both the empirical finding that some individuals bunch at AET kinks and with the practice in previous literature.

For individuals considering earning in a region well above the AET exempt amount, thus triggering benefit enhancement, the AET could also affect decisions for several reasons. The AET was roughly actuarially fair only beginning in the late 1990s. Furthermore, those whose expected life span is shorter than average should expect to collect OASI benefits for less long than average, implying that the AET is more financially punitive. Liquidity-constrained individuals or those who discount faster than average could also reduce work in response to the AET. Finally, many individuals also may not understand the AET benefit enhancement or other aspects of OASI (Liebman and Luttmer, 2011; Brown, Kapteyn, Mitchell, and Mattox, 2013). We follow previous work and do not distinguish among these potential reasons for a response to the AET in our main analysis.

For beneficiaries under NRA, the actuarial adjustment raises future benefits whenever an individual earns over the AET exempt amount (Social Security Administration, 2012, Section 728.2; Gruber and Orszag, 2003), by 0.55 percent per month of benefits withheld. Thus, beneficiaries in this age range do not face a pure kink in the budget set at the exempt amount. To address this, we limit the sample to ages above NRA in our estimates of elasticities and

adjustment costs.

3 Initial Bunching Framework

As a preliminary step, we begin with a model with no frictions to illustrate our technique for estimating bunching, which is also suited to measuring the bunching at the kink arising in a model with frictions. This model is well-known and described in detail in Saez (2010), but we briefly describe it in preparation for our initial descriptive evidence.[7] Agents maximize utility $u(c, z; a)$ over consumption c and pre-tax earnings z (where greater earnings are associated with greater disutility due to the cost of effort), subject to a budget constraint $c = (1-\tau)z + R$, where R is virtual income.[8] Agents can adjust earnings, for example, through a change in hours worked or effort, or in principle though a change in the reporting of earnings. The first-order condition, $(1-\tau)u_c + u_z = 0$, implicitly defines an earnings supply function $z((1-\tau), R; a)$.

The parameter a reflects heterogeneous "ability," *i.e.* the trade-off between consumption and earnings supply. Following previous literature, we assume rank preservation in earnings as a function of a.[9] Thus, a is isomorphic to the level of earnings that would occur in the absence of any tax.[10] Ability, a, is distributed according a smooth CDF. Under a constant marginal tax rate of τ_0, this implies a smooth distribution of earnings $H_0(\cdot)$, with pdf $h_0(\cdot)$.

Starting with a linear tax at a rate of τ_0, suppose the AET is additionally introduced, so that the marginal net-of-tax rate decreases to $1-\tau_1$ for earnings above a threshold z^*, where $\tau_1 > \tau_0$. Individuals earning in the neighborhood above z^* reduce their earnings. If ability is smoothly distributed, a range of individuals initially locating between z^* and $z^* + \triangle z^*$ will "bunch" exactly at z^*, due to the discontinuous jump in the marginal net-of-tax rate at z^*. In practice, previous literature finds empirically that individuals locate in the neighborhood

[7] Saez (2010) follows earlier work on estimation of labor supply responses on nonlinear budget sets, including Burtless and Hausman (1978) and Hausman (1981). Moffitt (1990) surveys these methods.

[8] More generally, we can write $c = z - T(z)$, where $T(z)$ is a general, nonlinear tax schedule. As is customary in the public finance literature (*e.g.* Hausman, 1981) we rewrite the budget constraint in linearized form, $c = (1-\tau)z + R$, where $\tau \equiv T'(z)$ is the marginal tax rate and $R \equiv T'(z) \cdot z - T(z)$ is virtual income, *i.e.* the intercept of a linear budget set that passes through the point $(z, T(z))$. Hausman (1981) shows that the optimal earnings response to this linearized tax schedule and the nonlinear tax schedule are locally equivalent.

[9] This rank preservation is a direct implication of the Spence-Mirrlees, single-crossing assumptions generally made in the optimal taxation literature.

[10] For example, if we assume a standard isoelastic and quasilinear utility function, $u(c, z; a) = c - (a/(1+1/\varepsilon))(z/a)^{1+1/\varepsilon}$, the optimal level of earnings is $z((1-\tau), R; a) = a(1-\tau)^{\varepsilon}$. Thus, when $\tau = 0$, we have $z = a$.

of z^* (rather than exactly at z^*).

To quantify the amount of bunching, *i.e.* "excess mass," we use a technique similar to Chetty *et al.* (2011) and Kleven and Waseem (2013). For each earnings bin z_i of width δ we calculate p_i, the proportion of all people with annual earnings in the range $[z_i - \delta/2, z_i + \delta/2)$. We then run the following regression:

$$p_i = \sum_{d=0}^{D} \beta_d (z_i - z^*)^d + \sum_{j=-k}^{k} \gamma_j 1\{z_i - z^* = j \cdot \delta\} + u_i \tag{1}$$

This expresses the annual earnings distribution as a degree D polynomial, plus a set of indicators for each bin with a midpoint within $k\delta$ of the kink.

Our measure of excess mass, or bunching, is $\hat{B} = \sum_{j=-k}^{k} \hat{\gamma}_j$, the estimated excess probability of locating at the kink (relative to the polynomial fit). To obtain a measure of excess mass that is comparable across different kinks, we scale by the counterfactual density at z^*, *i.e.* $\hat{h}_0(z^*) = \hat{\beta}_0/\delta$. Thus, our estimate of "normalized excess mass" is $\hat{b} = \hat{B} \Big/ \hat{h}_0(z^*) = \delta \hat{B} \Big/ \hat{\beta}_0$. In our empirical application, we choose $D = 7$, $\delta = 800$ and $k = 4$ as a baseline, implying that our estimate of bunching is driven by individuals with annual earnings within \$3,600 of the kink. We also show our results under alternative choices of D, δ, and k. We estimate bootstrapped standard errors.

4 Data

We primarily rely on a one percent random sample of Social Security numbers from the restricted-access Social Security Administration Master Earnings File (MEF), linked to the Master Beneficiary Record (MBR). The data contain a complete longitudinal earnings history with yearly information on earnings since 1951; the type and amount of yearly Social Security benefits an individual receives; year of birth; the year (if any) that claiming began; and sex (among other variables). Separate information is available on self-employment earnings and non-self-employment earnings. Starting in 1978, the earnings measure reflects total wage compensation, as reported on Internal Revenue Service forms.

In choosing our main sample, we take into account a number of considerations. It is desirable to show a constant sample in making comparisons of earnings densities. Meanwhile, the

AET only affects people who claim OASI, and thus we wish to focus on claimants. However, many individuals claim OASI at ages older than the Early Entitlement Age (EEA) of 62. Thus, to investigate a constant sample, we cannot simply limit the sample to claimants at each age, as many people move from not claiming to claiming. To balance these considerations, our main sample at each age and year consists of individuals who have ultimately claimed at an age less than or equal to 65. In our main analysis we exclude person-years with positive self-employment income. Because we focus on the intensive margin response, in our main analysis we further limit the sample in a given year to observations with positive earnings in that year.[11]

Several features of the data are worth noting. First, these administrative data allow large sample sizes and are subject to little measurement error. Second, earnings (as measured in the dataset) are taken from W-2 tax forms and are not subject to manipulation through tax deductions, credits, or exemptions. Third, because earnings are taken from the W-2 form, they are subject to third-party reporting (among the non-self-employed). Fourth, the data do not contain information on hours worked or amenities at individuals' jobs.

Table 1 shows summary statistics in our main sample for our main age and year range, 62-69 year-olds in 1990-1999. The sample has 376,431 observations, of which 57 percent is male. Mean earnings (conditional on positive earnings) is $28,892.63. Median earnings, $14,555.56, is not far from the AET exempt amount, which averages $16,738 for those NRA and older and $11,650 for those younger than NRA over this period.

Our second data source is the Longitudinal Employer Household Dynamics (LEHD) of the U.S. Census (McKinney and Vilhuber, 2008; Abowd *et al.*, 2009), which longitudinally follows workers' earnings. In covered states, the data have information on around nine-tenths of workers and their employers. We are only able to use data on a 20 percent random subsample of these individuals from 1990 to 1999. We use these data primarily because the sample size in the LEHD is much larger than in the SSA data. We use the LEHD only in the context of two figures (4 and 5) for which the larger sample is helpful; all other analysis is based on the SSA data.[12]

[11] We explore extensive margin decisions in Gelber, Jones, Sacks, and Song (2016).

[12] The LEHD lacks information on whether a given individual is claiming OASI, but the importance of this shortcoming is limited because we use the LEHD to study the evoluation of earnings from ages 69 to 71. In our SSA data, 94 percent of individuals claim by age 69.

5 Earnings Response to Policy Variation Across Ages

5.1 Primary Empirical Results

We first examine the pattern of bunching across ages. In this case, we focus on the period 1990-1999, when the AET applied from ages 62 to 69. The policy changes at ages 62 and 70—when the AET is imposed and removed, respectively—are "anticipated," in the sense that they would be anticipated by those who have knowledge of the relevant policies. Figure 2 plots earnings histograms for each age from 59 to 73 (connected dots), along with the estimated smooth counterfactual polynomial density (smooth line). Earnings are measured along the x-axis, relative to the exempt amount, which is shown by a vertical line. For ages younger than 62, we define the (placebo) kink in a given year as the kink that applies to pre-NRA individuals in that year. For individuals 70 and older, we define the (placebo) kink in a given year as the kink that applies to post-NRA individuals in that year.

Figure 2 shows clear visual evidence of substantial bunching from ages 62-69, when the AET is in effect, and no excess mass at earlier ages. At ages 70 and 71, which are not subject to the AET, there is still clear visual evidence of bunching in the region of the kink.

Figure 3 plots the estimates of normalized excess mass at each age. Bunching is statistically significantly different from zero at each age from 62 to 71 ($p < 0.01$ at all ages). Normalized excess mass rises from 62 to 63 and remains around this level until age 69 (with a dip at age 65 that we discuss below). We estimate that there is substantial excess mass at ages 70 and 71, which are not subject to the AET. Thus, "de-bunching" does not occur immediately for some individuals, where "de-bunching" refers to movement away from the former kink among those initially bunching at the kink.

Figures 4 and 5 show spikes near the exempt amount in the mean percentage change in earnings from ages 69 to 70 and 70 to 71, respectively, consistent with de-bunching from age 69 to 70, and from age 70 to 71, among those initially near the kink. This shows that those bunching are returning to higher earnings, as predicted by theory, and that this process continues at least until age 71. It is striking that we document adjustment frictions even among the group bunching prior to age 70, who were evidently able to adjust earnings to

the kink initially.

We classify claimants at age 70 based on the highest age they attain in the calendar year. As a result, some individuals will be classified as age 70 but will have been subject to the AET for a portion of the year (in the extreme case of a December 31 birthday, for all but one day). In principle this is one potential explanation for continued bunching at age 70 that does not rely on earnings adjustment frictions. However, other evidence is sufficient to document earnings adjustment frictions, namely: (1) the continued bunching at age 71, which cannot be explained through the coarse measure of age; (2) the continued adjustment away from the kink from age 70 to age 71 documented in Figure 5; and (3) the spike in the elasticity estimated using the Saez (2010) approach in 1990, documented in Figure 10 and explained below. Moreover, Appendix Table B.1 shows that those born in January to March—who are not subject to the AET for nearly the entire calendar year when they are age 70—also show statistically significant bunching at ages 70 ($p<0.05$) and 71 ($p<0.10$) from 1983 to 1999.[13] Finally, when we pool data from 1983 to 1999 in Appendix Figure B.1 and Panel B of Appendix Figure B.3—giving us more power than in our baseline sample over 1990 to 1999 when the AET does not change—bunching above age 70 is even more visually apparent, and excess mass at age 71 is highly significant and clearly greater than zero.

Figure 3 shows that bunching is substantially lower at age 65 than surrounding ages. The location of the kink changes substantially from age 64 to age 65; as Figure 1 shows, during this period the exempt amount is much higher for individuals NRA and older than for individuals younger than NRA. Individuals may have difficulty adjusting to the new location of the kink within one year. Prior to the divergence of the exempt amount for those younger and older than the NRA in 1978, we find no such dip in bunching at age 65; this "placebo" evidence further supports the hypothesis that the dip in bunching at age 65 arises from delayed adjustment to the increase in the exempt amount from ages 64 to 65 that emerges after 1978. This delay suggests that individuals also face adjustment frictions in this context. This interpretation of the patterns around ages 64 and 65 is consistent with Figure 6, which shows that conditional on age 64 earnings near the age 64 exempt amount, the age 65 earnings density shows a large spike at the kink that prevailed at age 64 and a

[13] Limiting the sample only to those born in January yields insignificant and imprecise results.

smaller spike at the current, age 65 kink. Also, conditional on age 65 earnings near the age 65 exempt amount, the density of age 64 earnings shows a spike near the exempt amount for age 64.[14]

In our context, the only "appearance" of a new kink that we observe is the appearance of a kink at age 62. The amount of time since the appearance of the kink at age 62 is correlated with age, and elasticities and adjustment costs could also be correlated with age—thus confounding analysis of the time necessary to adjust to the appearance of a kink. While recognizing these caveats, it is worth noting that the amount of bunching slowly rises from age 62 to 63, which suggests gradual adjustment.[15]

Each of these several pieces of evidence points to delayed adjustment. In Appendix Table B.2, we probe the robustness of these results by varying the bandwidth, the degree of the polynomial, and the excluded region. We conduct several additional analyses in Gelber, Jones, and Sacks (2013), including varying the time period examined. Overall, these additional analyses usually show similar patterns to our baseline. In a number of cases these estimates lose significance, but the weight of the evidence points toward adjustment frictions.

We find no evidence of adjustment in anticipation of future changes in policy, as those younger than 62 do not bunch. If the cost of adjustment in each year rose with the size of adjustment and this relationship were convex, we would expect anticipatory adjustment. Other literature on earnings adjustment frictions has shown that firms are important in coordinating bunching responses to taxation in Denmark (Chetty *et al.,* 2011), by documenting bunching among individuals not subject to the taxes. In our context, individuals younger than 62—who are not subject to the AET—do not show noticeable bunching at the kink, nor do those 72 and older (Figure 2). While we cannot rule out that firms play some role in our context, the available evidence in our sample does not directly support this hypothesis. Thus, we do not interpret continued bunching at ages 70 and 71 as relating to firm choices.[16]

[14] In principle, our coarse measure of age could affect these patterns: individuals turning 65 in a given calendar year face the age-65 exempt amount for only the part of the calendar year after they turn 65, which could serve as a partial explanation for continued bunching at age 65 at the exempt amount applying to age 64. However, we would then expect the age 64 and age 65 exempt amounts to display equal amounts of bunching, which is not the case.

[15] In principle, this could also relate to the fact that these graphs show the sample of those who have claimed by age 65, and the probability of claiming at a given age (conditional on claiming by age 65) rises from age 62 to 63. To address this issue, Appendix Figure B.2 shows that when the sample at a given age consists of those who have claimed by that age, we still find a substantial increase in bunching from 62 to 63.

[16] The fact that individuals often appear to jump from the age 64 kink to the age 65 kink also does not necessarily imply that the earnings menus offered by firms are the only factor driving individuals from one kink to another. Individuals could be

We interpret the continued bunching at ages 70 and 71 as reflecting frictions preventing adjustment. If this is the case, those bunching after the kink is removed should have been bunching prior to the removal (and those bunching before the kink is removed should be disproportionately represented among those bunching after the removal). A degree of such inertia has already been shown at ages 64 and 65 in Figure 6. Figure 7 further shows that indeed, conditional on earnings at ages 70 or 71 within $1,000 of the exempt amount, the density of earnings at age 69 spikes at the exempt amount (and conditional on earnings at age 69 within $1,000 of the exempt amount, the density of age 70 or age 71 earnings spikes near the exempt amount).

5.2 Suggestive Evidence on Source of Frictions

Frictions could fall into two broad categories: those relating to information or salience (*e.g.* not knowing about the policy or about changes in policy), and those relating to more concrete costs of changing hours or jobs (*e.g.* the time and financial cost of job search or contract renegotiation). Previous literature on estimating adjustment frictions typically has not distinguished these two sources of frictions. Like previous literature, our method of examining continued bunching at former kinks is well suited to document that individuals face adjustment frictions, but it is less well equipped to directly address the particular mechanisms that underlie these frictions.

Before examining our data, it is worth noting that though they did not estimate frictions using data, Liebman and Luttmer (2012) found descriptively that individuals often are not aware of AET parameters, raising the possibility that information costs may also be important when AET parameters change over time or across ages. To provide further suggestive evidence on the source of frictions in our setting, we examine two sources of evidence.

First, the self-employed typically have much more control over the number of hours they work than wage earners—for example, a self-employed small business owner can choose to work more hours in their business, but a wage earner may work a fixed number of hours for their employer like 40 hours per week (Levine and Rubinstein, 2013). Thus, if the costs of

driving this movement on their own. In particular, when the distance between the two kinks is less than normalized bunching b—which is true on average in our sample—it can be shown in the Saez (2010) model that the set of bunchers at both kinks should overlap.

negotiating a new contract with an employer or finding a new job with the desired number of hours are important drivers of adjustment frictions, then we would expect faster adjustment among the self-employed than among wage earners. Consistent with this hypothesis, Appendix Figure B.3 shows that although excess normalized mass is positive at ages 67 to 69 among wage earners and the self-employed, excess normalized mass dissipates much more quickly among the self-employed than among wage earners: excess normalized mass is statistically indistinguishable from zero at age 70 for the self-employed, but this takes until age 72 among wage earners. Consistent with these results, when we estimate our model below on the self-employed sample alone, we estimate smaller adjustment costs than in the wage earner sample alone.

Second, if informational frictions are important, then we might expect slower adjustment while individuals are learning about the AET than once they have been subject to the AET and may have become more familiar with it. In Figure 3 we observe that when individuals are first subject to the AET starting at the EEA of 62, normalized excess mass increases from age 61 to 62 and from 62 to 63, before roughly plateauing from 63 to 64.[17] This adjustment process does not take longer than the dissipation in bunching from ages 69 to 72, by which point these individuals have already been subject to the AET. This could suggest that informational frictions are not as important in explaining the results, as individuals do not respond more slowly to the introduction of the AET than to its removal. Indeed, if we estimate our dynamic model below on claimants at ages 61 to 63 and ages 69 to 71, we estimate adjustment costs that are similar, and insignificantly different, in the two cases (available upon request).

Both of these results point toward frictions like the costs of negotiating a new contract with an employer or finding a new job with the desired number of hours, rather than informational frictions, in explaining our results. However, this evidence is suggestive, not dispositive. The self-employed may be different from wage earners in ways other than the frictions they face. Moreover, we do not directly observe when individuals learn about the AET or what they learn, and it is possible that individuals in their early 60s may be differ-

[17] We observe a similar speed of adjustment in Appendix Figure B.2 when we limit the sample in any given year to those who have claimed by that year. Like Appendix Figure B.3, Appendix Figure B.2 shows evidence of a small amount of continued bunching past age 71. This also strengthens the case that adjustment frictions exist. Our main overall conclusions are that (1) adjustment frictions exist and (2) nearly all bunching dissipates within a few years.

ent from those in their early 70s in ways that would affect their speed of learning (or that individuals initially learn about certain features of the AET but not about its removal at age 70). Despite these important caveats, these two pieces of suggestive evidence work together because they are both consistent with the same source of adjustment frictions.

6 Estimation Method

The results thus far suggest a role for adjustment frictions in individuals' earnings choices. To develop a method to estimate elasticities and adjustment costs jointly, we build on the frictionless Saez (2010) model described in Section 3. There we considered a transition from a linear tax schedule with a constant marginal tax rate (MTR) τ_0 to a schedule with a convex kink, where the MTR below the kink earnings level z^* is τ_0, and the MTR above z^* is $\tau_1 > \tau_0$. We refer to this kink at z^* as K_1. Next, as in our empirical context, consider a decrease in the higher MTR above z^* to $\tau_2 < \tau_1$. We refer to this less sharply bent kink as K_2. In the presence of a kink K_j with marginal tax rate τ_0 below z^* and τ_j above z^*, $j \in \{1, 2\}$, the share of individuals bunching at z^*in the frictionless model will be:

$$B_j^* = \int_{z^*}^{z^* + \Delta z_j^*} h_0(\zeta)\, d\zeta \tag{2}$$

For relatively small changes in the tax rate, we can relate the elasticity of earnings with respect to the net-of-tax rate to the earnings change $\triangle z_j^*$ for the individual with the highest *ex ante* earnings who bunches *ex post*:

$$\varepsilon = \frac{\triangle z_j^* / z^*}{d\tau_j / (1 - \tau_0)} \tag{3}$$

where $d\tau_j = \tau_j - \tau_0$ and ε is the elasticity of pre-tax earnings with respect to the net-of-tax rate, $\varepsilon \equiv -(\partial z / z) / (\partial \tau / (1 - \tau))$.

6.1 Fixed Cost of Adjustment

We now extend the model to include a fixed cost of adjusting earnings. Following recent public finance literature on bunching including Saez (2010) and Kleven and Waseem (2013), our model is stylized to illustrate the relevant forces as transparently as possible. We assume

that to change earnings from an initial level, individuals must pay a fixed utility cost of ϕ. This could represent the information costs associated with navigating a new tax regime if, for example, individuals only make the effort to understand their earnings incentives when the utility gains from doing so are sufficiently large (*e.g.* Simon, 1955; Chetty *et al.,* 2007; Hoopes, Reck, and Slemrod, 2013). Alternatively, this cost may represent frictions such as the cost of negotiating a new contract with an employer or the time and financial cost of job search, assuming that these costs do not depend on the size of the desired earnings change. All of these factors may play a role even when individuals consider moving from one positive earnings level to another.

Our model of fixed costs relates to labor economics literature on constraints on hours worked, as well as public finance literature that explores frictions in earnings. One common feature of models of earnings frictions in labor economics (*e.g.* Cogan, 1981; Altonji and Paxson, 1990, 1992; Dickens and Lundberg, 1993) and public finance (*e.g.* Chetty *et al.*, 2011; Chetty, 2012) is that the decision-making setting is generally static. We begin by adopting this modeling convention.

There is an extensive literature on fixed costs and adjustment in other fields, including the "s-S" literature (see literature reviews in Dixit and Pindyck, 2004; Leahy, 2008; and Stokey, 2008). In "s-S" models, agents adjust behavior when the value of a state variable falls outside a "region of inactivity" around its "target" level, within a dynamic optimization problem. Our model shares a similar process, but unlike existing literature, we develop our model in the context of kinked budget sets and the determination of earnings.

6.2 Bunching in a Single Cross-Section with Adjustment Costs

Figure 8, Panel A illustrates how a fixed adjustment cost attenuates the level of bunching, relative to equation (2), and obscures the estimation of ε in a single cross-section that is possible in the Saez (2010) model. The figure shows the budget set before and after the kink, K_1, is introduced, as well as indifference curves that pass through key earnings levels. Consider the individual at point 0, who initially earns $\underline{z}_1$ along the linear budget constraint with tax rate τ_0. This individual faces a higher marginal tax rate after the kink is introduced, which increases the marginal tax rate to τ_1 above earnings level z^*. Because she

faces an adjustment cost, she may decide to keep her earnings at $\underline{z}_1$ and locate at point 1. Alternatively, with a sufficiently low adjustment cost, she would like to incur the adjustment cost and reduce her earnings to z^*, marked by point 2.

We assume that the benefit of relocating to the kink is increasing in distance from the kink for initial earnings in the range $[z^*, z^* + \Delta z_1^*]$. In general, this requires that the size of the optimal adjustment in earnings increases in a at a rate faster than the decrease in the marginal utility of consumption.[18] This assumption is true, for example, if utility is quasilinear, which is assumed in related recent public finance literature (*e.g.* Saez, 2010; Chetty *et al.*, 2011; Kleven, Landais, Saez, and Schultz, 2012; and Kleven and Waseem, 2013).

These assumptions imply that above a threshold level of initial earnings, $\underline{z}_1$, individuals adjust their earnings to the kink, and below this threshold individuals remain inert. In Figure 8, this individual is the marginal buncher who is indifferent between staying at the initial level of earnings $\underline{z}_1$ (point 1) and moving to the kink earnings level z^* (point 2) by paying the adjustment cost ϕ.

In Panel B of Figure 8, we show that the level of bunching is attenuated due to the adjustment cost. Panel B plots the counterfactual density of earnings, *i.e.* under a linear tax τ_0. Only individuals with initial earnings in the range $[\underline{z}_1, z^* + \Delta z_1]$ bunch at the kink K_1 (areas ii, iii, iv, and v)—whereas in the absence of an adjustment cost, individuals with initial earnings in the range $[z^*, z^* + \Delta z_1^*]$ bunch (areas i, ii, iii, iv, and v). The amount of bunching is given by the integral of the initial earnings density, $h_0(\cdot)$, over the range $[\underline{z}_1, z^* + \Delta z_1^*]$:

$$B_1(\boldsymbol{\tau}_1, z^*; \varepsilon, \phi) = \int_{\underline{z}_1}^{z^* + \Delta z_1^*} h_0(\zeta)\, d\zeta, \tag{4}$$

where $\boldsymbol{\tau}_1 = (\tau_0, \tau_1)$ measures the tax rates below and above z^*. The lower limit of the

[18] To see this, note that the utility gain from reoptimizing is $u((1-\tau_1)z_1 + R_1, z_1; a) - u((1-\tau_1)z_0 + R_1, z_0; a) \approx u_c \cdot (1-\tau_1)[z_1 - z_0] + u_z \cdot [z_1 - z_0] = u_c \cdot (\tau_1 - \tau_0)[z_0 - z_1]$, where in the first expression, we have used a first-order approximation for utility at $((1-\tau_0)z_0 + R_0, z_0)$, and in the second expression we have used the first order condtion $u_z = -u_c(1-\tau_0)$. The gain in utility is approximately equal to an expression that depends on the marginal utility of consumption, the change in tax rates, and the size of the earnings adjustment. The first term, u_c, is decreasing as a (and therefore initial earnings z_0) increases. Thus, in order for the gain in utility to be increasing in a, we need the size of earnings adjustment $[z_0 - z_1]$ to increase at a rate that dominates.

integral, $\underline{z}_1$, is implicitly defined by the indifference condition shown in Figure 8, Panel A:

$$\phi = u\left((1-\tau_1)z^* + R_1, z^*; \underline{a}_1\right) - u\left((1-\tau_1)\underline{z}_1 + R_1, \underline{z}_1; \underline{a}_1\right) \tag{5}$$

where R_1 is virtual income and $\underline{a}_1$ is the "ability" level of this marginal buncher.[19]

Bunching therefore depends on the preference parameters ε and ϕ, the tax rates below and above the kink, $\boldsymbol{\tau}_1 = (\tau_0, \tau_1)$, and the density $h_0(\cdot)$ near the exempt amount z^*. With only one kink and without further assumptions, we cannot estimate both ε and ϕ, as the level of bunching depends on both parameters.

6.3 Estimation Using Variation in Kink Size

We can estimate elasticities and adjustment costs when we observe bunching at a kink both before and after a change in $d\tau$, as we observe in our empirical applications. Suppose we observe a population that moves from facing a more pronounced kink K_1, with a marginal tax rate τ_1 above z^*, to facing a less pronounced kink K_2, with a marginal tax rate of $\tau_2 < \tau_1$ above z^*. To make progress, we assume that a is fixed over time from K_1 to K_2. Some individuals will remain bunching at the kink, even though they would prefer to move away from the kink in the absence of an adjustment cost, because the gain from de-bunching is not large enough to outweigh the adjustment cost. The fixed adjustment cost therefore attenuates the reduction in bunching, relative to a frictionless case.[20]

Attenuation in the change in bunching is driven by individuals in area iv of Panel B in Figure 8. Under a frictionless model, individuals in this range do not bunch under the smaller kink K_2. To see this, note that their counterfactual earnings are greater than $z^* + \triangle z_2^*$, *i.e.* the highest level of initial earnings among bunchers at K_2 when there are no frictions. However, when moving from K_1 to K_2 in the presence of frictions, those in area iv continue to bunch, as shown in Panel C of Figure 8. At point 0, we show an individual's initial earnings $\bar{z}_0 \in [z^*, z^* + \Delta z_1^*]$ under a constant marginal tax rate of τ_0. We now introduce the

[19] The threshold level of earnings $\underline{z}_1$ is an increasing function of ϕ. If adjustment costs are large enough, we may have $\underline{z}_1 > z^* + \triangle z_1^*$, in which case frictions eliminate bunching entirely. Since we observe bunching in our empirical setting, we ignore this case.

[20] If $d\tau_2 > d\tau_1$ instead – *i.e.* the kink becomes larger – then additional individuals will be induced to bunch, but the change in bunching will in general still be attenuated (due to the adjustment cost). This is governed by an analogous set of formulas to the case $d\tau_2 < d\tau_1$ that we explore.

first kink, K_1. The individual responds by bunching at z^* (point 1), since $\bar{z}_0 > \underline{z}_1$. Next, we transition to the less pronounced kink K_2. Since $\bar{z}_0 > z^* + \triangle z_2^*$, this individual would have chosen earnings $\bar{z}_2 > z^*$ (point 2) under K_2 in a frictionless setting. However, to move to point 2, this individual must pay a fixed cost of ϕ. We have drawn this individual as the marginal buncher who is indifferent between staying at z^* and moving to $\bar{z}_2$. Under similar logic, all individuals with initial earnings in the range $[z^* + \Delta z_2^*, \bar{z}_0]$ will remain at the kink.

Thus, bunching under K_2 is:

$$\tilde{B}_2(\tilde{\boldsymbol{\tau}}_2, z^*; \varepsilon, \phi) = \int_{\underline{z}_1}^{\bar{z}_0} h_0(\zeta)\, d\zeta, \tag{6}$$

where $\tilde{\boldsymbol{\tau}}_2 = (\tau_0, \tau_1, \tau_2)$ measures the tax rate below z^*, the initial tax rate above z^*, and the final tax rate above z^*, respectively, and the "$\sim$" indicates that the budget set with K_2 was preceded by a larger kink K_1. The critical earnings levels for the marginal buncher, $\bar{z}_0$ and $\bar{z}_2$, are implicitly defined by the following three conditions:[21]

$$\begin{aligned} -\frac{u_z(c_2, \bar{z}_2; \bar{a}_2)}{u_c(c_2, \bar{z}_2; \bar{a}_2)} &= (1-\tau_2) \\ u((1-\tau_2)\bar{z}_2 + R_2, \bar{z}_2; \bar{a}_2) - u((1-\tau_2)z^* + R_2, z^*; \bar{a}_2) &= \phi \\ -\frac{u_z(c_0, \bar{z}_0; \bar{a}_2)}{u_c(c_0, \bar{z}_0; \bar{a}_2)} &= (1-\tau_0). \end{aligned} \tag{7}$$

In words, the first line indicates that $\bar{z}_2$ is the optimal, frictionless level of earnings chosen by the top buncher in the presence of K_2, where $\bar{z}_2 > z^*$. The second line requires that when facing K_2, this agent is indifferent between remaining at z^*, or moving to $\bar{z}_2$ and paying the adjustment cost. The third line defines $\bar{z}_0$ as the initial level of earnings that this individual chooses when facing a constant marginal tax rate of τ_0 and no kink. The elasticity of taxable earnings is again related to the potential adjustment of the marginal buncher: $\varepsilon = \frac{\bar{z}_0 - \bar{z}_2}{\bar{z}_2} \frac{(1-\tau_0)}{d\tau_2}$.

The equations in (4), (5), (6) and (7) together pin down four unknowns ($\triangle z_1^*$, $\underline{z}_1$, $\bar{z}_0$ and $\bar{z}_2$), each of which is in turn a function of ε and ϕ. Bunching at each kink is therefore jointly

[21] We additionally require that $\bar{z}_0 \leq z^* + \triangle z_1^*$. When this inequality is binding, none of the bunchers move away from the kink at z^* when the kink is reduced from K_1 to K_2. Since we observe a reduction in bunching in our empirical setting, we ignore this inequality.

determined by ε and ϕ. Ultimately, we draw on two empirical moments in the data, B_1 and $\tilde{B}_2$, to identify our two key parameters, ε and ϕ.

Relative to the frictionless case represented by the Saez (2010) model, the change in bunching from the more pronounced kink K_1 to the less pronounced kink K_2 is now attenuated by the adjustment cost. As noted above, in the Saez (2010) model, bunching decreases by areas iv and v in Figure 8 when moving from K_1 to K_2. When moving sequentially from K_1 to K_2 in the presence of an adjustment cost, areas ii, iii, iv, and v bunch under K_1, whereas areas ii, iii, and iv bunch under K_2. Thus, bunching decreases only by area v, rather than by both areas iv and v as in the frictionless case. We show in Gelber, Jones, and Sacks (2013) that the absolute value of the decrease in bunching from K_1 to K_2 is decreasing in the adjustment cost—$\bar{z}_0$ is increasing in the adjustment cost, and therefore area v is decreasing in the adjustment cost. As in the frictionless case, the amount of bunching at K_1 is still increasing in the elasticity (*ceteris paribus*). We refer to this estimation strategy, using data just before and after a policy change, as the "comparative static method."

The features of the data that help drive our estimates of the elasticity and adjustment cost are intuitive. In the frictionless model of Saez (2010), bunching at a convex kink is approximately proportional to $d\tau$; thus, when $d\tau$ falls in this model, the degree of bunching at the kink falls proportionately. In our model, adjustment costs help to explain deviations from this pattern. As we move from the more sharply bent kink to the less sharply bent kink in our model with adjustment costs, bunching falls by a less-than-proportional amount—consistent with our empirical observation that individuals continue to bunch at the location of a former kink. In the extreme case in which a kink has been eliminated, we can attribute any residual bunching to adjustment costs. Moreover, the absolute value of the change in bunching is decreasing in the adjustment cost.

6.3.1 Accounting for the Claiming Decision

In our model and estimation we abstract from the claiming decision by examining a sample of those who have already claimed OASI (*i.e.* our estimates use a sample of those older than age 65 who had claimed by age 65); our model thus applies more broadly to understanding responses to kink points where the claiming decision is not relevant (as in, for example, most

other tax contexts, including those where Gudgeon and Trenkle (2016) and He, Peng, and Wang (2016) have applied our method). In the context of the AET, this modeling choice follows previous literature such as Friedberg (1998, 2000).

This decision is only a trivial abstraction in our context because nearly everyone has claimed by the ages we study in our main evidence, 69 to 71, implying minimal scope for the claiming decision—or conditioning on those who have claimed by 65, as we do in our primary estimates—to affect our results. By age 69, fully 94 percent of the sample has claimed OASI, and nearly everyone has claimed by 71. Moreover, the bunching we observe at ages after the AET ceases to apply is a tell-tale sign of adjustment frictions, regardless of the fraction claiming at these ages. By ages 66 to 68, which we later examine in the context of the 1990 change in the AET BRR, 92 percent have claimed OASI on average across ages in the time period we study. Moreover, over the period we examine from before to after 1990, the proportion claiming is stable at 92 percent in each year separately from 1988 to 1992, implying that the results should not be materially affected by changes in claiming over time. It is straightforward to show that as the percent of the sample claiming approaches 100, our bunching estimates—and therefore the resulting parameter estimates using our method of simulated moments estimator explained below—will converge to those we have calculated conditional on claiming (results available upon request). Consistent with this, we estimate very similar, and insignificantly different, elasticities and adjustment costs when bunching is estimated from the sample of only those who have claimed, as when the sample includes both claimants and non-claimants (results available upon request).

Moreover, empirically our evidence suggests the claiming decision during our ages of interest does not appear to interact notably with the AET. In particular, we add to previous literature by showing in Appendix Figure B.4 that the hazard of claiming at year $t+1$ is smooth around the exempt amount at year t, indicating no evidence that claimants come disproportionately from close to or far from the kink. Building on these results, Gelber, Jones, Sacks, and Song (2016) are studying claiming responses in greater depth (as studied using a different strategy in Gruber and Orszag, 2003).

6.4 Extensions

Our basic model can be extended in a number of ways, including by incorporating more dynamic elements into the model and by allowing for heterogeneity in our key parameters.[22]

6.4.1 Dynamic Extension of Model

By applying our approach thus far to study adjustment over a given time frame, the resulting parameters should be interpreted as meaning that bunching in this given time frame can be predicted if individuals behaved as if they faced the indicated adjustment cost and elasticity—in the spirit of Friedman (1953), who argued that economic models should predict behavior "as if" individuals followed the model. In practice, we apply our model to study the nature of immediate adjustment to a policy change, so the parameters we estimate pertain to the frictions faced in immediately adjusting to a policy change. This framework may be applied separately in each period to yield "as if" estimates separately in each period, thus yielding adjustment costs and elasticities in each period separately.

However, the comparative static model gives no account of how bunching may evolve over time. Section 5 shows evidence of lagged adjustment up to two years following a change in incentives. To capture such features of the data, we can nest our comparative static model within a framework incorporating more dynamic elements that allow us to account for this subsequent adjustment. We use a Calvo (1983) or "CalvoPlus" framework (*e.g.* Nakamura and Steinsson, 2010), in which there is a positive probability in each period of facing a finite, fixed adjustment cost.

Thus, we will assume that the adjustment cost in any period is drawn from a discrete distribution $\{0, \phi\}$. This generates a gradual response to policy, as agents may adjust only when a sufficiently low value of the fixed cost is drawn.[23] This element of the model is necessarily stylized, for the sake of tractability and consistency with previous literature.

[22] In Gelber, Jones, and Sacks (2013), we discuss a number of extensions of the model, including allowing for frictions to affect the distribution of earnings in the initial period, or allowing for adjustment costs that are linear in the size of the adjustment. We also discuss the possibility of using other moments of the data to inform the estimates. Finally, we discuss how under certain assumptions we can express the elasticity and adjustment cost as functions of observed levels of bunching that can be easily solved in closed form.

[23] We have alternatively modeled intertemporal shocks via a time-varying ability, a. This model will likewise generate delayed response, as agents will only reoptimize when the ability draw generates a preferred level of earnings far enough away from current earnings to justify paying the fixed cost of adjustment. Analogously, agents in our model reoptimize only when a sufficiently low cost of adjustment is drawn.

Such variation over time in the size of the adjustment cost from this discrete distribution could capture, for example, the job search process following a recent policy change: when a job offer arrives exogenously, the adjustment cost reaches zero, but when it is not available, search costs are positive and equal to ϕ. Future work could generalize this or other aspects of the model.

How we model dynamics is also influenced by a key feature observed in the data: the lack of an anticipatory response to policy changes. In Appendix A.2, we solve a completely forward-looking model, which nests the models presented in the main text. In this forward-looking model, the key results are less parsimonious, and the identification of the key parameters is less transparent. In practice, the data drive this unrestricted version of the model to place little to no weight on the future by estimating discount factors of zero or near zero. If agents were to place weight on the future in our forward-looking model, they should begin to bunch in anticipation of facing a kink, and they should begin to de-bunch in anticipation of the disappearance of a kink—neither of which we have observed in the data, as shown for example in Figure 2 or Figure 3. Note, however, that our confidence intervals do not allow us to rule out forward-looking behavior entirely; indeed, the confidence intervals on bunching before age 62—or on a test that bunching changes in anticipation of the removal of the AET at age 70—are rather large and therefore do not rule out substantial anticipatory behavior. Nonetheless, the evidence is broadly consistent with agents who are not very forward-looking.

Meanwhile, we observe a degree of delayed response to policy changes. We can capture both of these features of the data by assuming that it is stochastic whether an agent faces the cost of adjustment, but agents are not forward looking. We therefore focus on the case without forward-looking behavior in the main text, as it is sufficient to explain the patterns in the data. The lack of forward-looking behavior can be rationalized, for example, if agents are myopic or only learn about the AET through experience with it.

We acknowledge that if individuals are uncertain about their future earnings, they may avoid anticipatory adjustment to preserve the option value of making the choice whether to respond once the uncertainty is resolved. To fit the data showing no evidence of adjustment in advance of anticipated policy changes, such a model with uncertainty and adjustment

costs would necessarily have to predict little adjustment in advance. Thus, the observable implications of the two sets of assumptions are similar in our context; in this light, our model without anticipatory adjustment can be interpreted as meaning that bunching can be predicted if individuals behaved as if they faced the estimated parameters, in the spirit of Friedman (1953). Moreover, it is worth noting that when we estimate our more static model for those with high and low prior earnings volatility, we estimate insignificantly different results in the two groups. Although other papers in this line have not modeled uncertainty (*e.g.* Saez 2010; Chetty *et al.*, 2011, 2012; Kleven and Waseem, 2013; Marx, 2015), an interesting extension for future work of our basic starting point would be to formally estimate a dynamic model with budget set kinks in which individuals may respond in anticipation of future policy and also face uncertainty about future earnings opportunities.[24]

Formally, our main dynamic model (without forward-looking behavior) extends the notation from above as follows. As before, we assume that agents begin with their optimal frictionless level of earnings in period 0. Flow utility in each period is $v\left(c_{a,t}, z_{a,t}; a, z_{a,t-1}\right) = u\left(c_{a,t}, z_{a,t}; a\right) - \tilde{\phi}_t \cdot \mathbf{1}\left(z_{a,t} \neq z_{a,t-1}\right)$, where $\mathbf{1}\left(\cdot\right)$ is the indicator function. Individuals are again indexed by a time-invariant heterogeneity parameter, a, which captures ability. In each period, an individual draws a cost of adjustment, $\tilde{\phi}_t$, from a discrete distribution, which equals ϕ with probability π_{t-t^*} and equals 0 with probability $1 - \pi_{t-t^*}$. To capture the observed features of the data, in which the probability of adjusting (conditional on initially locating at the kink) appears to vary over time, we allow the probability π_{t-t^*} to be potentially a function of the time lapsed since the most recent policy change, *i.e.* $t - t^*$.

Individuals make decisions over a finite horizon. In period 0, the individuals face a linear tax schedule, $T_0\left(z\right) = \tau_0 z$, with marginal tax rate τ_0. In period 1, a kink, K_1, is introduced at the earnings level z^*. This tax schedule is implemented for $\mathcal{T}_1$ periods, after which the tax schedule features a smaller kink, K_2, at the earnings level z^*. As before, the kink K_j, $j \in \{1, 2\}$, features a top marginal tax rate of τ_j for earnings above z^*. For simplicity, we abstract from income effects, to focus on the dynamics created by the presence of adjustment costs. In particular, $u\left(c, z; a\right) = c - \frac{a}{1+1/\varepsilon}\left(\frac{z}{a}\right)^{1+1/\varepsilon}$. In each period, individuals draw a cost of

[24] Such a model could build on Werquin (2015), who models uncertainty but only features a smooth tax schedule, and therefore does not accommodate kinked budget sets.

adjustment, $\tilde{\phi}_t$, and then maximize flow utility, $v\left(c_{a,t}, z_{a,t}; a, z_{a,t-1}\right)$ subject to a per-period budget constraint $z_{a,t} - T_j\left(z_{a,t}\right) - c_{a,t} \geq m$, where $j = \mathbf{1}\left(t \geq 1\right) + \mathbf{1}\left(t > \mathcal{T}_1\right)$, and m represents a borrowing constraint.[25]

These assumptions generate a simple decision rule for agents in each period. Let $\tilde{z}_{a,t}$ be the optimal frictionless level of earnings for an individual with ability a in period t, which is a function of the tax schedule in that period. An agent will choose this level of earnings provided that the utility gain of moving from $z_{a,t-1}$ to $\tilde{z}_{a,t}$ exceeds the currently-drawn cost of adjustment, $\tilde{\phi}_t$. Otherwise, the agent remains at $z_{a,t-1}$. The agent only considers current payoffs because we have abstracted from forward-looking behavior.

The model yields two types of transitions following a policy change. First, there are agents who adjust immediately following a policy change, because the gain in utility exceeds the highest possible adjustment cost, ϕ. These are the same agents who adjust in our model in Section 6.3. Second, there are agents who only adjust once a zero cost of adjustment is drawn, which happens in each period with probability $1 - \pi_{t-t^*}$.

We can now generalize our expressions for bunching under K_1 and K_2, (4) and (6). Denote B_1^t as bunching at K_1 in period $t \in [1, \mathcal{T}_1]$. We have the following dynamic version of (4):

$$\begin{aligned} B_1^t &= \int_{\underline{z}_1}^{z^*+\Delta z_1^*} h_0\left(\zeta\right) d\zeta + \left(1 - \Pi_{j=1}^t \pi_j\right) \int_{z^*}^{\underline{z}_1} h_0\left(\zeta\right) d\zeta \\ &= \Pi_{j=1}^t \pi_j \cdot B_1 + \left(1 - \Pi_{j=1}^t \pi_j\right) B_1^* \end{aligned} \tag{8}$$

where $h_0\left(\cdot\right)$ is again the density of earnings under a linear tax τ_0, $\underline{z}_1$ is implicitly defined in equation (5), B_1 is defined in (4) and B_1^* is the frictionless level of bunching defined in (2) when $j = 1$. We see on the first line of (8) that bunching in period t at K_1 is composed of two parts. First, there are individuals who immediately adjust in period 1, *i.e.* areas $ii - v$ in Figure 8, Panel B. Second, there are individuals in area i of Figure 8, who only adjust if they have drawn a zero cost of adjustment. The probability that this occurs by period t is $1 - \Pi_{j=1}^t \pi_j$. Finally, we see in the second line of (8) that as t grows, $\Pi_{j=1}^t \pi_j$ converges to zero, and bunching converges to the frictionless level of bunching B_1^*, *i.e.* areas $i - v$ in

[25] Note, the quasilinearity assumption implies that the borrowing constraint does not directly affect the earnings decision. However, when agents are not forward looking, the borrowing constraint is necessary to rule out infinite borrowing.

Figure 8, Panel B.

We can similarly derive an expression for B_2^t, bunching at K_2 in period $t > \mathcal{T}_1$:

$$\begin{aligned} B_2^t &= \int_{\underline{z}_1}^{z^*+\Delta z_2^*} h_0(\zeta)\, d\zeta + \Pi_{j=1}^{t-\mathcal{T}_1}\pi_j \cdot \int_{z^*+\Delta z_2^*}^{\bar{z}_0} h_0(\zeta)\, d\zeta \\ &\quad + \left(1 - \Pi_{j=1}^{t-\mathcal{T}_1}\pi_j \cdot \Pi_{j=1}^{\mathcal{T}_1}\pi_j\right) \int_{z^*}^{\underline{z}_1} h_0(\zeta)\, d\zeta \\ &= \Pi_{j=1}^{t-\mathcal{T}_1}\pi_j \cdot \left[\tilde{B}_2 + \left(1 - \Pi_{j=1}^{\mathcal{T}_1}\pi_j\right)\left[B_1^* - B_1\right]\right] + \left(1 - \Pi_{j=1}^{t-\mathcal{T}_1}\pi_j\right) B_2^* \end{aligned} \tag{9}$$

where $\bar{z}_0$ is implicitly defined in (7), $\tilde{B}_2$ is defined in (6) and B_2^* is the frictionless level of bunching defined in (2) when $j = 2$. In the first two lines, bunching in period t at K_2 consists of three components. First, there are individuals who immediately bunched in period 1, and remain bunching at the smaller kink, *i.e.* areas $ii - iii$ in Figure 8, Panel B. Second, there are "excess bunchers" who immediately bunched in period 1, and now "de-bunch" when a zero cost of adjustment is drawn, *i.e.* individuals in area iv of the same figure. The probability of not having drawn a zero cost of adjustment between periods $\mathcal{T}_1 + 1$ and t is $\Pi_{j=1}^{t-\mathcal{T}_1}\pi_j$. Finally, there are individuals who would like to bunch under both K_1 and K_2, but will only do so once a zero cost of adjustment is drawn, *i.e.* those in area i of Figure 8, Panel B. A fraction of these agents, $\left(1 - \Pi_{j=1}^{\mathcal{T}_1}\pi_j\right)$, have drawn a zero cost of adjustment by period $\mathcal{T}_1$, and of the remaining $\Pi_{j=1}^{\mathcal{T}_1}\pi_j$, the probability of drawing a zero cost from period $\mathcal{T}_1 + 1$ to t is $\left(1 - \Pi_{j=1}^{t-\mathcal{T}_1}\pi_j\right)$, yielding a total share of $\left(1 - \Pi_{j=1}^{\mathcal{T}_1}\pi_j\right) + \Pi_{j=1}^{\mathcal{T}_1}\pi_j \cdot \left(1 - \Pi_{j=1}^{t-\mathcal{T}_1}\pi_j\right) = \left(1 - \Pi_{j=1}^{t-\mathcal{T}_1}\pi_j \cdot \Pi_{j=1}^{\mathcal{T}_1}\pi_j\right)$. On the third line, we once again see that as the time between period t and $\mathcal{T}_1$ grows, $\Pi_{j=1}^{t-\mathcal{T}_1}\pi_j$ converges to zero, and the level of bunching converges to the frictionless amount, B_2^*, *i.e.* areas $i - iii$ in Figure 8, Panel B.

This dynamic version of the model nests the comparative static model from Sections 6.2-6.3. When $\pi_j = 1, \forall j$, (8) shows that $B_1^t = B_1$, and (9) shows that $B_2^t = \tilde{B}_2$. Furthermore, when $\pi_j = 0, \forall j$, or $\phi = 0$, the model returns the predictions from the frictionless, Saez (2010) model. This dynamic model also has implications for how we wish to interpret the comparative static model. The amount of bunching after a kink is removed (and therefore the estimated adjustment cost) should depend on the amount of time the kink has been in place, as in (9). As the kink has been in place for longer (before it is removed), a different

set of individuals (whose adjustment takes longer on average) bunches at the kink. Thus, removal of the kink should be associated with greater inertia at the kink when the kink has been in place for a longer time than when it has been in place for a shorter time. In this light, we interpret the results of the static model in a "reduced form" sense, namely as representing the elasticity and adjustment cost that can predict behavior within a given time frame for the *particular* individuals initially bunching at the kink. The dynamic model helps shed further light on how these bunching amounts should vary over time, and therefore helps us interpret the results of the static estimation.

Relative to the "comparative static" model in Section 6.3, the dynamic model has both strengths and weaknesses. The comparative static model transparently illustrates the basic forces determining the elasticity and adjustment cost. The estimation of the more dynamic model requires more moments from the data to estimate more parameters. We assume that ability is fixed throughout the window of estimation, which may be more plausible in the case of the static model in Section 6.3—when we only use two cross-sections from adjacent time periods—than when we use a dynamic model and study a longer time frame. However, this affords us tractability in the dynamic model, while allowing us to account for the time pattern of bunching.

6.4.2 Heterogeneity in Elasticities and Costs of Adjustment

The previous analysis assumed homogeneous elasticities and adjustment costs, but we can extend the model to accommodate heterogeneity. Suppose $(\varepsilon_i, \phi_i, a_i)$ is jointly distributed according to a smooth CDF, which translates into a smooth, joint distribution of elasticities, fixed adjustment costs, and earnings in the presence of a linear tax, $h_0^*(\varepsilon, \phi, z_0)$. In Appendix A.1 we derive generalized formulae for bunching that allow us to interpret our estimates as the average behavioral response and attenuation due to adjustment costs among the set of bunchers.[26] Appendix A.1.2 further discusses how the dynamic model can be interpreted in the presence of heterogeneity in these parameters and the vector $\boldsymbol{\pi}_i$.

Our estimates of elasticities and adjustment costs, and our earlier descriptive evidence documenting the speed of adjustment, are specific to the population that is observed bunch-

[26] We are grateful to Henrik Kleven for suggesting the approach that led to this derivation.

ing at the kinks. At the same time, note that for any value of ϕ_i, there exists a value of ε_i that generates positive bunching. Thus, while our estimates are local to the observed set of bunchers, they need not be confined to a subpopulation with small values of ϕ_i—for example, if ε_i and ϕ_i are positively correlated. Nevertheless, there may be a set of individuals for whom ε_i is small enough relative to ϕ_i to preclude bunching under either K_1 or K_2, and who therefore do not contribute to our parameter estimates.

It is important to note that while this may be a limitation of our particular policy setting, it is not a general methodological limitation in the sense that with sufficiently large variation in tax rates it may be possible to estimate population average parameters. Our estimation procedure relies on estimating bunching at more than one kink; over all such kinks, the limits of the integrals used to calculate bunching could in principle jointly cover much of the earnings distribution. Loosely speaking, the greater the variation in tax rates, the more of the population we will observe who bunch at kinks and therefore contribute to our estimates. In that light, our policy variation is useful because it varies over a large range of BRRs (from 50 percent to 33.33 percent to zero percent). Moreover, it is perhaps reassuring that we will find similar elasticity and adjustment cost estimates when we examine a larger change in the BRR (from 33.33 percent to zero percent) as when we examine a smaller change (from 50 percent to 33.33 percent). Extrapolating our estimates from bunchers to non-bunchers would require assumptions on the joint distribution of ε and ϕ.

6.5 Econometric Estimation of the Model

6.5.1 Comparative Static Model

We begin by describing our econometric estimation procedure under our basic comparative static model of Sections 6.2 and 6.3. Let $B = (B_1, B_2, \ldots, B_L)$ be a vector of (estimated) bunching amounts, using the method described in Section 3. Let $\boldsymbol{\tau} = (\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_L)$ be the tax schedule at each kink. The triplet $\boldsymbol{\tau}_l = (\tau_{0,l}, \tau_{1,l}, \tau_{2,l})$ denotes the tax rate below the kink ($\tau_{0,l}$), above the kink ($\tau_{1,l}$), and the *ex post* marginal tax rate above the kink after it has been reduced ($\tau_{2,l}$), as in Section 6.3. Let $\mathbf{z}^* = (z_1^*, \ldots, z_L^*)$ be the earnings levels associated with each kink. In principle, it would be possible to estimate bunching separately for each age group at a given kink. In practice and for simplicity, we pool across a constant set of

ages to estimate bunching at a given kink—for example, when examining the 1990 policy change we examine 66-68 year-olds both before and after the change—so it is not necessary to index the bunching amounts by age as well.[27]

In our baseline, we use a non-parametric density for the counterfactual earnings distribution, H_0. Once H_0 is known, we use (4) and (6) to obtain predicted bunching from the model. To recover H_0 non-parametrically we take the empirical earnings distribution for 72 year-olds in \$800 bins as the counterfactual distribution. 72 year-olds' earnings density represents a reasonable counterfactual because they no longer face the AET, no longer show bunching, and are close in age to those aged 70 or 71. Letting z_i index the bins, our estimate of the distribution is $\hat{H}_0(z_i) = \sum_{j \leq i} Pr(z \in z_j)$. This function is only defined at the midpoints of the bins, so we use linear interpolation for other values of z. In a robustness check, we instead assume that the earnings distribution over the range $[z^*, z^* + \Delta z]$ is uniform, a common assumption in the literature (*e.g.* Chetty *et al.,* 2011, Kleven and Waseem, 2013). Using the nonparametrically-estimated distribution of earnings from age 72 is helpful because it does not entail distributional assumptions, but relative to assuming a uniform distribution, using the age-72 distribution comes at the cost of using a different age (*i.e.* 72) to generate the earnings distribution.

To estimate (ε, ϕ), we seek the values of the parameters that make predicted bunching $\hat{B}$ and actual (estimated) bunching B as close as possible on average. Letting $\hat{B}(\varepsilon, \phi) \equiv (\hat{B}(\boldsymbol{\tau}_1, z_1^*, \varepsilon, \phi), \ldots, \hat{B}(\boldsymbol{\tau}_L, z_L^*, \varepsilon, \phi))$, our estimator is:

$$\left(\hat{\varepsilon}, \hat{\phi}\right) = \text{argmin}_{(\varepsilon,\phi)} \left(\hat{B}(\varepsilon, \phi) - B\right)' W \left(\hat{B}(\varepsilon, \phi) - B\right), \tag{10}$$

where W is a $K \times K$ identity matrix. This estimation procedure runs parallel to our theoretical model, as the bunching amounts $\hat{B}$ are those predicted by the theory (and the estimated counterparts B are found using the procedure outlined in Section 3).[28] When we pool data across multiple time periods, we assume that ε and ϕ are constant across these time periods.

[27] Analogously, when we examine bunching at each age around 70 when the AET is eliminated, we pool across calendar years (namely 1990-1999) to estimate bunching, so that we do not also have to index the bunching amounts by calendar year. We find comparable results when we estimate bunching separately at each age and year.

[28] Without loss of generality, we use normalized bunching, $\hat{b} = \delta \hat{B} / h_0(z^*)$, so that the moments are identical to what is reported elsewhere in the text.

We obtain our estimates by minimizing (10) numerically. Solving this problem requires evaluating $\hat{B}$ at each trial guess of (ε, ϕ).[29] Our estimator assumes a quasilinear utility function, $u(c, z; a) = c - \frac{a}{1+1/\varepsilon}\left(\frac{z}{a}\right)^{1+1/\varepsilon}$, following Saez (2010), Chetty *et al.* (2011) and Kleven and Waseem (2013). Note that because we have assumed quasilinearity, $\Delta z_{1,l} = z_l^* \left(\left(\frac{1-\tau_{1,l}}{1-\tau_{0,l}}\right)^{\varepsilon} - 1\right)$ and $a = z(\tau)/(1-\tau)^{\varepsilon}$, where $z(\tau)$ are the optimal, interior earnings under a linear tax of τ. Typically there is no closed form solution for $\underline{z}_{1,l}$ or $\bar{z}_{0,l}$. Instead, given ε and ϕ, we find $\underline{z}_{1,l}$ and $\bar{z}_{0,l}$ numerically as the solution to the relevant indifference conditions in (5) and (7). For example, $\underline{z}_{1,l}$ is defined implicitly by:

$$\underbrace{u((1-\tau_{1,l})z_l^* + R_{1,l}, z_l^*; \underline{z}_{1,l}/(1-\tau_{0,l})^{\varepsilon})}_{\text{utility from adjusting to kink}} - \underbrace{u((1-\tau_{1,l})\underline{z}_{1,l} + R_{1,l}, \underline{z}_{1,l}; \underline{z}_{1,l}/(1-\tau_{0,l})^{\varepsilon})}_{\text{utility from not adjusting}} = \phi, \tag{11}$$

This equation is continuously differentiable and has a unique solution for $\underline{z}_{1,l}$.[30]

6.5.2 Dynamic Model

Our estimation method is easily amended to accommodate the dynamic extension of our model in Section 6.4.1. As in (8) and (9), the bunching expressions in the dynamic model are weighted sums of B_1 and $\tilde{B}_2$, which are calculated as in Section 6.5.1, and two measures of frictionless bunching, B_1^* and B_2^*. Frictionless bunching under either kink can be calculated conditional on H_0 and ε using (2).

We must also estimate the probability of drawing a positive fixed cost as a function of the time since the last policy shock, π_{t-t^*}.[31] For given values of ε, ϕ, and the vector $\boldsymbol{\pi}$ of π_{t-t^*}'s, we can evaluate (8) and (9). Our vector of predicted bunching, $\hat{B}$, will now be a function of these additional parameters, as well as the relevant time indices: $\hat{B}(\varepsilon, \phi, \boldsymbol{\pi}) \equiv (\hat{B}(\boldsymbol{\tau}_1, z_1^*, t_1, \mathcal{T}_{1,l}, \varepsilon, \phi, \boldsymbol{\pi}), \ldots, \hat{B}(\boldsymbol{\tau}_L, z_L^*, t_L, \mathcal{T}_{1,L}, \varepsilon, \phi, \boldsymbol{\pi}))$, where t_l is the time elapsed since the first kink, $K_{1,l}$, was introduced, and $\mathcal{T}_{1,l}$ is the length of time before the second kink, $K_{2,l}$, is introduced. Once again we use the minimum distance estimator (10).

Equations (8) and (9) illustrate how we estimate the elasticity and adjustment cost

[29] In solving (10), we impose that $\phi \geq 0$. When $\phi < 0$, every individual adjusts her earnings by at least some arbitrarily small amount, regardless of the size of ϕ. This implies that ϕ is not identified if it is less than zero.

[30] Note that some combinations of $\boldsymbol{\tau}_l$, z_l^*, ε, and ϕ imply $\underline{z}_{1,l} > z_l^* + \triangle z_{1,l}$. In this case, the lowest-earning adjuster does not adjust to the kink. Whenever this happens, we set $\hat{B}_l = 0$.

[31] We have also tried using a flexible, logistic functional form, $\pi_j = \exp(\alpha + \beta \cdot j)/(1 + \exp(\alpha + \beta \cdot j))$, and we found comparable results (available upon request).

in this richer setting. We require as many observations of bunching as the parameters, $(\varepsilon, \phi, \pi_1, ..., \pi_J)$, and these moments must span a change in $d\tau$.[32] Suppose we observe the pattern of bunching over time around two or more different policy changes. Loosely speaking, the $\boldsymbol{\pi}$'s are estimated relative to one another from the time pattern of bunching over time: a delay in adjustment in a given period will generally correspond to a higher probability of facing the adjustment cost (all else equal). Note that the relationship is linear; the degree of "inertia" in bunching in (for example) period 1 increases linearly in π_1. Meanwhile, a higher ϕ implies a larger amount of inertia in *all* periods until bunching has fully dissipated (in a way that depends on the earnings distribution, the elasticity, and the size of the tax change). Finally, a higher ε will correspond to a larger amount of bunching once bunching has had time to adjust fully to the policy changes. Intuitively, these features of the data help us to identify the parameters using our dynamic model.

6.6 Inference

We again estimate bootstrapped standard errors to perform inference about our parameters. For example, in our comparative static model, the estimated vector of parameters $(\hat{\varepsilon}, \hat{\phi})$ is a function of the estimated amount of bunching. We use the bootstrap procedure of Chetty *et al.* (2011) to obtain 200 bootstrap samples of B. For each bootstrap sample, we compute $\hat{\varepsilon}$ and $\hat{\phi}$ as the solution to (10). We determine whether an estimate of ϕ is significantly different from zero by assessing how frequently the constraint $\phi \geq 0$ binds in our estimation. This share is doubled in order to construct p-values from a two-sided test.

6.7 Identification

We have already discussed the features of the data that intuitively help us identify the parameters in the comparative static and dynamic models. We have also shown how conditions can be derived in which a model with heterogeneous parameters reduces to the equations governing the homogeneous case.

It is also possible to demonstrate identification more formally. In Appendix A.3 we give formal conditions for identification in both the comparative static and dynamic cases, and

[32] The number of moments is not itself sufficient. We also require non-trivial variation in bunching before and after the tax change in order to point identify ϕ. As in footnote 21, this requires $\bar{z}_0 < z^* + \triangle z_1^*$.

we show that these conditions hold in our data in both cases. These conditions essentially require that we have sufficient tax variation relative to the adjustment cost. Intuitively, without sufficient tax variation, there will not be any bunching or de-bunching, and so we cannot identify the parameters—a case that is not relevant in our data, as we observe both bunching and de-bunching.

7 Estimates of Elasticity and Adjustment Cost

7.1 Estimates using the Comparative Static Method

To estimate ε and ϕ using our "comparative static" method, we first examine the reduction in the rate in 1990 and next turn to the elimination of the AET at ages 70 and older. We use the 1990 change as a baseline because this allows us to compare our method to the Saez (2010) method, whereas we cannot apply the Saez method at age 70 or later because the BRR is zero above the exempt amount.

We begin with graphical depictions of the patterns driving the parameter estimates for the 1990 change. We follow a group of 66-68 year-olds, so that we can examine an age group that moved over time from being affected by the 50 percent tax rate before the policy change to the 33.33 percent tax rate after the policy change. Although 69-year-olds are also subject to the AET, an individual who is 69 in 1989 would have turned 70 by 1990 and therefore would not have been affected by the 33.33 percent tax rate in 1990—preventing us from examining those age 69 if we wish to examine a constant age group. Examining a constant age group (*i.e.* those aged 66-68 in each calendar year) is crucial because different age groups have persistently different amounts of bunching, which we do not wish to confound with the effect of the policy change. Figure 9 shows bunching among 66-68 year-olds, for whom the BRR fell from 50 percent to 33.33 percent in 1990. Bunching fell slightly from 1989 to 1990 but fell more subsequent to 1990.

Estimating these parameters requires estimates of the implicit marginal tax rate: both the "baseline" marginal tax rate, τ_0—the rate that individuals near the AET threshold face in the absence of the AET due to federal and state taxes—and the implicit marginal tax rate associated with the AET. We begin by using a marginal tax rate incorporating the AET BRR

as well as the average federal and state income and payroll marginal tax rates, calculated using the TAXSIM calculator of the National Bureau of Economic Research (Feenberg and Coutts, 1993) and information on individuals within $2,000 of the kink in the Statistics of Income data in the years we examine.

Table 2 presents estimates of our static model, examining 66-68 year-olds in 1989 and 1990. We estimate an elasticity of 0.35 in Column (1) of Table 2 and an adjustment cost of $278 in Column (2), both significantly different from zero ($p < 0.01$). This specification examines data in 1989 and 1990; thus, our estimated adjustment cost represents the cost of adjusting earnings in the first year after the policy change. We interpret our estimates as meaning that when considering a given time frame (in this particular case, from 1989 to 1990), bunching amounts can be predicted if individuals behaved as if they had the estimated elasticity and adjustment cost (in this particular case, 0.35 and $278, respectively).

When we constrain the adjustment cost to zero using 1990 data in Column (3), as most previous literature has implicitly done, we estimate a substantially larger elasticity of 0.58.[33] Consistent with our discussion above, it makes sense that the estimated elasticity is higher when we do not allow for adjustment costs than when we do, as adjustment costs keep individuals bunching at the kink even though tax rates have fallen. The difference in the constrained and unconstrained estimates of the elasticity is substantial (66 percent higher in the constrained case) and statistically significant ($p < 0.01$).

Other specifications in Table 2 show similar results. Adjusting the marginal tax rate to take account of benefit enhancement raises the estimated elasticity but yields similar qualitative patterns across the constrained and unconstrained estimates. The next rows show that our estimates are similar under other specifications: excluding FICA taxes from the baseline tax rate; using a locally uniform density; other bandwidths; and other years of analysis.[34] The point estimates in Appendix Table B.3 show that across groups, elasticities tend to be similar, but women have higher adjustment costs than men, those with low prior lifetime real earnings have higher adjustment costs than those with high prior earnings, and

[33] Friedberg (2000) finds uncompensated elasticity estimates of 0.22 and 0.32 in different samples. However, differences in the estimation strategies imply that these results are not directly comparable to ours.

[34] In Gelber, Jones, and Sacks (2013) we find similar estimates when we apply our method to the 1990 policy change but assume that bunching in 1989 is not attenuated by adjustment frictions, under the rationale that bunching could have reached a "steady state" in 1989.

those with high and low volatility of prior earnings have similar adjustment costs.[35]

In Table 2, the variation driving our identification is not confounded with age variation, because we hold ages constant from before to after the policy change (examining 66-68 year-olds in both cases). The variation instead comes from the time series shown in Figure 9 for 66-68 year-olds—specifically, we compare the amount of bunching in 1990 to the amount in 1989 among 66-68 year-olds. Although this particular pattern is in principle indistinguishable from time dummies, we believe that three additional observations make our identification credible.

First, Figure 9 shows that in a "control group" of 62-64 year-olds who do not experience a policy change in 1990, bunching is *extremely* stable in the years before and after 1990, suggesting that the 66-68 year-old group will be sufficient to pick up changes in bunching due to the policy change. The relative comparison demonstrates that bunching fell only slightly among 66-68 year-olds relative to the "control" group of 62-64 year-olds in 1990, but fell more after 1990—suggesting an incomplete reaction to the policy change among the 66-68 year-old group in 1990. Figure 9 also shows that there are no apparent pre-existing trends in the 66-68 year-old group relative to the 62-64 year-old group.

Table 3 verifies that bunching among 66-68 year-olds falls insignificantly in 1990 relative to before 1990, and bunching is significantly smaller in years after 1990, both with and without a linear trend. It also shows that in a "differences-in-differences specification" comparing 66-68 year-olds to 62-64 year-olds, bunching among 66-68 year-olds again falls insignificantly in 1990 relative to before 1990, and bunching is significantly smaller among 66-68 year-olds in years after 1990, again both with and without separate linear trends. Importantly, the "differences-in-differences" estimates comparing ages 66-68 to 62-64 over time (Columns 3 and 4) are very similar to the time series estimates comparing only 66-68 year-olds over time (Columns 1 and 2), further suggesting that the time series analysis of 66-68 year-olds is not substantially confounded by unrelated shocks to bunching over time.

Second, Table 4 shows that when we examine the removal of the kink at age 70 (pooling years 1990-1999), we estimate comparable results to those in Table 2.[36] The Table 4 results

[35] The estimates by group are comparable for the dynamic model estimated below.

[36] Table 4 also shows that we estimate similar results when we use data only on those born in January to March from 1983 to 1999.

cannot be driven by shocks across ages, because there is no reason for bunching at ages over 70 except delayed adjustment to the removal of the kink. When comparing adjustment at age 70 to adjustment in 1990, a key pattern in the data consistent with our model is that the decrease in normalized excess mass from 1989 to 1990 shown in Figure 9 is much smaller (in absolute and percentage terms) than the decrease in normalized excess mass from age 69 to age 70 shown in Figure 3. With an adjustment cost preventing immediate adjustment as in our model, normalized excess mass should fall less when the jump in marginal tax rates at the kink falls less (in the change from a 50 percent to a 33.33 percent BRR in 1990) than when the jump in marginal tax rates at the kink falls more (in the change from a 33.33 percent to a 0 percent BRR at age 70). Table 5 verifies that we also estimate similar results when we pool data from the age 69 to age 71 transition with data from the 1989 to 1990 transition.

Third, Figure 10 shows that the elasticity we estimate among 66-68 year-olds using the Saez (2010) method—constraining the adjustment cost to be zero—shows a sharp, sudden upward spike in bunching in 1990 but subsequently reverts to near its previous level. This relates directly to our theory, which predicts that following a reduction in the change in the MTR at the kink, there may be excess bunching due to inertia (corresponding to area iv in Figure 8, Panel B). Once we allow for an adjustment cost, this excess bunching is attributed to optimization frictions.[37] Thus, the spike in 1990 is consistent with our interpretation that the excess bunching in 1990 reflects delayed adjustment to the policy change, as opposed to some other shock.[38]

7.2 Estimates using the Dynamic Method

Table 6 shows the estimates of the dynamic model described in Sections 6.4.2 and 6.5.2. There are several parameters to estimate—ε, ϕ, and the vector of observed π_{t-t^*}'s—but a limited number of years in the data with useful variation in bunching. Namely, bunching varies little from year to year prior to the policy changes in 1990 or at age 70, and bunching

[37] This figure also serves as additional evidence that adjustment frictions drive our results, rather than the probability of claiming. To explain the 1990 spike in the Saez (2010) elasticity using changes in the probability of claiming, this probability would also have to spike in 1990, but it is constant at 92 percent before, during, and after 1990.

[38] Note that this figure shows that in a context in which individuals have not yet had a chance to adjust and the effective marginal tax rate has fallen, frictions may lead to larger elasticity estimates under the Saez (2010) method. Interestingly, this case yields the opposite of the usual presumption that adjustment frictions lead to attenuation of elasticity estimates that do not account for frictions.

fully dissipates by at most three years after the policy changes. As in Table 5, we pool data on bunching at ages 67, 68, 69, 70, 71, and 72 (pooling 1990 to 1999), with data on bunching among 66-68 year-olds in 1987, 1988, 1989, 1990, 1991, and 1992.[39] This gives us twelve moments (six moments for each of two policy changes) with which to estimate seven parameters (ε, ϕ, π_1, $\pi_1\pi_2$, $\pi_1\pi_2\pi_3$, $\pi_1\pi_2\pi_3\pi_4$, and $\pi_1\pi_2\pi_3\pi_4\pi_5$).[40] We pool the 1990 and age 70 transitions so that we have a sufficient number of moments to estimate the parameters.[41]

In the baseline dynamic specification, we estimate $\varepsilon = 0.36$ and $\phi = \$243$. The estimates of ε are remarkably similar—usually within several percent for a given specification—under the static and dynamic models applied to comparable data, *i.e.* Table 5 and Table 6, respectively. The estimates of ϕ are also in the same range. The point estimate of π_1 varies across specifications from 0.64 in the baseline to 1, indicating that at most a minority of individuals are able to adjust in the year of the policy change. This mirrors our empirical finding that while some individuals adjust in the year of a policy change, particularly to the change from age 69 to 70, there are still many who do not (as observed both from age 69 to 70 and from 1989 to 1990). The point estimate of $\pi_1\pi_2$ varies across specifications from 0.00 to 0.47 (with an estimate of 0.22 in the baseline), indicating that a majority of individuals are able to adjust by the year following a policy change. Again, this mirrors our empirical finding that substantial adjustment occurs from 1990 to 1991. In all specifications, $\pi_1\pi_2\pi_3$ is estimated to be zero, with confidence intervals that rule out more than a modest positive value. Thus, our estimates indicate that individuals are fully able to adjust by the third year after a policy change. Again, this mirrors our empirical finding that adjustment has fully taken place by three years after the policy change, both in 1990 and at age 70. It therefore makes sense that $\pi_1\pi_2\pi_3\pi_4$ and $\pi_1\pi_2\pi_3\pi_4\pi_5$ are also estimated to be zero, with confidence intervals that rule out more than a modest positive value of each.

Given our estimates of the π_j's, it makes sense that we estimate comparable results from the static and dynamic models. If, hypothetically, adjustment were completely constrained

[39] Both at age 70 and in 1990, individuals experienced the previous policy change several years earlier (*e.g.* at age 65, or in 1983) and therefore had ample time to adjust fully.

[40] Note that π_1, π_2, π_3, π_4, and π_5 are not all separately identified; only the cumulative probabilities are identified, *i.e.* π_1, $\pi_1\pi_2$, $\pi_1\pi_2\pi_3$, $\pi_1\pi_2\pi_3\pi_4$, and $\pi_1\pi_2\pi_3\pi_4\pi_5$. The reason is that once one of π_1, π_2, π_3, π_4, or π_5 equals zero, none of the subsequent probabilities is identified. For example, if $\pi_1 = 0$, then any value of π_2 leads to the same predicted level of bunching.

[41] The age 70 and 1990 transitions are slightly different contexts and thus could be governed by different parameters; our pooled estimates will be influenced by variation around both policy changes. However, in both cases, we are examining a change that is anticipated long in advance.

in years 1 and 2 after the policy change and subsequently completely unconstrained, then we should estimate essentially identical results in the static and dynamic models because the static model effectively assumes that the only barrier to adjustment is the adjustment cost ϕ (effectively similar to assuming that $\pi_j = 1$ for the periods over which adjustment is estimated). The dynamic model shows results that are not very different: π_1 is well over 50 percent, and $\pi_1\pi_2$ is substantial but under 50 percent. Meanwhile, subsequent probabilities of facing the adjustment cost are zero. Thus, the immediate adjustment to policy in the first 1-2 years is substantially constrained, and our estimates of the static model during this time frame should show results that are not far from the dynamic model—as is borne out in the estimates.

Given the estimates of (ε,ϕ,$\boldsymbol{\pi}$) from the dynamic model, we can simulate the amount of bunching we should observe at each age. In other words, we simulate how bunching should evolve at ages 62 and over once the AET is introduced at age 62; how bunching should subsequently evolve at ages 65 and over once the AET exempt amount increases markedly from ages 64 to 65; and how bunching should evolve at ages 70 and older, after the elimination of the AET at age 70. These simulation results are shown in Figure 11 Panel A, along with the actual bunching amounts. The pattern in simulated bunching generally tracks the pattern of actual bunching, including in the out-of-sample ages not used for estimation.[42] Similarly, the simulated bunching amounts generally track the data around the policy change from 1989 to 1990 (Figure 11 Panel B): predicted bunching falls somewhat in 1990 and more in 1991, as in the data, and the out-of-sample predictions are also not far from the data.

Figure 2 shows that more individuals appear to "bunch" below the exempt amount than above it. Appendix A.4 explains using simulations how this can arise under our parameter estimates, as in the simulated distribution in Appendix Figure B.5.

8 Conclusion

We investigate earnings adjustment frictions in the context of the Social Security Annual Earnings Test. We introduce a new method for documenting adjustment frictions, which may be more broadly applicable in other economic applications: examining the speed of

[42] The fit across ages is weakest at age 65, but the interpretation of bunching at that age is potentially affected by our coarse measure of age.

adjustment to the disappearance of convex kinks in the effective tax schedule. We document delays in adjustment, consistent with the existence of earnings adjustment frictions in the U.S. Despite the presence of frictions, we find that adjustment is rapid, as the vast majority of adjustment occurs within at most three years of budget set changes. The lack of immediate response suggests that the short-run impact of changes in the effective marginal tax rate can be substantially attenuated, even with large changes in the marginal tax rate such as the 17 percentage point change in 1990 that was accompanied by little immediate change in bunching. We interpret the observed frictions as reflecting a cost of making an immediate adjustment to policy.

Next, we develop a method to estimate earnings elasticities and adjustment costs, using transparent identification relying on bunching at convex budget set kinks, which are commonly encountered in public programs and other economic applications. Examining data in the year of a policy change, we estimate that the elasticity is 0.35 and the adjustment cost is around $280. We interpret our adjustment cost as meaning that bunching in this time frame can be predicted if individuals behaved as if they faced an adjustment cost around $280 and an elasticity of 0.35. When we constrain adjustment costs to zero in the baseline specification, the elasticity we estimate in 1990 (0.58) is substantially (66 percent) larger, demonstrating the potential importance of taking account of adjustment costs. We extend our methods to a dynamic context and continue to find comparable results.

Our estimates demonstrate the applicability of the methodology and the potential importance of allowing for adjustment costs when estimating elasticities. While the methodology we develop is applicable outside our particular context, adjustment costs and elasticities may be substantially different in other contexts. For example, individuals may pay particular attention to policy rules in the context of Social Security and retirement, particularly given that the 50 percent BRR may be very salient. The finding that adjustment costs matter even among those flexible enough to bunch at the kink, and even in the context of a salient policy, suggests the importance of taking frictions into account in other contexts.

Although the difference in the constrained and unconstrained elasticities we estimate is large in percentage terms (66 percent), the absolute difference between the elasticities (0.58 and 0.35) is moderate. However, as demonstrated by the data showing little immediate

reaction even to large policy changes, even modest fixed adjustment costs—like the $280 cost we estimate in our baseline—can greatly impede short-run adjustment to large reforms because the costs of deviating from the frictionless optimum are second order. Adjustment costs can therefore make a dramatic difference in the predictions. This could frustrate the goal of immediately impacting short-run earnings, as envisioned in many recent policy discussions.

Further analysis could enrich our findings. First, we consider our framework for estimating elasticities and adjustment costs to be a natural first step (following papers such as Saez, 2010; Chetty *et al.*, 2009, 2011, 2012a,b; Chetty, 2012; and Kleven and Waseem, 2013), but a next step could involve estimating a model that incorporates additional dynamic considerations (including the potential for anticipatory behavior as we have begun to model). Second, following most previous literature, in our formal model we have treated the adjustment cost as a "black box," without modeling the process that underlies this cost, such as information acquisition or job search. Further work distinguishing among the possible reasons for reaction to the AET, including misperceptions, remains an important issue. Third, further investigation of claiming responses to the AET would be valuable, as Gelber, Jones, Sacks, and Song (2016) are investigating. Finally, an important outstanding question, also suggested by Best and Kleven (2014), is why bunching disappears so quickly after the removal of the kink.

References

Abel, Andrew B. and Janice C. Eberly (1994), "A Unified Model of Investment Under Uncertainty," *American Economic Review*, 84, 1369-1384.

Abowd, John M., Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon Woodcock (2009), "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators." In *Producer Dynamics: New Evidence from Micro Data* (Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts, eds.), 149–230, University of Chicago Press.

Altonji, Joseph G. and Christina H. Paxson (1988), "Labor Supply Preferences, Hours Constraints, and Hours-Wage Trade-Offs," *Journal of Labor Economics*, 6, 254-276.

Arrow, K., T. Harris and J. Marschak (1951), "Optimal Inventory Policy," *Econometrica*, 19, 205-272.

Attanasio, Orazio (2000), "Consumer Durables and Inertial Behavior: Estimation and Aggregation of (S,s) Rules for Automobile Purchases," *Review of Economic Studies*, 67, 667-696.

Baumol, William J. (1952), "The transactions demand for cash: an inventory theoretic approach," *Quarterly Journal of Economics*, 66, 545-556.

Best, Michael, and Henrik Kleven (2014), "Housing Market Responses to Transaction Taxes: Evidence from Notches and Stimulus in the UK," London School of Economics Working Paper.

Blinder, Alan (1981), "Retail inventory behavior and business fluctuations," *Brookings Papers on Economic Activity*, 2, 442-505.

Blundell, Richard, and Hilary Hoynes (2004), "Has 'In-Work' Benefit Reform Helped the Labor Market?" In *Seeking a Premier Economy: The Economic Effects of British Economic reforms, 1980–2000,* (David Card, Richard Blundell, and Richard B. Freeman, eds.), 411–459, University of Chicago Press.

Brown, Jeffrey, Arie Kapteyn, Olivia Mitchell, and Teryn Mattox (2013), "Framing the Social Security Earnings Test," Wharton Pension Research Council Working Paper 2013-06.

Burtless, Gary and Robert A. Moffitt (1985), "The Joint Choice of Retirement Age and Postretirement Hours of Work," *Journal of Labor Economics*, 3, 209–236.

Calvo, Guillermo (1983), "Staggered Prices in a Utility-Maximizing Framework," *Journal of Monetary Economics*, 12, 383-398.

Caplin, Andrew (1985), "The variability of aggregate demand with (S,s) inventory policies," *Econometrica*, 59, 1659-1686.

Caplin, Andrew and John Leahy (1991), "State-dependent pricing and the dynamics of money and output," *Quarterly Journal of Economics*, 106, 683-708.

Caplin, Andrew and Daniel Spulber (1987), "Menu costs and the neutrality of money," *Quarterly Journal of Economics*, 102, 703-725.

Chetty, Raj (2012), "Bounds on Elasticities With Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply," *Econometrica*, 80, 969–1018.

Chetty, Raj, John N. Friedman, Tore Olsen, and Luigi Pistaferri (2011), "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records," *Quarterly Journal of Economics*, 126, 749-804.

Chetty, Raj, Adam Guren, Day Manoli, and Andrea Weber (2012), "Does Indivisible Labor Explain the Difference Between Micro and Macro Elasticities? A Meta-Analysis of Extensive Margin Elasticities." In *NBER Macroeconomics Annual 2012* (Daron Acemoglu, Jonathan Parker, and Michael Woodford, eds.), vol. 27, University of Chicago Press.

Chetty, Raj, Adam Looney, and Kory Kroft (2007), "Salience and Taxation: Theory and Evidence," NBER Working Paper No. 13330.

Chetty, Raj, Adam Looney, and Kory Kroft (2009), "Salience and Taxation: Theory and Evidence," *American Economic Review*, 99, 1145–1177.

Chetty, Raj, John N. Friedman, and Emmanuel Saez (2012), "Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings," forthcoming, *American Economic Review*.

Coile, Courtney, and Jonathan Gruber (2001), "Social security incentives for retirement." In *Themes in the Economics of Aging* (David Wise, ed.), 311-354, University of Chicago Press.

Cooper, Russell W. and John C. Haltiwanger (2006), "On the Nature of Adjustment Costs," *Review of Economic Studies*, 73, 611-633.

Diamond, Peter, and Jonathan Gruber (1999), "Social Security and Retirement in the United States." In *Social Security and Retirement around the World* (Jonathan Gruber and David Wise, eds.), 437-473, University of Chicago Press.

Dixit, Avinash and Robert Pindyck (1994), *Investment Under Uncertainty*, Princeton University Press.

Engelhardt, Gary, and Anil Kumar (2014), "Taxes and the Labor Supply of Older Americans: Recent Evidence from the Social Security Earnings Test," *National Tax Journal*, 67(2), 443-458.

Eissa, Nada, and Jeffrey B. Liebman (1996), "Labor Supply Response to the Earned Income Tax Credit," *Quarterly Journal of Economics*, 111, 605-637.

Farhi, Emmanuel and Xavier Gabaix (2015), "Optimal Taxation with Behavioral Agents," New York University Working Paper.

Feenberg, Daniel, and Elizabeth Coutts (1993), "An Introduction to the TAXSIM Model," *Journal of Policy Analysis and Management*, 12, 189-194.

Friedberg, Leora (1998), "The Social Security earnings test and labor supply of older men." In *Tax Policy and the Economy* (James M. Poterba, ed.), 121–150, University of Chicago Press.

Friedberg, Leora (2000), "The Labor Supply Effects of the Social Security Earnings Test," *Review of Economics and Statistics*, 82, 48–63.

Gelber, Alexander, Damon Jones, and Daniel Sacks (2013), "Earnings Adjustment Frictions: Evidence from the Social Security Earnings Test," NBER Working Paper 19491.

Gelber, Alexander, Damon Jones, Daniel Sacks, and Jae Song (2016), "Extensive Margin Responses to the Social Security Earnings Test." Mimeo, Indiana University.

Grossman, Sanford and Guy Laroque (1990), "Asset pricing and optimal portfolio choice in the presence of illiquid durable consumption goods," *Econometrica*, 58, 25-51.

Gruber, Jonathan (2013), *Public Finance and Public Policy*. New York: Worth Publishers.

Gruber, Jonathan and Peter Orszag (2003), "Does the Social Security Earnings Test Affect Labor Supply and Benefits Receipt?" *National Tax Journal*, 56, 755–773.

Gudgeon, Matthew, and Simon Trenkle (2016), "Frictions in Adjusting Earnings: Evidence from Notches in German Mini Jobs," Boston University Working Paper.

Hausman, Jerry A. (1981), "Labor Supply." In *How Taxes Affect Economic Behavior* (Henry J. Aaron and Joseph A. Pechman, eds.), 27–71, Brookings Institution.

He, Daixin, Langchuan Peng, and Xiaxin Wang (2016), "Understanding China's Personal Income Tax," UC San Diego Working Paper.

Hoopes, Jeffrey, Daniel Reck, and Joel Slemrod (2013), "Taxpayer Search for Information: Implications for Rational Attention," University of Michigan Working Paper.

Kleven, Henrik (2016), "Bunching," *Annual Review of Economics*, Volume 8.

Kleven, Henrik, Martin Knudsen, Claus Kreiner, Søren Pedersen, and Emmanuel Saez (2011), "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark," *Econometrica*, 79, 651-692.

Kleven, Henrik, Camille Landais, Emmanuel Saez, and Esben Schultz (2012), "Taxation and International Migration of Top Earners: Evidence from the Foreigner Tax Scheme in Denmark," forthcoming, *American Economic Review*.

Kleven, Henrik and Mazhar Waseem (2013), "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan," *Quarterly Journal of Economics*, 128, 669-723.

Kline, Patrick, and Christopher Walters (2016), "Evaluating Public Programs with Close Substitutes: The Case of Head Start," forthcoming, *Quarterly Journal of Economics.*

Leahy, John (2008), "s-S Models." In *The New Palgrave Dictionary of Economics, Second Edition* (Steven N. Durlauf and Lawrence E. Blume, eds.), Palgrave MacMillan.

Levine, Ross, and Yona Rubinstein (2013), "Does Entrepreneurship Pay? The Michael Bloombergs, the Hot Dog Vendors, and the Returns to Self-Employment." UC Berkeley Working Paper.

Liebman, Jeffrey B., and Erzo F.P. Luttmer (2011), "Would People Behave Differently If They Better Understood Social Security? Evidence From a Field Experiment," NBER Working Paper 17287.

Liebman, Jeffrey, & Erzo F.P. Luttmer (2012), "The Perception of Social Security Incentives for Labor Supply and Retirement: The Median Voter Knows More Than You'd Think." In *Tax Policy and the Economy* 26, 1-42.

Liebman, Jeffrey B., Erzo F.P. Luttmer, and David Seif (2009), "Labor Supply Responses to Marginal Social Security Benefits: Evidence from Discontinuities," *Journal of Public Economics*, 93, 1208-1223.

McKinney, Kevin L. and Lars Vilhuber (2008), "LEHD Infrastructure Files in the Census RDC - Overview Revision: 219," U.S. Census Bureau, LEHD Program.

Meyer, Bruce D. and Dan T. Rosenbaum (2001), "Welfare, the Earned Income Tax Credit, and the Labor Supply of Single Mothers," *Quarterly Journal of Economics*, 116, 1063–1114.

Moffitt, Robert (1990), "The Econometrics of Kinked Budget Constraints," *Journal of Economic Perspectives*, 4, 119–39.

Nakamura, Emi and Jón Steinsson (2010), "Monetary Non-Neutrality in a Multisector Menu Cost Model," *Quarterly Journal of Economics*, 125, 961-1013.

Newey, Whitney, and Daniel McFadden (1994), "Large Sample Estimation and Hypothesis Testing." In R.F. Engle and D. McFadden, eds., *Handbook of Econometrics*, Vol. 4.

Reiss, Peter, and Matthew White (2005), "Household Electricity Demand, Revisited," *Review of Economic Studies*, 72, 853-883.

Saez, Emmanuel (2010), "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy*, 2, 180–212.

Saez, Emmanuel, Joel Slemrod, and Seth H. Giertz (2012), "The Elasticity of Taxable Income with Respect to Marginal Tax Rates: A Critical Review," *Journal of Economic Literature*, 50, 3–50.

Simon, Herbert A. (1955), "A Behavioral Model of Rational Choice," *Quarterly Journal of Economic*s, 69, 99–118.

Sheshinski, Eytan and Yoram Weiss (1977), "Inflation and the cost of price adjustment," *Review of Economic Studies*, 44, 287-303.

Social Security Administration (2012a), *Annual Statistical Supplement.* Washington, D.C.

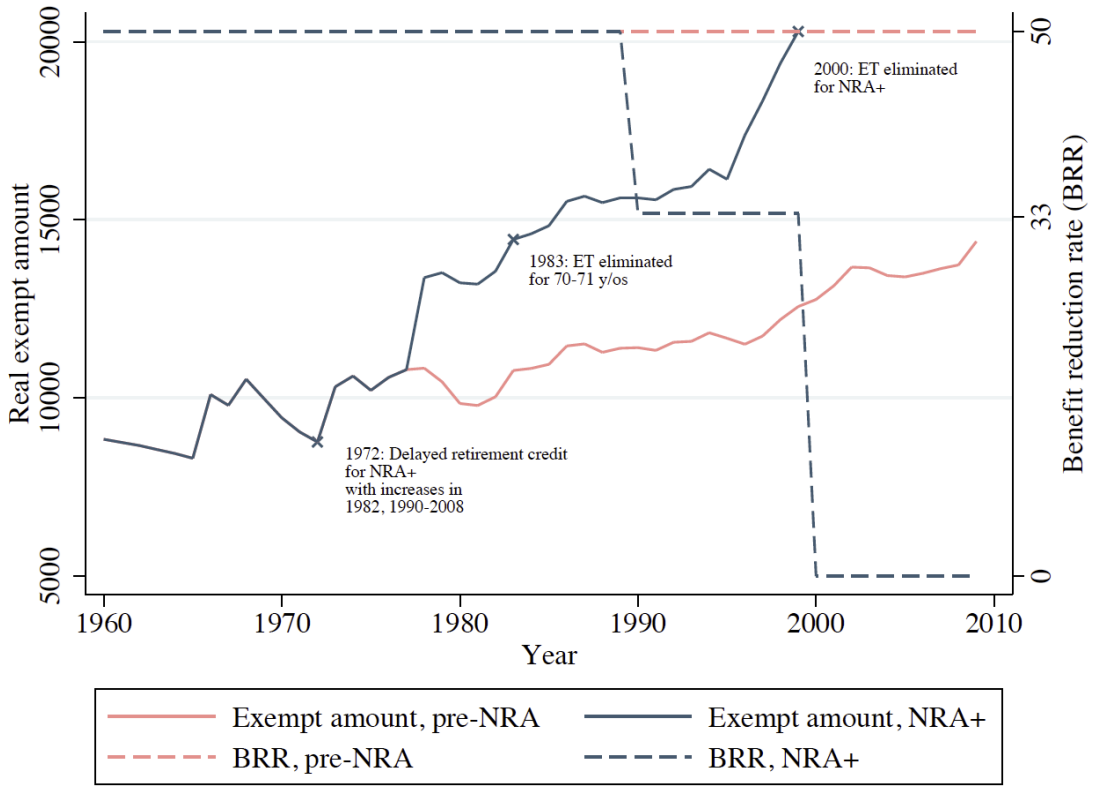
Social Security Administration (2012b), *Social Security Handbook*. Washington, D.C.

Song, Jae G. and Joyce Manchester (2007), "New evidence on earnings and benefit claims following changes in the retirement earnings test in 2000," *Journal of Public Economics*, 91, 669–700.

Stokey, Nancy (2008), *The Economics of Inaction*, Princeton University Press.

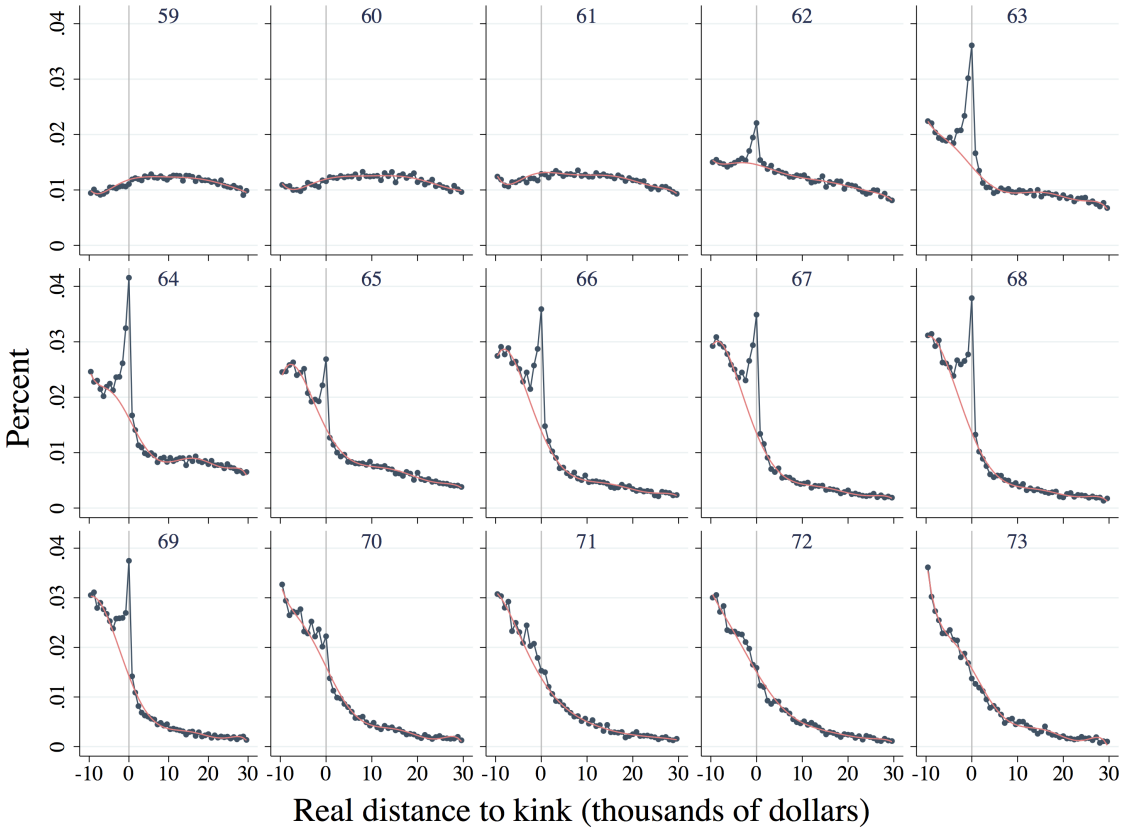
Werquin, Nicolas (2015), "Income Taxation with Frictional Labor Supply," Toulouse School of Economics Working Paper.

Figure 1: Key Earnings Test Rules, 1961-2009



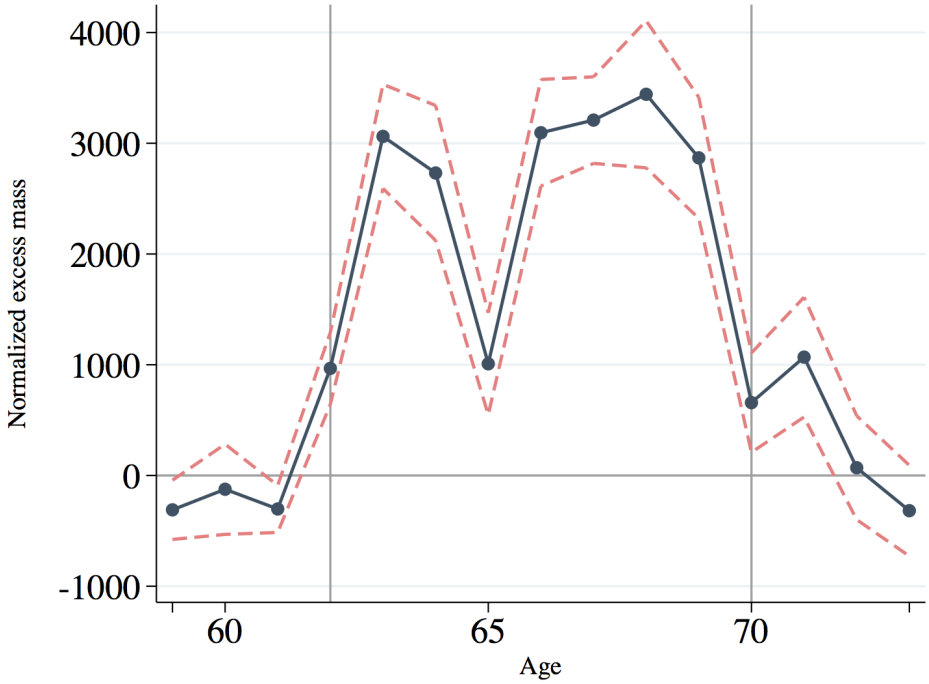
Notes: The right vertical axis measures the benefit reduction rate (BRR) in OASI payments for every dollar earned beyond the exempt amount. The left vertical axis measures the real value of the exempt amount over time.

Figure 2: Histograms of Earnings, 59-73-year-olds Claiming OASI by Age 65, 1990-1999



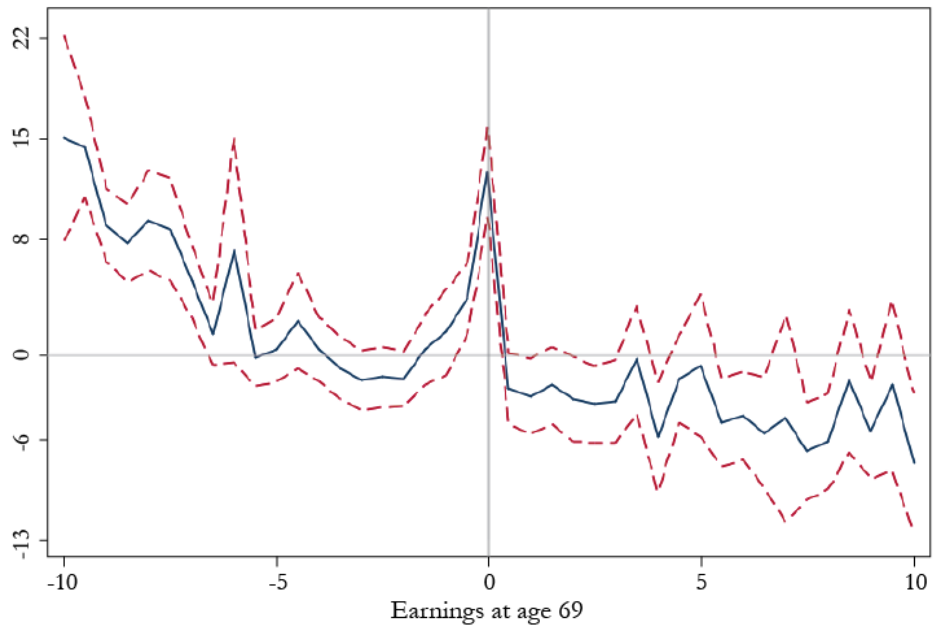
Notes: The sample is a one-percent random sample of all Social Security numbers, limited to individuals who claim OASI benefits by age 65. We exclude person-years with self-employment income or with zero non-self-employment earnings. The bin width is $800. The earnings level zero, shown by the vertical lines, denotes the kink. The dots show the histograms using the raw data, and the polynomial curves show the estimated counterfactual densities estimated using data away from the kink.

Figure 3: Adjustment Across Ages: Normalized Excess Mass, 59-73 Year-Olds Claiming OASI by Age 65, 1990-1999



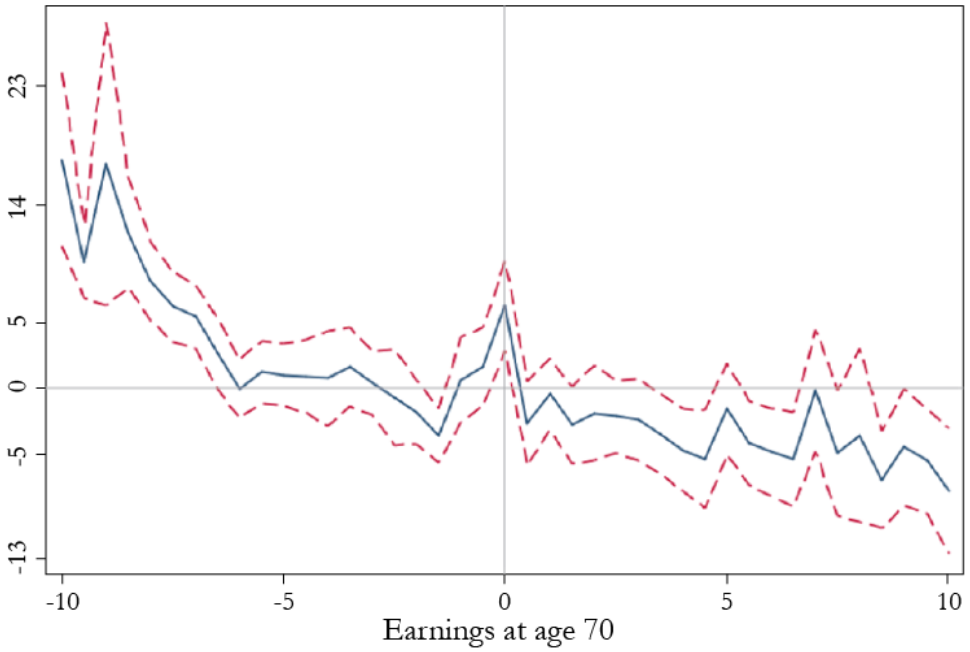
Notes: The figure shows normalized bunching at the AET kink, calculated as described in Section 3. Dashed lines denote 95% confidence intervals. The vertical lines show the ages at which the AET first applies (62) and the age at which the AET ceases to apply (70). See also notes from Figure 2.

Figure 4: Mean Percentage Change in Earnings from Age 69 to 70, by Earnings at 69, 1990-1998



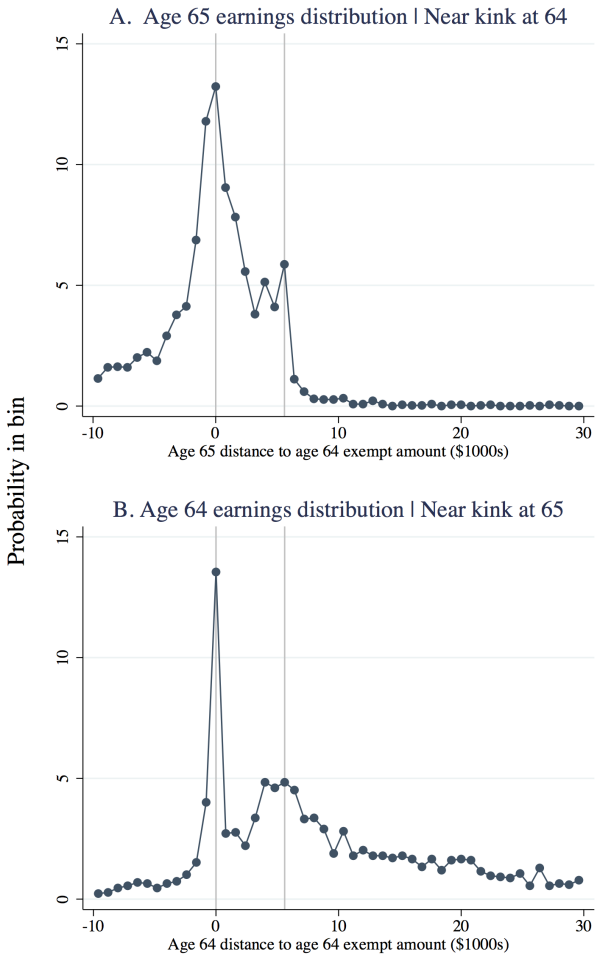
Notes: The figure shows the mean percentage change in earnings from age 69 to age 70 (y-axis), against earnings at age 69 (x-axis). Dashed lines denote 95 percent confidence intervals. Earnings are measured relative to the kink, shown at zero on the x-axis. The data are a 20 percent random sample of 69-year-olds in the LEHD in 1990-1998. We exclude 1999 as a base year in this and similar graphs because the AET is eliminated for those older than NRA in 2000. Higher earnings growth far below the kink reflects mean reversion visible in this part of the earnings distribution at all ages.

Figure 5: Mean Percentage Change in Earnings from Age 70 to 71, by Earnings at 70, 1990-1998



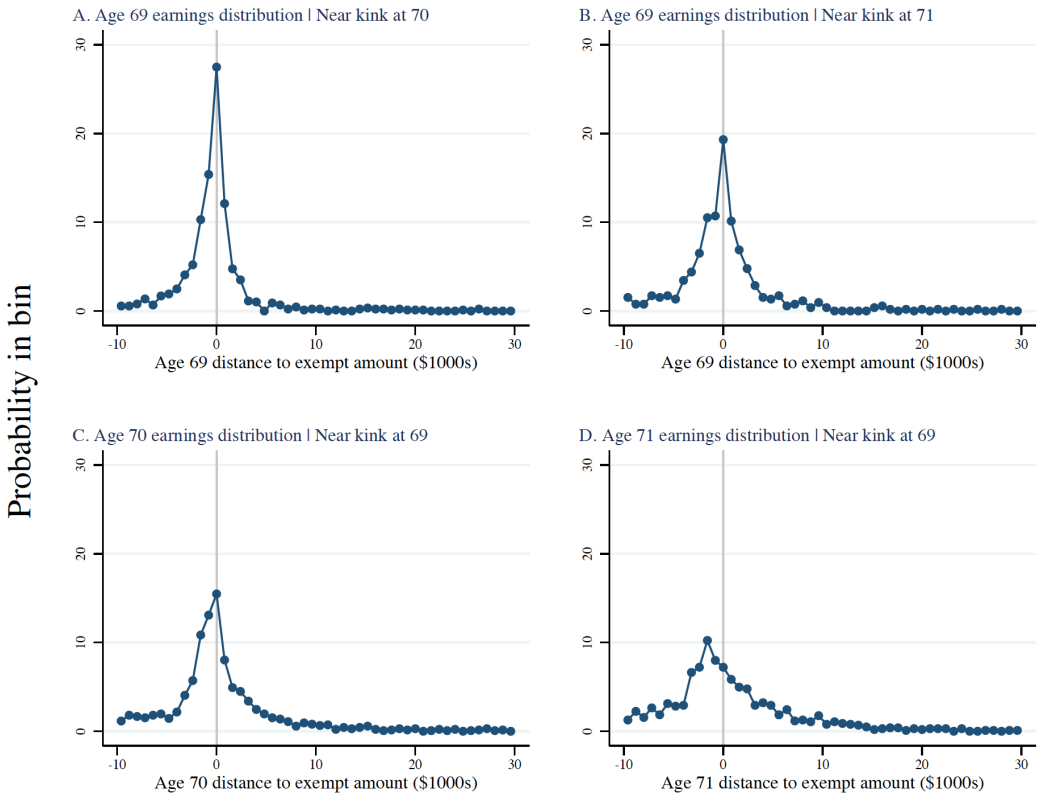
Notes: The figure shows the mean percentage change in earnings from age 70 to age 71 (y-axis), against earnings at age 70 (x-axis). Dashed lines denote 95 percent confidence intervals. Earnings are measured relative to the kink, shown at zero on the x-axis. The data are a 20 percent random sample of 70-year-olds in the LEHD in 1990-1998. We exclude 1999 as a base year in this and similar graphs because the AET is eliminated for those older than NRA in 2000. Higher earnings growth far below the kink reflects mean reversion visible in this part of the earnings distribution at all ages.

Figure 6: Inertia in Bunching from 64 to 65



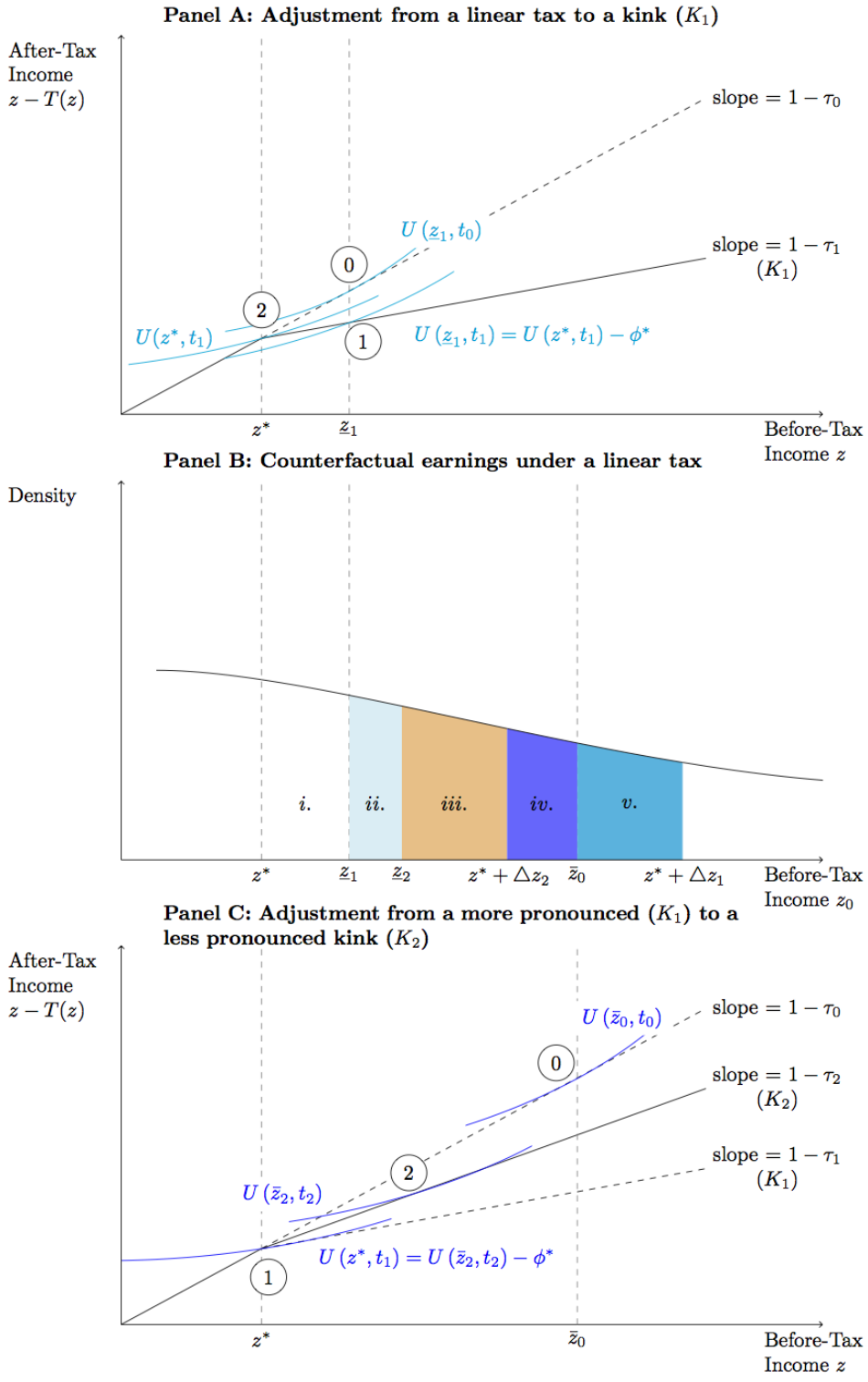
Notes: Panel A of the figure shows that when they are age 65, those previously bunching at age 64 tend to either (a) remain near the age 64 exempt amount or (b) move to the age 65 exempt amount. A greater fraction remains near the age 64 exempt amount than the fraction that moves to the age 65 exempt amount. Panel B of the figure shows that those bunching at age 65 were usually bunching at age 64 in the previous year (or were near the age 65 exempt amount in the previous year). Having earnings "near the kink" at a given age is defined as having earnings within $1,000 of the kink at that age; in other words, "near kink at 64" means that the individual has age 64 earnings within $1,000 of the exempt amount applying at age 64, and "near kink at 65" means that the individual has age 65 earnings within $1,000 of the exempt amount applying at age 65. The first vertical line at zero shows the location of the age 64 exempt amount (normalized to zero), and the second vertical line shows the average location of the age 65 exempt amount (relative to the age 64 exempt amount that has been normalized to zero). The years of data used are 1990 to 1999.

Figure 7: Inertia in Bunching from 69 to 70 and 71



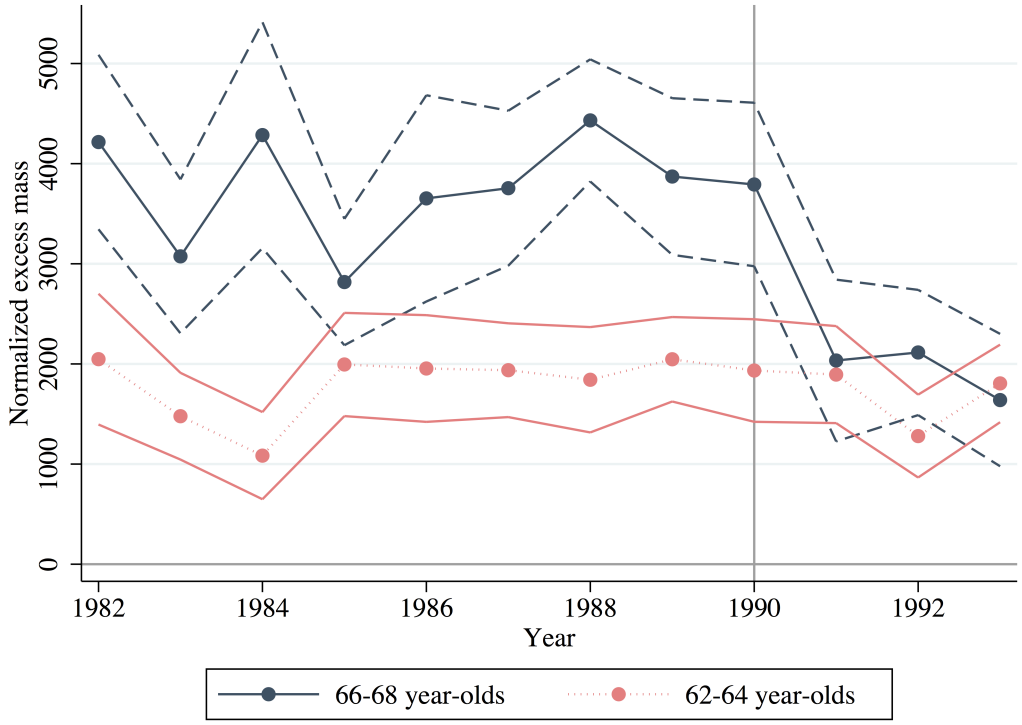
Notes: Using data from 1990 to 1999, the figure shows that those bunching at age 69 tend to remain near the kink at ages 70 and 71, and that those bunching at ages 70 and 71 were also bunching at age 69. Specifically, the figure shows the density of earnings at age 69 conditional on having earnings near the kink at age 70 (Panel A), the density of earnings at age 69 conditional on having earnings near the kink at age 71 (Panel B), the density of earnings at age 70 conditional on having earnings near the kink at age 69 (Panel C), and the density of earnings at age 71 conditional on having earnings near the kink at age 69 (Panel D). Having earnings "near the kink" is defined as having earnings within $1,000 of the exempt amount applying to that age. See also notes from Figure 6.

Figure 8: Bunching Responses to a Convex Kink, with Fixed Adjustment Costs



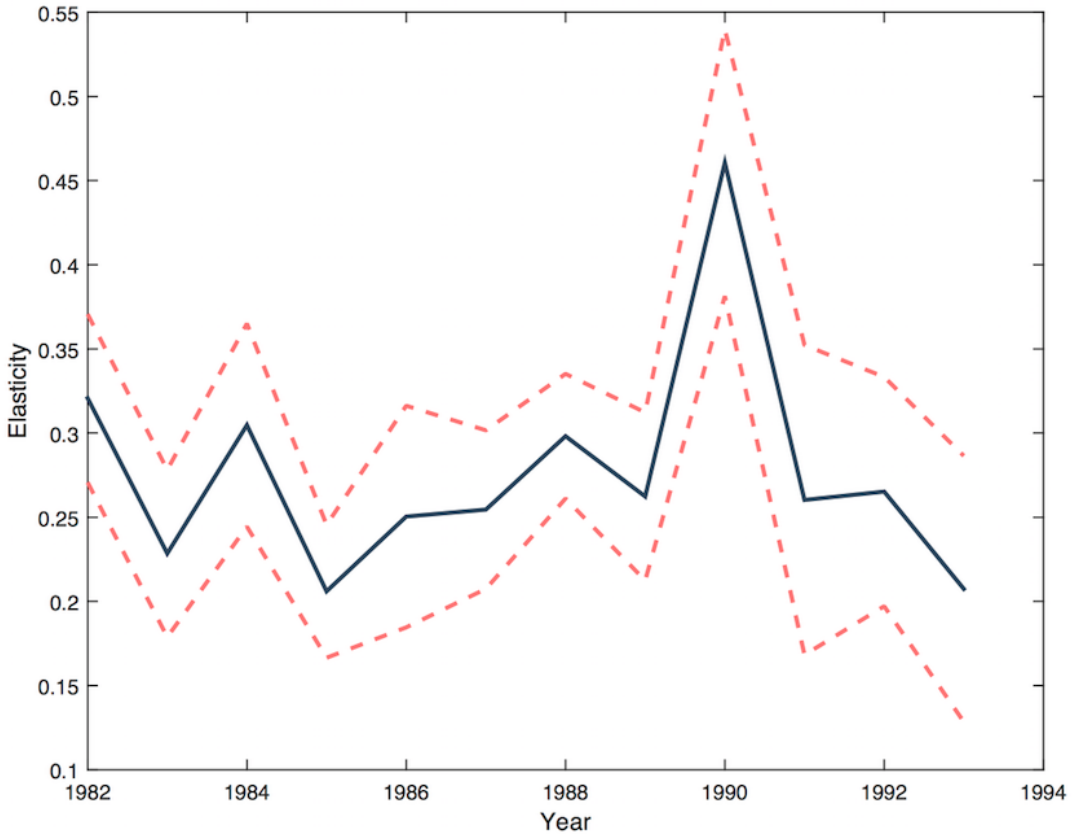
Note: See Section 6 for an explanation of the figures.

Figure 9: Comparison of Normalized Excess Mass Among 62-64 Year-Olds and 66-68 Year-Olds, 1982-1993



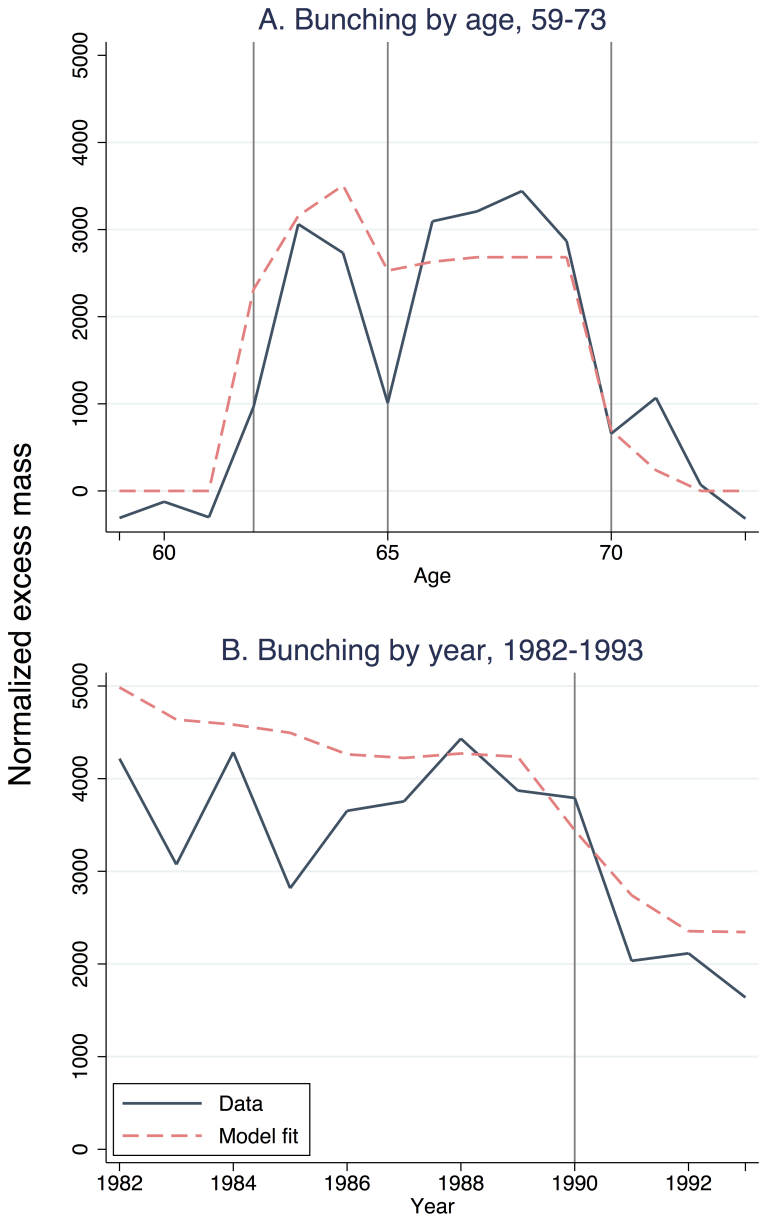
Notes: The figure shows normalized bunching among 62-64 year-olds and 66-68 year-olds in each year from 1982 to 1993. See other notes from Figure 2.

Figure 10: Elasticity Estimates by Year, Saez (2010) Method, 1982-1993



Notes: The figure shows elasticities estimated using the Saez (2010) method, by year from 1982 to 1993, among 66-68 year-old OASI claimants. Dashed lines denote 95 percent confidence intervals. We use our methods for estimating normalized excess mass but use Saez' (2010) formula to calculate elasticities, under a constant density. This method yields the following formula: $\varepsilon = \left[\log\left(\frac{b}{z^*}+1\right)\right] / \left[\log\left(\frac{1-\tau_0}{1-\tau_1}\right)\right]$.

Figure 11: Simulated Bunching Amounts from Dynamic Estimation Method



Notes: The figure shows the bunching amounts estimated from the data as the solid line, along with the simulated bunching amounts predicted by our dynamic model as the dashed line. Panel A shows that the predicted amounts generally track the rise in bunching at age 62 and the gradual dissipation at ages 70 and 71. Panel B shows that predicted bunching falls somewhat in 1990 and more in 1991, as in the data. Predicted bunching falls in the years prior to 1990 because the exempt amount is rising in real terms over these years. This implies that a more sparsely populated part of the earnings distribution is affected by the earnings test, thus reducing predicted bunching over time.

Table 1: Summary Statistics, Social Security Administration Master Earnings File

	Ages 62-69
Mean Earnings	28,892.63
	(78,842.99)
10th Percentile	1,193.64
25th Percentile	5,887.75
50th Percentile	14,555.56
75th Percentile	35,073.00
90th Percentile	64,647.40
Fraction Male	0.57
Observations	376,431

Notes: The data are taken from a one percent random sample of the SSA Master Earnings File and Master Beneficiary Record. The data cover those in 1990-1999 who are aged 62-69, claim by age 65, do not report self-employment earnings, and have positive earnings. Earnings are expressed in 2010 dollars. Numbers in parentheses are standard deviations.

Table 2: Estimates of Elasticity and Adjustment Cost: Variation Around 1990 Policy Change

	(1)	(2)	(3)	(4)
	ε	ϕ	$\varepsilon \vert \phi = 0$	
			1990	1989
Baseline	0.35	\$278	0.58	0.31
	[0.31, 0.43]***	[58, 391]***	[0.45, 0.73]***	[0.24, 0.39]***
Uniform Density	0.21	\$162	0.36	0.19
	[0.18, 0.24]***	[55, 211]***	[0.30, 0.43]***	[0.16, 0.23]***
Benefit Enhancement	0.58	\$151	0.87	0.52
	[0.50, 0.72]***	[17, 226]***	[0.69, 1.11]***	[0.41, 0.66]***
Excluding FICA	0.49	\$318	0.74	0.42
	[0.44, 0.59]***	[60, 364]***	[0.58, 0.94]***	[0.33, 0.54]***
Bandwidth = \$400	0.45	\$103	0.62	0.43
	[0.36, 0.58]***	[0, 478]*	[0.47, 0.81]***	[0.32, 0.56]***
Bandwidth = \$1,600	0.33	\$251	0.55	0.30
	[0.29, 0.43]***	[34, 407]***	[0.43, 0.72]***	[0.23, 0.40]***

Notes: The table shows estimates of the elasticity and adjustment cost using the method described in Section 6.3. We report bootstrapped 95 percent confidence intervals shown in parentheses. We investigate the 1990 reduction in the AET BRR from 50 percent to 33.33 percent. The baseline specification uses a nonparametric density taken from the age 72 earnings distribution, calculates the effective MTR by including the effects of the AET BRR and federal and state income and FICA taxes, uses data from 1989 and 1990, and calculates bunching using a bin width of \$800. Alternative specifications deviate from the baseline as noted. The estimates that include benefit enhancement use effective marginal tax rates due to the AET based on the authors' calculations relying on Coile and Gruber (2001) (assuming that individuals are considering earning just enough to trigger benefit enhancement). This translates the BRR before and after the 1990 policy change to 36% and 24%, respectively. Columns (1) and (2) report joint estimates with $\phi \geq 0$ imposed (consistent with theory, as described in the Appendix), while Columns (3) and (4) impose the restriction $\phi = 0$. The constrained estimate in Column (3) only uses data from 1990, whereas that in Column (4) uses only data from 1989. *** indicates that the left endpoint of the 99 percent confidence interval is greater than zero, ** the 95 percent confidence interval and * the 90 percent confidence interval.

Table 3: Estimates of Changes in Bunching Around 1990

Sample	Old only	Old only, linear trend	DD	DD, separate linear trend
old x 1990 dummy	28.9	-165.1	-107.3	-69.2
	(249.1)	(411.0)	(306.7)	(411.7)
old x 1991 dummy	-1728.9	-1966.0	-1824.5	-1777.9
	(249.1)***	(500.6)***	(306.7)***	(481.3)***
old x 1992 dummy	-1648.8	-1928.9	-1130.2	-1075.1
	(249.1)***	(594.9)***	(306.7)***	(558.1)*
old x 1993 dummy	-2123.8	-2447.1	-2131.2	-2067.6
	(249.1)***	(692.1)***	(306.7)***	(639.7)***
Ages	66-68	66-68	62-64, 66-68	62-64, 66-68
Year FE?	No	No	Yes	Yes
Linear time trend (in year)	No	Yes	No	No
Separate linear trend for "old"	No	No	No	Yes

Notes: The table shows that the estimated change in bunching amounts from before to after 1990 in the age 66-68 age group are similar under several specifications. The dummy variable "old" indicates the older age group (66-68). The sample in Columns (1) and (2) includes only 66-68 year-olds, and in Columns (3) and (4) it also includes 62-64 year-olds. Additional controls include a linear time trend (in year) in column (2), year fixed effects in columns (3) and (4), and the linear time trend interacted with the "old" dummy in column (4). Robust standard errors are in parentheses. Under all the specifications, the coefficient on old x 1990 is insignificantly different from zero: bunching in 1990 is not significantly different from prior bunching, indicating that adjustment does not immediately occur. However, the coefficients on old x 1991, old x 1992, old x 1993 are negative and significant, indicating that bunching falls significantly after 1990—*i.e.* a reduction in bunching does eventually occur (but not immediately in 1990). The fact that the results are similar under all these various specifications indicates that the results are little changed by controlling for a linear trend (Column 2), comparing 66-68 year-olds to a reasonable control group of 62-64 year-olds (Column 3), and additionally controlling for a separate linear trend for the older group (Column 4). In Columns 1 and 3, the standard errors are the same across all of the interaction coefficients shown because there is only one observation underlying each dummy, and the dummies are exactly identified. See also notes from Table 2.

Table 4: Estimates of Elasticity and Adjustment Cost: Disappearance of Kink at Age 70

	(1)	(2)	(3)
	ε	ϕ	$\varepsilon \vert \phi = 0$, Age 69
Baseline	0.42	\$90	0.38
	[0.35, 0.53]***	[20, 349]***	[0.32, 0.47]***
Uniform Density	0.28	\$90	0.25
	[0.24, 0.33]***	[21, 238]***	[0.22, 0.30]***
Benefit Enhancement	0.62	\$59	0.58
	[0.53, 0.77]***	[13, 205]***	[0.49, 0.71]***
Excluding FICA	0.53	\$83	0.49
	[0.45, 0.66]***	[19, 305]***	[0.42, 0.61]***
Bandwidth = \$400	0.39	\$62	0.36
	[0.31, 0.48]***	[25, 133]***	[0.28, 0.45]***
Bandwidth = \$1,600	0.45	\$100	0.41
	[0.37, 0.56]***	[20, 444]***	[0.33, 0.49]***
68-70 year-olds	0.44	\$42	0.43
	[0.38, 0.58]***	[0.49, 267]**	[0.37, 0.50]***
69, 71 year-olds	0.45	\$175	0.38
	[0.36, 0.86]***	[30, 1053]***	[0.32, 0.47]***
Born January-March	0.48	\$86	0.49
	[0.36, 0.76]***	[10, 1008]***	[0.37, 0.71]***

Notes: The table estimates elasticities and adjustment costs using the removal of the AET at age 70, using data on 69-71 year-olds in 1990-1999. We cannot estimate the constrained elasticity using only data on age 70 because the benefit reduction rate is zero at that age. The estimates of bunching at age 70 are potentially affected by the coarse measure of age that we use, as explained in the main text. To address this issue, we use both age 70 and age 71 in estimating these results. To further address this issue, in the second-to-last row of the table, we use only ages 69 and 71, which shows very similar results—this is unsurprising because Figure 3 shows that normalized excess mass is similar at ages 70 and 71. The row labeled "68-70 year-olds" uses data from ages within this range. The final row provides the estimates only for those born in January to March, again to avoid issues relating to the coarse measurement of age (as explained in the main text). For this sample, we pool 1983-1989 and 1990-1999 (accounting for the different benefit reduction rates in each period) to maximize statistical power. See also notes from Table 2.

Table 5: Estimates of Elasticity and Adjustment Cost Using Comparative Static Method and Pooling 69/70 Transition and 1989/1990 Transition

	(1)	(2)
	ε	ϕ
Baseline	0.39	\$160
	[0.34, 0.46]***	[59, 362]***
Uniform Density	0.22	\$105
	[0.20, 0.25]***	[47, 185]***
Benefit Enhancement	0.62	\$100
	[0.55, 0.75]***	[33, 211]***
Excluding FICA	0.41	\$67
	[0.37, 0.56]***	[9, 192]***
Bandwidth = \$400	0.46	\$94
	[0.39, 0.56]***	[25, 399]***
Bandwidth = \$1,600	0.37	\$135
	[0.32, 0.45]***	[43, 299]***

Notes: This table implements our "comparative static" method, applied to *pooled* data from two policy changes: (1) around the 1989/1990 transition analyzed in Table 2, and (2) around the age 69/70 transition analyzed in Table 4. The table shows extremely similar results to the dynamic specification in Table 6, where we also pool data from around these two policy changes. See also notes from Tables 2 and 4.

Table 6: Estimates of Elasticity and Adjustment Cost Using Dynamic Model

	(1)	(2)	(3)	(4)	(5)
	ε	ϕ	π_1	$\pi_1\pi_2$	$\pi_1\pi_2\pi_3$
Baseline	0.36	\$243	0.64	0.22	0.00
	[0.34, 0.40]***	[34, 671]***	[0.39, 1.00]***	[0.00, 0.94]*	[0.00, 0.14]*
Uniform Density	0.21	\$81	1.00	0.31	0.00
	[0.20, 0.23]***	[31, 183]***	[0.72, 1.00]***	[0.00, 0.92]***	[0.00, 0.16]***
Benefit Enhancement	0.59	\$53	1.00	0.37	0.00
	[0.54, 0.64]***	[18, 169]***	[0.76, 1.00]***	[0.00, 1.00]***	[0.00, 0.084]***
Excluding FICA	0.40	\$55	1.00	0.00	0.00
	[0.37, 0.43]***	[9, 165]***	[1.00, 1.00]***	[0.00, 0.00]***	[0.00, 0.00]***
Bandwidth = \$400	0.40	\$74	1.00	0.47	0.00
	[0.36, 0.44]***	[20, 271]***	[0.74, 1.00]***	[0.094, 0.94]***	[0.00, 0.20]***
Bandwidth = \$1,600	0.36	\$99	0.88	0.52	0.00
	[0.34, 0.39]***	[19, 401]***	[0.40, 1.00]***	[0.043, 1.00]***	[0.00, 0.071]***

Notes: The table shows estimates of the elasticity and adjustment cost using the dynamic method described in Section 7. The table reports the elasticity ε, the adjustment cost ϕ, and the cumulative probability in each period t of having drawn $\phi_t > 0$ for each period following the policy change, *i.e.* π_1 as well as $\pi_1\pi_2$. The model is estimated by matching predicted and observed bunching, using bunching on 66-68 year-olds (pooled) for each year 1987-1992, and bunching on 1990-1999 (pooled) for each age 67-72. Estimates of $\pi_1\pi_2\pi_3$ are statistically significantly different from zero, even though the reported point estimates are 0.00, because the point estimates are positive but round to zero. $\pi_1\pi_2\pi_3\pi_4$ and $\pi_1\pi_2\pi_3\pi_4\pi_5$ are always estimated to 0.00, with a confidence interval that rules out more than a small value (results available upon request). The results are comparable when we investigate only the 1989/1990 or 69/70 policy changes alone using the dynamic specification (results available upon request). *** indicates *p<0.01*;**p<0.05*; * *p<0.10*.

A Appendix: Model of Earnings Response (for online publication)

A.1 Derivation of Bunching Formulae with Heterogeneity

A.1.1 Comparative Static Model

Under heterogenous preferences, our estimates can be interpreted as reflecting average parameters among the set of bunchers (as in Saez, 2010, and Kleven and Waseem, 2013). As described in Section 6.4.2, suppose $(\varepsilon_i, \phi_i, a_i)$ is jointly distributed according to a smooth CDF, which translates to a smooth, joint distribution of elasticities, fixed costs and earnings. Let the joint density of earnings, adjustment costs and elasticities be $h_0^*(z, \varepsilon, \phi)$ under a linear tax of τ_0. Assume that the density of earnings is constant over the interval $[z^*, z^* + \Delta z^*]$, conditional on ε and ϕ. When moving from no kink to a kink, we derive a formula for bunching at K_1 in the presence of heterogeneity as follows:

$$\begin{aligned} B_1 &= \iiint_{\underline{z}_1}^{z^*+\Delta z_1^*} h_0^*(\zeta, \epsilon, \varphi)\, d\zeta d\epsilon d\varphi \\ &= \iint [z^* + \Delta z_1^* - \underline{z}_1]\, h_0^*(z^*, \epsilon, \varphi)\, d\epsilon d\varphi \\ &= h_0(z^*) \cdot \iint [z^* + \Delta z_1^* - \underline{z}_1] \frac{h_0^*(z^*, \epsilon, \varphi)}{h_0(z^*)} d\epsilon d\varphi \\ &= h_0(z^*) \cdot \mathbb{E}[z^* + \Delta z_1^* - \underline{z}_1], \end{aligned} \tag{A.1}$$

where we have used the assumption of constant $h_0^*(\cdot)$ in line two, $h_0(z^*) = \iint h_0^*(z^*, \epsilon, \varphi)\, d\epsilon d\varphi$, and ζ, ϵ and φ are dummies of integration. The expectation $\mathbb{E}[\cdot]$ is taken over the set of bunchers, under the various combinations of ε and ϕ throughout the support. It follows that normalized bunching can be expressed as follows:

$$b_1 = z^* + \mathbb{E}[\Delta z_1^*] - \mathbb{E}[\underline{z}_1]. \tag{A.2}$$

Under heterogeneity, the level of bunching identifies the average behavioral response, Δz^*, and threshold earnings, $\underline{z}_1$, among the marginal bunchers under each possible combination of parameters ε and ϕ. Under certain parameter values, there is no bunching, and thus, the values of the elasticity and adjustment cost in these cases do not contribute our estimates.

When we move sequentially from a larger kink, K_1 to a smaller kink, K_2, our formula for

bunching under K_2 in the presence of heterogeneity is likewise derived as follows:

$$\begin{aligned}
\tilde{B}_2 &= \iiint_{\underline{z}_1}^{\bar{z}_0} h_0^*(\zeta, \epsilon, \varphi)\, d\zeta d\epsilon d\varphi \\
&= \iint [\bar{z}_0 - \underline{z}_1]\, h_0^*(z^*, \epsilon, \varphi)\, d\epsilon d\varphi \\
&= h_0(z^*) \cdot \iint [\bar{z}_0 - \underline{z}_1] \frac{h_0^*(z^*, \epsilon, \varphi)}{h_0(z^*)} d\epsilon d\varphi \\
&= h_0(z^*) \cdot \mathbb{E}[\bar{z}_0 - \underline{z}_1]. \qquad \text{(A.3)}
\end{aligned}$$

Similarly, normalized bunching can now be expressed as follows:

$$\tilde{b}_2 = \mathbb{E}[\bar{z}_0] - \mathbb{E}[\underline{z}_1]. \qquad \text{(A.4)}$$

Once again, the expectations are taken over the population of bunchers.

Following the approach in Kleven and Waseem (2013, pg. 682), the average value of the parameters Δz_1^*, $\underline{z}_1$ and $\bar{z}_0$ can then be related to ε and ϕ, assuming a quasi-linear utility function and using (5) and (7) and the identities $\triangle z_1^* = \varepsilon z^* d\tau_1/(1-\tau_0)$ and $\bar{z}_0 - \bar{z}_2 = \varepsilon \bar{z}_2 d\tau_2/(1-\tau_0)$.

A.1.2 Dynamic Model

A similar interpretation of our results holds when we turn to our more dynamic framework in Section 6.4.1. Suppose now that $(\varepsilon_i, \phi_i, a_i, \boldsymbol{\pi}_i)$ is jointly distributed according to a smooth CDF, which results in a smooth, joint distribution of elasticities, fixed costs, earnings, and probabilities of drawing a positive fixed cost. In order to gain tractability, we assume that the profile $\boldsymbol{\pi}_i$ is independent of the parameters $(\varepsilon_i, \phi_i, a_i)$. The result is that the joint density of these parameters, under a linear tax of τ_0, can be expressed as a product of two densities: $h_0^*(z, \varepsilon, \phi)\, g(\boldsymbol{\pi}_i)$. We maintain the assumption that the density of earnings is constant over the interval $[z^*, z^* + \Delta z^*]$, conditional on ε and ϕ. Bunching at K_1 in period $t \in [1, \mathcal{T}_1]$ will

now be:

$$\begin{aligned}
B_1^t &= \iiiint_{\underline{z}_1}^{z^*+\Delta z_1^*} h_0^*(\zeta,\epsilon,\varphi)\, g(\boldsymbol{\pi})\, d\zeta d\epsilon d\varphi d\boldsymbol{\pi} \\
&\quad + \iiiint_{z^*}^{\bar{z}_1} \left(1-\Pi_{j=1}^t \pi_j\right) h_0^*(\zeta,\epsilon,\varphi)\, g(\boldsymbol{\pi})\, d\zeta d\epsilon d\varphi d\boldsymbol{\pi} \\
&= \iint \left[z^*+\Delta z_1^*-\underline{z}_1\right] h_0^*(z^*,\epsilon,\varphi) \left(\int g(\boldsymbol{\pi})\, d\boldsymbol{\pi}\right) d\epsilon d\varphi \\
&\quad + \iint \left[\underline{z}_1-z^*\right] h_0^*(z^*,\epsilon,\varphi) \left(\int \left(1-\Pi_{j=1}^t \pi_j\right) g(\boldsymbol{\pi})\, d\boldsymbol{\pi}\right) d\epsilon d\varphi \\
&= h_0(z^*) \left\{ \iint \left[z^*+\Delta z_1^*-\underline{z}_1\right] \frac{h_0^*(z^*,\epsilon,\varphi)}{h_0(z^*)} d\epsilon d\varphi \right. \\
&\qquad \left. + \left(1-\mathbb{E}\left[\Pi_{j=1}^t \pi_j\right]\right) \iint \left[\underline{z}_1-z^*\right] \frac{h_0^*(z^*,\epsilon,\varphi)}{h_0(z^*)} d\epsilon d\varphi \right\} \\
&= h_0(z^*) \left\{ z^* + \mathbb{E}\left[\Delta z_1^*\right] - \mathbb{E}\left[\underline{z}_1\right] + \left(1-\mathbb{E}\left[\Pi_{j=1}^t \pi_j\right]\right) \left(\mathbb{E}\left[\underline{z}_1\right]-z^*\right) \right\} \\
&= h_0(z^*) \left\{ \mathbb{E}\left[\Delta z_1^*\right] - \mathbb{E}\left[\Pi_{j=1}^t \pi_j\right] \left(\mathbb{E}\left[\underline{z}_1\right]-z^*\right) \right\}, \qquad \text{(A.5)}
\end{aligned}$$

where now $h_0(z^*) = \iiint h_0^*(z^*,\epsilon,\varphi)\, g(\boldsymbol{\pi})\, d\epsilon d\varphi d\boldsymbol{\pi}$. In the second line, we have again made use of a constant $h_0^*(\cdot)$ and also the independence of $\boldsymbol{\pi}_i$. Normalized bunching at K_1 in period t will then be:

$$b_1^t = \mathbb{E}\left[\Delta z_1^*\right] - \mathbb{E}\left[\Pi_{j=1}^t \pi_j\right] \left(\mathbb{E}\left[\underline{z}_1\right]-z^*\right). \qquad \text{(A.6)}$$

Using similar steps, we can show that bunching in period $t > \mathcal{T}_1$ at K_2, when moving sequentially from K_1, can be written as:

$$\begin{aligned}
B_2^t &= \iiiint_{\underline{z}_1}^{z^*+\Delta z_2^*} h_0^*(\zeta,\epsilon,\varphi)\, g(\boldsymbol{\pi})\, d\zeta d\epsilon d\varphi d\boldsymbol{\pi} \\
&\quad + \iiiint_{z^*+\Delta z_2^*}^{\bar{z}_0} \left(\Pi_{j=1}^{t-\mathcal{T}_1} \pi_j\right) h_0^*(\zeta,\epsilon,\varphi)\, g(\boldsymbol{\pi})\, d\zeta d\epsilon d\varphi d\boldsymbol{\pi} \\
&\quad + \iiiint_{z^*}^{\bar{z}_1} \left(1-\Pi_{j=1}^{t-\mathcal{T}_1} \pi_j \cdot \Pi_{j=1}^{\mathcal{T}_1} \pi_j\right) h_0^*(\zeta,\epsilon,\varphi)\, g(\boldsymbol{\pi})\, d\zeta d\epsilon d\varphi d\boldsymbol{\pi} \\
&= h_0(z^*) \left\{ \left(1-\mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1} \pi_j\right]\right) \mathbb{E}\left[\Delta z_2^*\right] + \mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1} \pi_j\right] \mathbb{E}\left[\bar{z}_0\right] \right. \\
&\qquad \left. - \mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1} \pi_j \cdot \Pi_{j=1}^{\mathcal{T}_1} \pi_j\right] \mathbb{E}\left[\underline{z}_1\right] - \left(\mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1} \pi_j\right] - \mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1} \pi_j \cdot \Pi_{j=1}^{\mathcal{T}_1} \pi_j\right]\right) z^* \right\}. \qquad \text{(A.7)}
\end{aligned}$$

Likewise, normalized bunching at K_2 will be:

$$\begin{aligned} b_2^t &= \left(1 - \mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1}\pi_j\right]\right)\mathbb{E}\left[\Delta z_2^*\right] + \mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1}\pi_j\right]\mathbb{E}\left[\bar{z}_0\right] - \mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1}\pi_j \cdot \Pi_{j=1}^{\mathcal{T}_1}\pi_j\right]\mathbb{E}\left[\underline{z}_1\right] \\ &\quad - \left(\mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1}\pi_j\right] - \mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1}\pi_j \cdot \Pi_{j=1}^{\mathcal{T}_1}\pi_j\right]\right) z^*. \end{aligned} \tag{A.8}$$

The levels of bunching at the kink before and after the transition are now functions of average behavioral responses, $(\Delta z_1^*, \Delta z_2^*)$, the average thresholds for marginal bunchers, $(\underline{z}_1, \bar{z}_0)$, and average survival probabilities, $\left(\Pi_{j=1}^{t}\pi_j, \Pi_{j=1}^{t-\mathcal{T}_1}\pi_j \cdot \Pi_{j=1}^{\mathcal{T}_1}\pi_j\right)$. Relative to our baseline dynamic model in Section 6.4.1, the number of intermediate parameters to be identified is increasing in the number of post-transition periods, due to the terms of the form $\mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1}\pi_j \cdot \Pi_{j=1}^{\mathcal{T}_1}\pi_j\right]$. A sufficient condition that allows us to retain identification while only using two transitions in kinks is that the expectation of this product simplifies to a product of expectations: $\mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1}\pi_j \cdot \Pi_{j=1}^{\mathcal{T}_1}\pi_j\right] = \mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1}\pi_j\right]\mathbb{E}\left[\Pi_{j=1}^{\mathcal{T}_1}\pi_j\right]$. There are two cases of interest that satisfy this condition. First, if $\pi_j = 0$ for some $j < \mathcal{T}_1$, then $\Pi_{j=1}^{\mathcal{T}_1}\pi_j = 0$, and the condition holds. This empirically appears to be the case in our context: adjustment takes roughly two years, while $\mathcal{T}_1 \geq 3$ in our two main applications. Second, if there is no heterogeneity in $\boldsymbol{\pi}$ across agents, the condition also holds.

If we relax the assumption that $\mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1}\pi_j \cdot \Pi_{j=1}^{\mathcal{T}_1}\pi_j\right] = \mathbb{E}\left[\Pi_{j=1}^{t-\mathcal{T}_1}\pi_j\right]\mathbb{E}\left[\Pi_{j=1}^{\mathcal{T}_1}\pi_j\right]$, we will require additional transitions in kinks in order to achieve identification. Furthermore, if we relax the assumption that the profile $\boldsymbol{\pi}_i$ is independent of $(\varepsilon_i, \phi_i, a_i)$, identification is more complicated, as the expectations in the above expressions will then feature weights that vary with t. In that case, more parametric structure on the joint distribution of $(\varepsilon_i, \phi_i, a_i, \boldsymbol{\pi}_i)$ is needed to achieve identification. We discuss identification further in section A.3 of the Appendix.

A.2 Dynamic Model with Forward-Looking Behavior

We present in this appendix a version of the dynamic model in Section 6.4.1 in which we allow for forward-looking behavior. The key difference in implications is that in addition to a gradual, lagged response to policy changes, this version of the model also predicts anticipatory adjustment by agents when policy changes are anticipated in advance. We have essentially the same setting as in Section 6.4.1, except that we will alter three of the assumptions. First, in each period, an individual draws a cost of adjustment, $\tilde{\phi}_t$, from a discrete distribution, which takes a value of ϕ with probability π and a value of 0 with probability $1 - \pi$.[43] Second, individuals make decisions over a finite horizon, living until Period $\overline{\mathcal{T}}$. In period 0, the individuals face a linear tax schedule, $T_0(z) = \tau_0 z$, with marginal tax rate τ_0. In period 1, a kink, K_1, is introduced at the earnings level z^*. This tax schedule is implemented for $\mathcal{T}_1$ periods, after which the tax schedule features a smaller kink, K_2, at the earnings level z^*. The smaller kink is present until period $\mathcal{T}_2$, after which we return to the linear tax schedule, T_0. As before, the kink K_j, $j \in \{1, 2\}$, features a top marginal tax rate of τ_j for earnings above z^*.[44] Finally, in each period, individuals solve this maximization

[43] For expositional purposes, we constrain the probability of drawing a nonzero fixed costs to be π in all periods. Thus, the terms from Section 6.4.1 of the form $\prod \pi_j$ simplify to π^j in this appendix. All results go through with the more flexible distribution of adjustment costs in Section 6.4.1.

[44] In Section 6.4.1, we do not specify time $\mathcal{T}_2$, when the smaller kink, K_2, is removed, as it is not relevant to the case where individuals are not forward-looking.

problem:

$$\max_{(c_{a,t},z_{a,t})} v\left(c_{a,t},z_{a,t};a,z_{a,t-1}\right)+\delta V_{a,t+1}\left(z_{a,t},A_{a,t}\right), \tag{A.9}$$

where $v\left(c_{a,t},z_{a,t};a,z_{a,t-1}\right)\equiv u\left(c_{a,t},z_{a,t};a\right)-\tilde{\phi}_t\cdot\mathbf{1}\left(z_{a,t}\neq z_{a,t-1}\right)$, δ is the discount factor, and $V_{a,t+1}$ is the value function moving forward in Period $t+1$:

$$V_{a,t+1}\left(\zeta,A_{a,t}\right)=\mathbb{E}_{\phi}\left[\max_{(c_{a,t+1},z_{a,t+1})} v\left(c_{a,t+1},z_{a,t+1};a,\zeta\right)+\delta V_{a,t+2}\left(z_{a,t+1},A_{a,t+1}\right)\right]. \tag{A.10}$$

$V_{a,t+1}$ is a function of where the individual has chosen to earn in Period t and assets $A_{a,t}$. The expectation $\mathbb{E}_{\phi}\left[\cdot\right]$ is taken over the distribution of $\tilde{\phi}_t$. The intertemporal budget constraint is:

$$A_{a,t}=\left(1+r\right)\left(A_{a,t-1}+z_{a,t}-T\left(z_{a,t}\right)-c_{a,t}\right). \tag{A.11}$$

We assume that $\delta\left(1+r\right)=1$. Because individuals have quasilinear preferences, this implies that consumption can be set to disposable income in each period: $c_{a,t}=z_{a,t}-T\left(z_{a,t}\right)$. We therefore use the following shorthand:

$$\begin{aligned} u_a^j\left(z\right) &= u\left(z-T_j\left(z\right),z;a\right) \\ V_{a,t}\left(z\right) &= V_{a,t}\left(z,A_{a,t-1}\right) \end{aligned} \tag{A.12}$$

Next, we define two operators that measure the utility gain (or loss) following a discrete change in earnings:

$$\begin{aligned} \triangle u_a^j\left(z,z'\right) &= u_a^j\left(z\right)-u_a^j\left(z'\right) \\ \triangle V_{a,t}\left(z,z'\right) &= V_{a,t}\left(z\right)-V_{a,t}\left(z'\right) \end{aligned} \tag{A.13}$$

In each case above, the utility and utility differential depend on the tax schedule. We define z_a^j as the optimal level of earnings under a frictionless, static optimization problem, facing the tax schedule T_j. We will refer to the frictionless, *dynamic* optimum in any given period as $\tilde{z}_{a,t}$. This is the optimal level of earnings when there is a fixed cost of zero drawn in the *current* period, but a nonzero fixed cost may be drawn in future periods. We will also make a distinction between two types of earnings adjustments: *active* and *passive*. An *active* earnings adjustment takes place in the presence of a nonzero fixed cost, while a *passive* earnings adjustment takes place only when a fixed cost of zero is drawn. We solve the model recursively, beginning in the regime after time $\mathcal{T}_2$, when the smaller kink, K_2, has been removed, continuing with the solution while the kink K_2 is present between times $\mathcal{T}_1$ and $\mathcal{T}_2$, and finally considering the first regime when the kink K_1 is present between time period 1 and $\mathcal{T}_1$.[45]

A.2.1 Earnings between $\mathcal{T}_2$ and $\overline{T}$

We will now derive the value function $V_{a,\mathcal{T}_2+1}\left(z\right)$. We begin with the following result: If an individual with initial earnings z makes an active adjustment in period $t>\mathcal{T}_2+1$, then it

[45] Our recursive method can be extended to the case of multiple, successive kinks. The effect on bunching of a sequence of more kinks depends on the relative size of the successive kinks.

must be the case that

$$\frac{1-(\delta\pi)^{\mathcal{T}_1+1-t}}{1-\delta\pi}\triangle u_a^0\left(z_a^0,z\right)\geq\phi. \tag{A.14}$$

We demonstrate this result with a constructive proof, showing the result for periods $\overline{\mathcal{T}}$ and $\overline{\mathcal{T}}-1$. Because the tax schedule is constant throughout this terminal period, the frictionless, dynamic optimum is equal to the static optimum: $\tilde{z}_{a,t}=z_a^0$. First, consider an agent in period $\overline{\mathcal{T}}$, with initial earnings z, who is considering maintaining earnings at z or paying the fixed cost ϕ and making an active adjustment to z_a^0, the frictionless, dynamic optimum in period $\overline{\mathcal{T}}$. The agent will make the adjustment if:

$$\begin{aligned}\triangle u_a^0\left(z_a^0,z\right) &\geq \phi \\ &= \frac{1-\delta\pi}{1-\delta\pi}\phi.\end{aligned} \tag{A.15}$$

Rearranging terms, we have satisfied the inequality in (A.14).

Now consider agents in period $\overline{\mathcal{T}}-1$ with initial earnings z. There are two types, those who would make an active adjustment to z_a^0 in period $\overline{\mathcal{T}}$ if the earnings z are carried forward and those who would not. Consider those who would not. If the agent remains with earnings of z, then utility will be $u_a^0(z)+\delta V_{a,\overline{\mathcal{T}}}(z)=u_a^0(z)+\delta\left[\pi\left(u_a^0(z)\right)+(1-\pi)\,u_a^0(z_a^0)\right]$. If the agent actively adjusts to z_a^0, then utility will be $u_a^0(z_a^0)-\phi+\delta u_a^0(z_a^0)$. The agent will actively adjust in period $\overline{\mathcal{T}}-1$ if:

$$\begin{aligned}\triangle u_a^0\left(z_a^0,z\right) &\geq \frac{1}{1+\delta\pi}\phi \\ &= \frac{1-\delta\pi}{1-(\delta\pi)^2}\phi.\end{aligned} \tag{A.16}$$

Once again, rearranging terms confirms that (A.14) holds. Finally, consider agents who would actively adjust from z to z_a^0 if earnings level z is carried forward. In this case, the agent's utility when remaining at z is:

$$\begin{aligned}u_a^0(z)+\delta V_{a,\overline{\mathcal{T}}}(z) &= u_a^0(z)+\delta\left[\pi\left(u_a^0\left(z_a^0\right)-\phi\right)+(1-\pi)\,u_a^0\left(z_a^0\right)\right] \\ &= u_a^0(z)+\delta\left(u_a^0\left(z_a^0\right)-\pi\phi\right).\end{aligned} \tag{A.17}$$

Intuitively, the agent will receive the optimal level of utility in the next period, and with probability π the agent will have to pay the fixed cost to achieve it. Similarly, the agent's utility after actively adjusting to z_a^0 in period $\overline{\mathcal{T}}-1$ is $u_a^0(z_a^0)-\phi+\delta u_a^0(z_a^0)$. The agent will therefore adjust in period $\overline{\mathcal{T}}$ if:

$$\triangle u_a^0\left(z_a^0,z\right)\geq(1-\delta\pi)\,\phi. \tag{A.18}$$

However, we know from (A.15) that this already holds for the agent who actively adjusts in period $\overline{\mathcal{T}}$. Finally, note that (A.15) implies (A.16). It follows that in period $\overline{\mathcal{T}}-1$, adjustment implies (A.15). We can similarly show the result for earlier periods by considering separately: (a) those who would actively adjust in the current period, but not in any future period; and (b) those who would adjust in some future period. Both types will satisfy the key inequality.

As a corollary, note that if an individual with initial earnings z makes an active adjustment in period $t > \mathcal{T}_2 + 1$, then she will also find it optimal to do so in any period t', where $\mathcal{T}_2 < t' < t$. To see this, note that if (A.14) holds for t, then it also holds for $t' < t$. It follows that the agent would also actively adjust in period t'.

Now consider an agent who earns z in period $\mathcal{T}_2$. Note that our results above imply that any active adjustment that takes place after $\mathcal{T}_2$ will only happen in period $\mathcal{T}_2 + 1$. These agents will receive a stream of discounted payoffs of $u_a^0(z_a^0)$ for $\overline{\mathcal{T}} - \mathcal{T}_2$ periods, *i.e.* $\sum_{j=0}^{\overline{\mathcal{T}}-\mathcal{T}_2-1} \delta^j u_a^0(z_a^0) = \frac{1-\delta^{\overline{\mathcal{T}}-\mathcal{T}_2}}{1-\delta} u_a^0(z_a^0)$, and pay a fixed cost of ϕ in period $\mathcal{T}_2$ with probability π. Otherwise, an agent will adjust to the dynamic frictionless optimum z_a^0 only when a fixed cost of zero is drawn. In the latter case, the agent receives a payoff of $u_a^0(z)$ until a fixed cost of zero is drawn, after which, the agent receives $u_a^0(z_a^0)$. We can therefore derive the following value function:[46]

$$V_{a,\mathcal{T}_2+1}(z) = \begin{cases} \frac{1-\delta^{\overline{\mathcal{T}}-\mathcal{T}_2}}{1-\delta} u_a^0(z_a^0) - \pi\phi & \text{if } \frac{1-(\delta\pi)^{\overline{\mathcal{T}}-\mathcal{T}_2}}{1-\delta\pi} \triangle u_a^0(z_a^0, z) \geq \phi \\ \frac{1-\delta^{\overline{\mathcal{T}}-\mathcal{T}_2}}{1-\delta} u_a^0(z_a^0) - \pi \frac{1-(\delta\pi)^{\overline{\mathcal{T}}-\mathcal{T}_2}}{1-\delta\pi} \triangle u_a^0(z_a^0, z) & \text{otherwise} \end{cases}. \tag{A.19}$$

To gain some intuition for (A.14), note that the left side of (A.14) is the net present value of the stream of the utility differential once the agent adjusts from z to z_a^0. If this exceeds the up-front cost of adjustment, ϕ, then the agent actively adjusts. The discount factor for j periods in the future, however, is $(\delta\pi)^j$, instead of only δ^j. The reason is that current adjustment only affects future utility j periods from now if j consecutive nonzero fixed costs are drawn, which happens with probability π^j. To better understand our second result regarding the timing of active changes, note that if the gains from adjustment over $\overline{\mathcal{T}} - t$ periods exceed the up-front cost, then the agent should also be willing to adjust in period $t' < t$ and accrue $\overline{\mathcal{T}} - t'$ periods of this gain, for the same up-front cost of ϕ.

A.2.2 Earnings between $\mathcal{T}_1$ and $\mathcal{T}_2$

We now derive the value function $V_{a,\mathcal{T}_1+1}(z)$. In this case, the dynamic frictionless optimum in each period, $\tilde{z}_{a,t}$, is not constant. Intuitively, the agent trades off the gains from adjusting earnings in response to K_2 with the effect of this adjustment on the value function $V_{a,\mathcal{T}_2+1}$. In general, the optimum is defined as:

$$\tilde{z}_{a,t} = \arg \max_{z \in [z_a^2, z_a^0]} \frac{1-(\delta\pi)^{\mathcal{T}_2+1-t}}{1-\delta\pi} u_a^2(z) + \delta^{\mathcal{T}_2+1-t} \pi^{\mathcal{T}_2-t} V_{a,\mathcal{T}_2+1}(z). \tag{A.20}$$

We restrict the maximization to the interval $[z_a^2, z_a^0]$, since reducing earnings below z_a^2 or raising earnings above z_a^0 weakly reduces utility in any current and all future periods for $t > \mathcal{T}_1$. From (A.19), we know that $V_{a,\mathcal{T}_2+1}$ is continuous, and thus the solution in (A.20) exists.[47] We present two results analogous to those in Section A.2.1, without proof. The proofs, nearly identical to those in the previous section, are available upon request. First, if

[46] The expected utility for passive adjusters is constructed recursively, working backward from period $\overline{\mathcal{T}}$ to period $\mathcal{T}_2 + 1$.

[47] Technically, we can see from (A.19) that while the function V_{a,P_2+1} is continuous, it is kinked, which creates a nonconvexity. Thus, the solution in (A.20) may not always be single-valued. In such cases, we define $\tilde{z}_{a,t}$ as the lowest level of earnings that maximizes utility.

an individual with initial earnings z makes an active adjustment in period t, $\mathcal{T}_1 < t \leq \mathcal{T}_2$, then:

$$\frac{1-(\delta\pi)^{\mathcal{T}_2+1-t}}{1-\delta\pi}\triangle u_a^2\left(\tilde{z}_{a,t},z\right)+\delta^{\mathcal{T}_2+1-t}\pi^{\mathcal{T}_2-t}\triangle V_{a,\mathcal{T}_2+1}\left(\tilde{z}_{a,t},z\right)\geq\phi. \tag{A.21}$$

Furthermore, if an individual with initial earnings z makes an active adjustment in period t, $\mathcal{T}_1 < t \leq \mathcal{T}_2$, then she will also find it optimal to do so in any period t', where $\mathcal{T}_1 < t' < t$.

The condition in (A.21) differs from that in (A.14) because the effect of adjustment on the utility beyond period $\mathcal{T}_2$ is taken into account, in addition to the up-front cost of adjustment, ϕ. Any adjustment in this time interval, active or passive, will be to the dynamic, frictionless optimum for the current period, $\tilde{z}_{a,t}$. As before, (A.21) implies that all active adjustment occurring between $\mathcal{T}_1+1$ and $\mathcal{T}_2$ takes place in period $\mathcal{T}_2+1$. Those who adjust in period $\mathcal{T}_2+1$ will earn $\tilde{z}_{a,\mathcal{T}_1+1}$. Thereafter, they only adjust to $\tilde{z}_{a,t}$ when a fixed cost of zero is drawn. Likewise, those who only adjust passively earn $z_{a,\mathcal{T}_1}$ in period $\mathcal{T}_1+1$, and thereafter adjust to $\tilde{z}_{a,t}$ when a fixed cost of zero is drawn. We can therefore derive the following value function:

$$V_{a,\mathcal{T}_1+1}(z)=\begin{cases}\begin{array}{l}\sum_{j=0}^{\mathcal{T}_2-\mathcal{T}_1-1}\delta^j u_a^2\left(\tilde{z}_{a,\mathcal{T}_1+1+j}\right)+\delta^{\mathcal{T}_2-\mathcal{T}_1}\triangle V_{a,\mathcal{T}_2+1}\left(\tilde{z}_{a,\mathcal{T}_2}\right)\\ -\sum_{j=0}^{\mathcal{T}_2-\mathcal{T}_1-2}\frac{(\delta\pi)^{\mathcal{T}_2-\mathcal{T}_1}}{\pi^{j+1}}\triangle V_{a,\mathcal{T}_2+1}\left(\tilde{z}_{a,\mathcal{T}_1+2+j},\tilde{z}_{a,\mathcal{T}_1+1+j}\right)\\ -\sum_{j=0}^{\mathcal{T}_2-\mathcal{T}_1-2}\frac{1-(\delta\pi)^{\mathcal{T}_2-\mathcal{T}_1-1-j}}{1-\delta\pi}\delta^{j+1}\pi\triangle u_a^2\left(\tilde{z}_{a,\mathcal{T}_1+2+j},\tilde{z}_{a,\mathcal{T}_1+1+j}\right)\\ -\pi\phi\end{array} & \begin{array}{l}\text{if (A.21) is satisfied}\\ \text{when } t=\mathcal{T}_1+1\end{array}\\ \\ \begin{array}{l}\sum_{j=0}^{\mathcal{T}_2-\mathcal{T}_1-1}\delta^j u_a^2\left(\tilde{z}_{a,\mathcal{T}_1+1+j}\right)+\delta^{\mathcal{T}_2-\mathcal{T}_1}\triangle V_{a,\mathcal{T}_2+1}\left(\tilde{z}_{a,\mathcal{T}_2}\right)\\ -\sum_{j=0}^{\mathcal{T}_2-\mathcal{T}_1-2}\frac{(\delta\pi)^{\mathcal{T}_2-\mathcal{T}_1}}{\pi^{j+1}}\triangle V_{a,\mathcal{T}_2+1}\left(\tilde{z}_{a,\mathcal{T}_1+2+j},\tilde{z}_{a,\mathcal{T}_1+1+j}\right)\\ -\sum_{j=0}^{\mathcal{T}_2-\mathcal{T}_1-2}\frac{1-(\delta\pi)^{\mathcal{T}_2-\mathcal{T}_1-1-j}}{1-\delta\pi}\delta^{j+1}\pi\triangle u_a^2\left(\tilde{z}_{a,\mathcal{T}_1+2+j},\tilde{z}_{a,\mathcal{T}_1+1+j}\right)\\ -\pi\left\{\begin{array}{l}\sum_{j=0}^{\mathcal{T}_2-\mathcal{T}_1-1}(\delta\pi)^j\triangle u_a^2\left(\tilde{z}_{a,\mathcal{T}_1+1},z\right)\\ -\delta^{\mathcal{T}_2-\mathcal{T}_1}\pi^{\mathcal{T}_2+1-\mathcal{T}_1}\triangle V_{a,\mathcal{T}_2+1}\left(\tilde{z}_{a,\mathcal{T}_1+1},z\right)\end{array}\right.\end{array} & \text{otherwise}\end{cases} \tag{A.22}$$

The first case in (A.22) applies to those who actively adjust in period $\mathcal{T}_1+1$ and passively adjust thereafter. The first line is the utility that would accrue if a fixed cost of zero were drawn in each period. The next two lines represent the deviation from this stream of utility, due to nonzero fixed costs potentially drawn in periods $\mathcal{T}_1+1$ through $\mathcal{T}_2$. The final line represents the fixed cost that is paid in period $\mathcal{T}_1+1$ with probability π. The second case in (A.22) applies to those who only passively adjust. The first three lines remain the same. The final two lines represent a loss in utility attributed to fact that earnings in period $\mathcal{T}_1+1$ may not be $\tilde{z}_{a,T_1+1}$. Note that earnings in period $\mathcal{T}_1$ can only affect utility through this last

channel.

A.2.3 Earnings between Period 1 and $\mathcal{T}_1$

Earnings during the first period, when the kink K_1 is present, can be derived similarly. The dynamic, frictionless optimum is now defined as:

$$\tilde{z}_{a,t} = \arg \max_{z \in [z_a^1, z_a^0]} \frac{1-(\delta\pi)^{\mathcal{T}_1+1-t}}{1-\delta\pi} u_a^1(z) + \delta^{\mathcal{T}_1+1-t}\pi^{\mathcal{T}_1-t} V_{a,T_1+1}(z).^{48} \tag{A.23}$$

Similar to the other cases, if an individual with initial earnings z makes an active adjustment in period t, $0 < t \leq \mathcal{T}_1$, then it must be the case that

$$\frac{1-(\delta\pi)^{\mathcal{T}_1+1-t}}{1-\delta\pi} \triangle u_a^1(\tilde{z}_{a,t}, z) + \delta^{\mathcal{T}_1+1-t}\pi^{\mathcal{T}_1-t} \triangle V_{a,\mathcal{T}_1+1}(\tilde{z}_{a,t}, z) \geq \phi. \tag{A.24}$$

Furthermore, if an individual with initial earnings z makes an active adjustment in period t, $0 < t \leq T_1$, then she will also find it optimal to do so in any period t', where $0 < t' < t$. Again, this implies that all active adjustment will take place in period 1. Since individuals begin with earnings of z_a^0, we know that all active adjustment will be downward. Thereafter, it can be shown that $\tilde{z}_{a,t}$ is weakly increasing, and upward adjustment will occur passively.

A.2.4 Characterizing Bunching

Given these results, we can now derive expressions for excess mass at z^* analogous to (8) and (9). For notational convenience, we define $\mathcal{A}_j(z)$ as the set of individuals, a, with initial earnings z who actively adjust in period j. Again, denote B_1^t as bunching at K_1 in period $t \in [1, \mathcal{T}_1]$. We have the following generalized version of (8):

$$\begin{aligned} B_1^t = \int_{z^*}^{z^*+\triangle z_1^*} \Bigg[& \mathbf{1}\{\tilde{z}_{a,1} = z^*, a \in \mathcal{A}_1(\zeta)\} \\ & + \sum_{j=1}^{t} \left(1-\pi^j\right) \pi^{t-j} \mathbf{1}\{\sup\{l | \, l \leq t, \tilde{z}_{a,l} = z^*\} = j, a \notin \mathcal{A}_1(\zeta)\} \\ & - \sum_{j=1}^{t-1} \left(1-\pi^{t-j}\right) \mathbf{1}\{\sup\{l | \, l \leq t, \tilde{z}_{a,l} = z^*\} = j, a \in \mathcal{A}_1(\zeta)\} \Bigg] h_0(\zeta) \, d\zeta. \end{aligned} \tag{A.25}$$

We have partitioned the set of potential bunchers into three groups in (A.25). In the first line, we have the set of active bunchers in period 1. In the second line, we capture individuals who are passive bunchers, *i.e.* $a \notin \mathcal{A}_1(z_a^0)$. For $j \in [1, t-1]$, the indicator function selects the individual who has $\tilde{z}_{a,j} = z^*$ but $\tilde{z}_{a,j+1} \neq z^*$. Since $\tilde{z}_{a,t}$ is weakly increasing, the optimal earnings for this individual is z^* in periods 1 through $j-1$. The probability that the individual bunches by period j is $1-\pi^j$. Thereafter, the individual will de-bunch if a fixed cost of zero is drawn. The probability of only drawing nonzero fixed costs thereafter is

[48] Note, the objective function now features two potential nonconvexities. In cases where the solution is multi-valued, we again define $\tilde{z}_{a,t}$ as the lowest earnings level from the set of solutions.

π^{t-j}. For $j = t$, the indicator function selects agents for whom $\tilde{z}_{a,t} = z^*$. Their probability of passively bunching by period t is $1 - \pi^t$. The third line captures the outflow of active bunchers, for whom $\tilde{z}_{a,t}$ ceases to be z^* starting in period j. The probability of having drawn a nonzero fixed cost and de-bunching since period j is $1 - \pi^{t-j}$.

Equation (A.25) differs from (8) in three key ways. First, the set of active bunchers in period 1 is different, as can be seen by comparing (A.24) and the relevant condition for active bunchers in Section 6.4.1, $\triangle u_a^1(z^*, z_a^0) \geq \phi$. The utility gain accrues for multiple periods in the forward-looking case, increasing the probability of actively bunching, but the effect of adjustment on future payoffs via $V_{a,\mathcal{T}_1+1}$ may either reinforce or offset this incentive. Furthermore, passive bunchers are (weakly) less likely to remain bunching, as they de-bunch in anticipation of policy changes in future periods. To see this, note that the π^{t-j} factor is decreasing in t. Finally, the set of active bunchers similarly de-bunch passively, in anticipation of future policy changes. The model therefore predicts a gradual outflow from the set of bunchers, in anticipation of the shift from K_1 to K_2. Nonetheless, the overall net change in bunching over time is ambiguous.

We now turn to bunching starting in period $\mathcal{T} + 1$. It can be shown, similarly to the cases above, that if an agent would be willing to actively bunch in period $\mathcal{T}_1 + 1$, she will also be willing to actively bunch in earlier periods. Thus, the only active adjustment occurring that affects bunching will be de-bunching. The set of individuals who actively de-bunch, $\mathcal{A}_{\mathcal{T}_1+1}(z^*)$, are those for whom (A.21) is satisfied, when evaluated at $t = \mathcal{T}_1 + 1$ and $z = z^*$. The remaining changes in bunching between $\mathcal{T}_1$ and $\mathcal{T}_2$ consist of passive adjustment among those who were bunching at the end of period $\mathcal{T}_1$. We can thus characterize B_2^t, bunching at K_2 in period $t \in \left[\mathcal{T}_1 + 1, \overline{\mathcal{T}}\right]$, in a manner analogous to (9):[49]

$$
\begin{aligned}
B_2^t \;=\; & \int_{z^*}^{z^*+\triangle z_1^*} \Bigg[\mathbf{1}\left\{a \notin \mathcal{A}_{\mathcal{T}_1+1}(z^*)\right\} \\
& \times \left\{ \pi^{t-\mathcal{T}_1} \mathbf{1}\left\{\tilde{z}_{a,\mathcal{T}_1+1} \neq z^*\right\} + \sum_{j=\mathcal{T}_1+1}^{t} \pi^{t-j} \mathbf{1}\left\{\sup\left\{l \middle| \, l \leq t, \tilde{z}_{a,l} = z^*\right\} = j\right\} \right\} \\
& \times \Bigg\{ \mathbf{1}\left\{\tilde{z}_{a,1} = z^*, a \in \mathcal{A}_1(\zeta)\right\} \\
& + \sum_{j=1}^{\mathcal{T}_1} \left(1 - \pi^j\right) \pi^{\mathcal{T}_1 - j} \mathbf{1}\left\{\sup\left\{l \middle| \, l \leq \mathcal{T}_1, \tilde{z}_{a,l} = z^*\right\} = j, a \notin \mathcal{A}_1(\zeta)\right\} \\
& - \sum_{j=1}^{\mathcal{T}_1 - 1} \left(1 - \pi^{\mathcal{T}_1 - j}\right) \mathbf{1}\left\{\sup\left\{l \middle| \, l \leq \mathcal{T}_1, \tilde{z}_{a,l} = z^*\right\} = j, a \in \mathcal{A}_1(\zeta)\right\} \Bigg\} \Bigg] h_0(\zeta) \, d\zeta.
\end{aligned}
\tag{A.26}
$$

The first line of this expression selects only those agents who do not actively de-bunch immediately in period $\mathcal{T}_1 + 1$. The second line selects the set of agents who would like to passively de-bunch beginning at some period $j > \mathcal{T}_1 + 1$. They are weighted by the probability

[49] When $\mathcal{T}_1 = 1$, we set the very last summation to zero.

of continuing to bunch due to consecutive draws of nonzero fixed costs. The final three lines select agents from the set of bunchers at the end of period T_1. As with our simpler model in Section 6.4.1, bunching gradually decreases following a reduction in the size of the kink from K_1 to K_2. However, in this case, the reduction is due to both fixed costs of adjustment and anticipation of the removal of the kink K_2 in period $\mathcal{T}_2 + 1$.

As in Section 6.4.1, the richer model in this appendix nests the dynamic model without forward looking behavior when we set $\delta = 0$, collapses to the comparative static model of Sections 6.2-6.3 if we additionally assume that $\pi = 1$ and is equivalent to the frictionless model when either $\phi = 0$ or $\pi = 0$.

A.3 Identification

Our estimator is a minimum distance estimator (MDE); Newey and McFadden (1994) give conditions for identification, consistency, and asymptotic normality. An MDE is defined as:

$$\begin{aligned} \hat{\theta} &= \arg\min_{\theta} \hat{Q}(\theta) \\ \hat{Q}(\theta) &= [B - m(\theta)]' \hat{W} [B - m(\theta)] \end{aligned}$$

In our case, B is a vector of L estimated bunching amounts from before and after a policy change, and $m(\theta)$ is a vector of predicted bunching amounts. $\hat{W}$ is a weighting matrix. We consider our comparative static, and dynamic, models, in turn.

A.3.1 Comparative Static Model

We focus on the exactly identified case with two bunching moments, which is relevant in our empirical application of the comparative static model. We have:

$$\begin{aligned} m(\theta) &= (B_1(\varepsilon, \phi), \tilde{B}_2(\varepsilon, \phi)) \\ B_1 &= \int_{\underline{z}_1}^{z^* + \Delta z_1^*} h(\xi) d\xi \\ \tilde{B}_2 &= \int_{\underline{z}_1}^{\bar{z}_0} h(\xi) d\xi \end{aligned}$$

where B_1 and $\tilde{B}_2$ refer to bunching before and after the policy change, and $\theta \equiv (\varepsilon, \phi)$.

The upper cutoff in B_1 is defined as

$$z^* + \Delta z_1^* = z^* \left(\frac{1 - \tau_0}{1 - \tau_1} \right)^{\varepsilon}.$$

A necessary condition for identification is that solutions for $\underline{z}_1$ and $\bar{z}_0$ exist; if they do not, then no bunching occurs. It is straightforward to show that a solution for $\underline{z}_1$ exists if

$$z^* \left[(1 - \tau_1) - \left(\frac{1 - \tau_0}{1 - \tau_1} \right)^{\varepsilon} \left((1 - \tau_1) - \varepsilon (\tau_1 - \tau_0) \right) \right] > \phi (\varepsilon + 1).$$

This ensures that the "top" buncher wants to adjust to the kink. A solution for $\bar{z}_0$ exists as long as some debunching occurs. It is straightforward to show that this requires that:

$$z^* \left[\frac{(1-\tau_2)^{\varepsilon+1} - (1-\tau_1)^{\varepsilon+1}}{(1-\tau_1)^{\varepsilon}} \right] > \phi(\varepsilon+1).$$

As long as $\tau_0 < \tau_2 < \tau_1$, $\varepsilon > 0$, and $\phi > 0$, there exists a range of values of ε and ϕ for which these inequalities hold.

Provided that $\bar{z}_0$ and $\underline{z}_1$ exist, identification requires that $m(\theta) = B$ has a unique solution. Following previous literature (*e.g.* Kline and Walters 2016), we establish local uniqueness by linearizing $m(\cdot)$ around a solution $m(\theta_0) = B$. Let θ_0 be a solution to $m(\theta) = B$. Linearizing $m(\cdot)$ around θ_0, we have:

$$m(\theta) \approx m(\theta_0) + \nabla m(\theta_0)(\theta - \theta_0).$$

It follows that a unique solution requires $\mathbf{J_m}(\theta_0)$ to have full rank, where $\mathbf{J_m}(\theta_0)$ is the Jacobian of $m(\cdot)$ evaluated at θ_0:

$$\mathbf{J_m}(\theta_0) = \begin{bmatrix} \frac{\partial B_1}{\partial \varepsilon} & \frac{\partial B_1}{\partial \phi} \\ \frac{\partial \tilde{B}_2}{\partial \varepsilon} & \frac{\partial \tilde{B}_2}{\partial \phi} \end{bmatrix}.$$

We calculate the elements of this matrix analytically by differentiating the expressions above for B_1 and $\tilde{B}_2$, which is straightforward.[50] Thus, given $\hat{\theta}$, $\underline{z}_1$, and $\bar{z}_0$, we can calculate the Jacobian analytically (although $\underline{z}_1$ and $\bar{z}_0$ must be found numerically).

$\mathbf{J_m}$ has full rank only if it has a non-zero determinant. We find in all of our bootstrap iterations that $\det(\mathbf{J_m}) < 0$, demonstrating that the determinant is significantly different from zero. We have also shown analytically that the determinant is generically non-zero (results available upon request).

A.3.2 Dynamic Model

To identify the dynamic model, we need to observe at least as many moments as the number of parameters we seek to estimate. In our case this means that we must observe bunching across multiple policy changes, specifically the reductions in the BRR above the exempt amount in 1990 and at age 70. Let l index different such policy changes (in our case, $l \in \{1990, 70\}$). Let $B^t_{1,l}$ be bunching at kink l and period t *before* the policy change, let $B^t_{2,l}$ be bunching at kink l and period t *after* the policy change, let time t measure the time since the introduction of the first kink, $K_{1,l}$, and let the policy change at kink l take place at time $\mathcal{T}_{1,l}$. The parameter vector θ now consists of $(\varepsilon, \phi, \pi_1, \pi_{2,}..., \pi_5)$. We match 12 bunching amounts in our estimates: 1987 to 1992 (pooling 66 to 68 year olds) and ages 67 to 72 (pooling years 1990 to 1999).

Bunching before the policy change is

$$B^t_{1,l} = \Pi^t_{j=1}\pi_j \cdot B_{1,l} + (1 - \Pi^t_{j=1}\pi_j) B^*_{1,l}$$

[50] We can specify functions implicitly defining the lower and upper cutoffs $\underline{z}_1$ and $\bar{z}_0$, respectively, as functions of the other parameters, given our quasilinear and isoelastic case. These enter the expressions for each element of the Jacobian (more details are available upon request).

where $B_{1,l} = \int_{\underline{z}_{1,l}}^{z_l^*+\Delta z_{1,l}^*} h(\xi)d\xi$ and $B_{1,l}^* = \int_{z_l^*}^{z_l^*+\Delta z_{1,l}^*} h(\xi)d\xi$, and the limits of integration are defined similarly to the static case (but with the additional subscript l to allow for analysis across multiple policy changes, as in our empirical application of the dynamic model). If the policy change happens $\mathcal{T}_{1,l}$ periods after the kink is initially introduced, then bunching under the new policy in period t is

$$B_{2,l}^t = \Pi_{j=1}^{t-\mathcal{T}_{1,l}}\pi_j \cdot \tilde{B}_{2,l} + \left(1 - \Pi_{j=1}^{t-\mathcal{T}_{1,l}}\pi_j\right) B_{2,l}^* + \Pi_{j=1}^{t-\mathcal{T}_{1,l}}\pi_j \left(1 - \Pi_{j=1}^{\mathcal{T}_{1,l}}\pi_j\right)\left(B_{1,l}^* - B_{1,l}\right)$$

where $\tilde{B}_{2,l} = \int_{\underline{z}_{1,l}}^{\bar{z}_{0,l}} h(\xi)d\xi$, $B_{2,l}^* = \int_{z_l^*}^{z_l^*+\Delta z_{2,l}^*} h(\xi)d\xi$, and the limits of integration again are defined similarly to the static case but with the additional subscript l.

We calculate the elements of the resulting Jacobian analytically by differentiating the expressions above for $B_{1,l}^t$ and $B_{2,l}^t$ with respect to ε, ϕ, π_1, π_2, π_3, π_4, and π_5, which is again straightforward. Thus, given $\hat{\theta}$, $\underline{z}_{1,l}$ and $\bar{z}_{0,l}$, we can again calculate the Jacobian analytically.

Identification requires that this Jacobian have full rank. To test for full rank of the Jacobian, we use the method of Kleibergen and Papp (2006). We use the bootstrap to obtain an estimate of $Var[\mathbf{J_m}(\hat{\theta})]$. In each iteration of our bootstrap, we also calculate $\mathbf{J_m}(\hat{\theta})$, and we estimate $Var[\mathbf{J_m}(\hat{\theta})]$ from the bootstrap variance-covariance matrix. The RK test easily rejects under-identification, with $p < 0.001$.

A.4 Explaining Bunching Below and Above the Kink

Figure 2 shows an intriguing pattern: bunching appears asymmetric around the exempt amount. In particular, it is evident that more excess mass appears just below the exempt amount, relative to just above it. We show in this appendix that our current model can explain this pattern. In particular, the observed pattern can result simply from a downward-sloping counterfactual density and symmetric noise in realized earnings, relative to desired earnings.

We demonstrate this in a simple model without adjustment costs or heterogeneity. In the upper-left panel of Appendix Figure B.5, we show the earnings density prior to the introduction of a kink and partition earners into three groups: A, B, and C. In the bottom-left panel, the kink has been introduced. Group A, who earns below z^*, does not respond. Group B is the set of bunchers, who locate near z^*. Finally, group C is comprised of earners who reduce their earnings in response to the kink, but do not bunch at z^*. To match the data, we assume that bunching is diffuse around z^*, rather than occurring only at z^*—*i.e.* realized earnings are equal to desired earnings plus some random, mean-zero noise (Friedberg, 2000). As can be seen, the downward slope of the initial density causes the ex-post density, net of bunchers, to be much higher just below the exempt amount, relative to just above the it. As a result, when the set of bunchers create excess mass near z^*, it appears that more excess mass is located just below the exempt amount, relative to just above it.

We confirm this in the right panel of Appendix Figure B.5, where we have plotted a counterfactual density and a simulated earnings density. We use the age 72 earnings density as the counterfactual density under a linear budget set, as in our main results. Next, we simulate the density that would be predicted at age 69 with the baseline elasticity of 0.36 from our dynamic model. We ignore the adjustment cost under the assumption—for illustrative purposes—that we are operating in a steady state in which enough time has elapsed for full

adjustment to take place, which is plausible by age 69. Finally, we assume, for illustrative purposes, that relative to desired earnings, a normally distributed, mean-zero error is added, with a standard deviation equal to \$1,800, *i.e.* half of the baseline excluded region above and below the exempt amount. This ensures that among those intending to locate at the exempt amount, over 99 percent locate within the baseline excluded region of \$3,600 on either side of the exempt amount.[51]

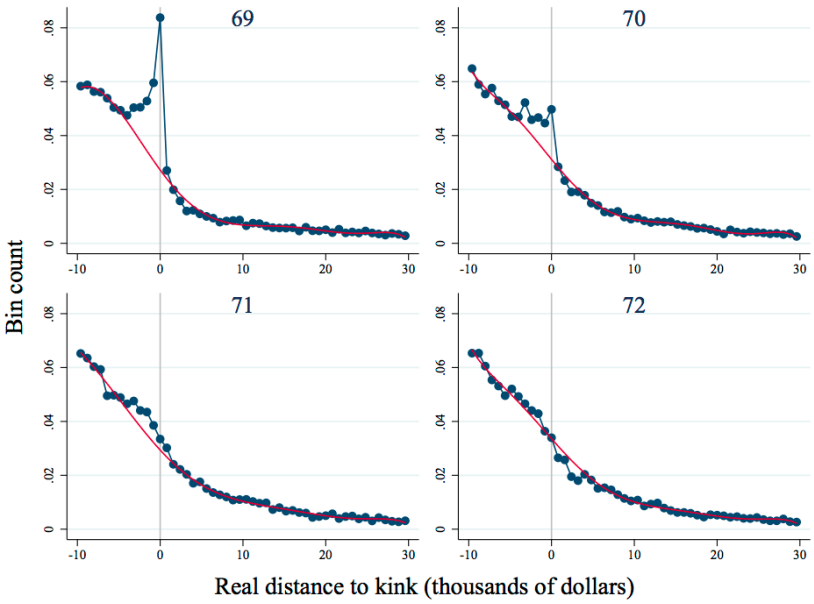
It is evident in Appendix Figure B.5 that more excess mass appears just below the exempt amount than just above it, consistent with the observed pattern at age 69 in Figure 2. Indeed, we can calculate measures of excess mass in the regions below and above the exempt amount, respectively. We do so by estimating excess mass, limiting the sample only to observations below and above the exempt amount, respectively. This calculation on the actual density for age 69 shows that normalized excess mass below the exempt amount accounts for 63 percent of the total amount of excess mass (in the bins centered below, at, and above the exempt amount). This is insignificantly different from the 55 percent predicted in the simulation. Note that in both the actual and simulated distributions, the percent bunching in the bin centered *at* the kink is substantial and includes some individuals who have earnings below the kink (but still within the central bin's width).

These qualitative conclusions are robust to alternative assumptions about the parameters (results available upon request). Thus, we can reproduce the patterns of bunching below and above the kink using only the fact that the counterfactual density is downward sloping, and following previous literature in adding noise to realized earnings, relative to desired earnings.

[51] The results are comparable throughout a range of assumptions about this standard deviation.

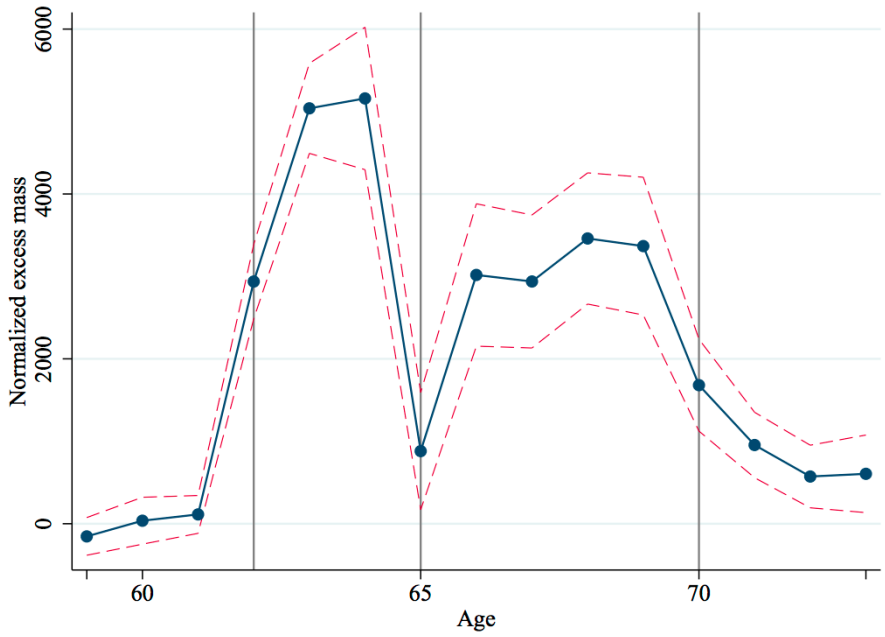
B Appendix: Additional Empirical Results (for online publication)

Figure B.1: Normalized Excess Mass of Claimants, Ages 69 to 72, 1983 to 1999



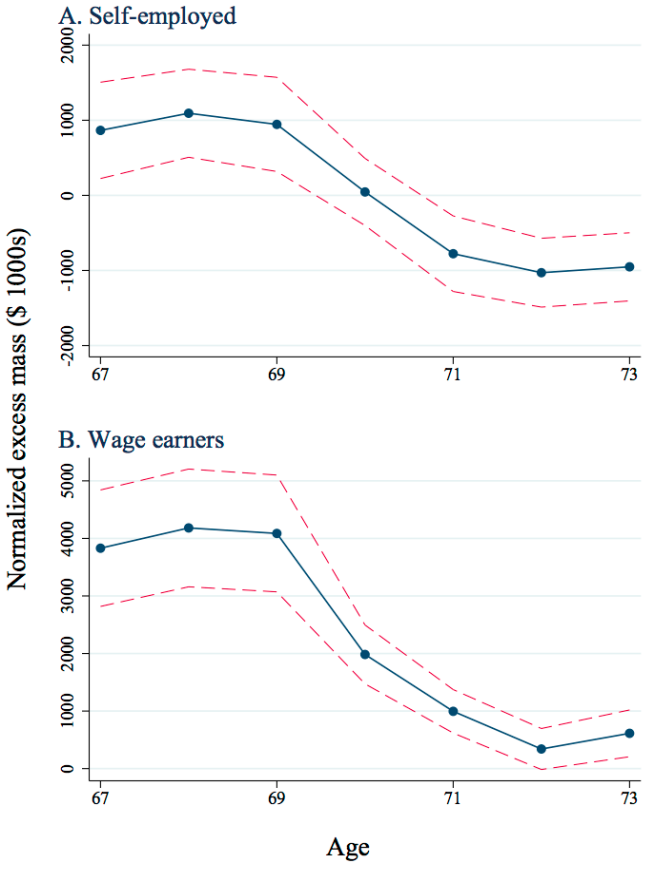
Note: See notes from Figure 2. This figure differs from Figure 2 because here we pool 1983 to 1999 to gain extra statistical power. The continued bunching at age 71 is more evident. In the main sample, we pool only 1990 to 1999 because the BRR was constant over this period, avoiding issues relating to the transition to a lower BRR in 1990.

Figure B.2: Normalized Excess Mass of Claimants, Ages 59 to 73, 1990 to 1999



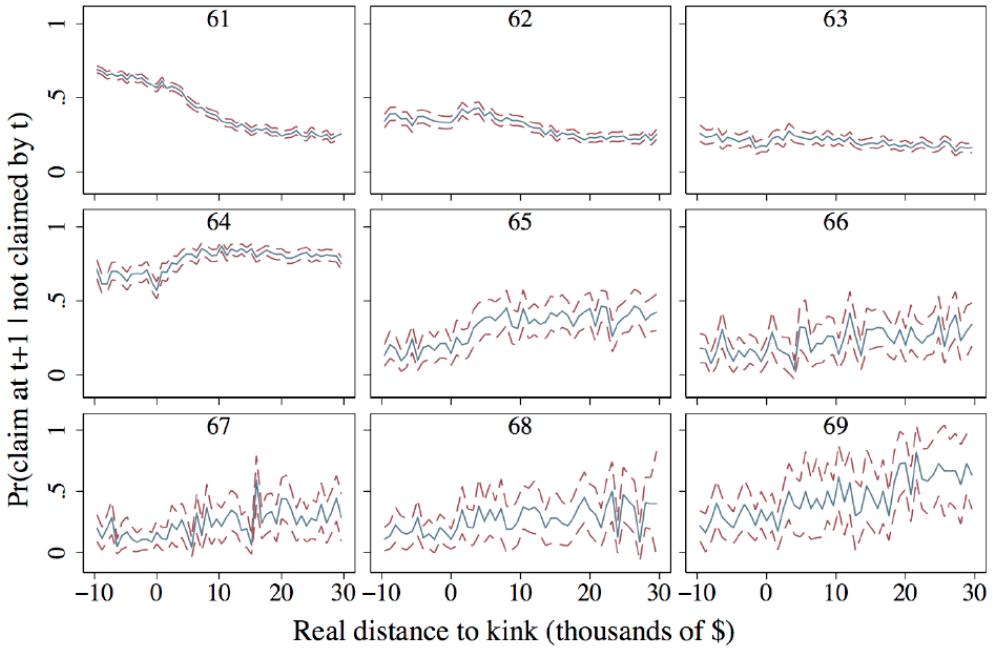
Note: See notes to Figure 3. This figure differs from Figure 3 because here the sample in year t consists only of people who have claimed OASI in year t or before (whereas in Figure 3 it consists of those who claimed by age 65).

Figure B.3: Excess Normalized Mass Among Self-Employed and Wage Earners, 1983-1999



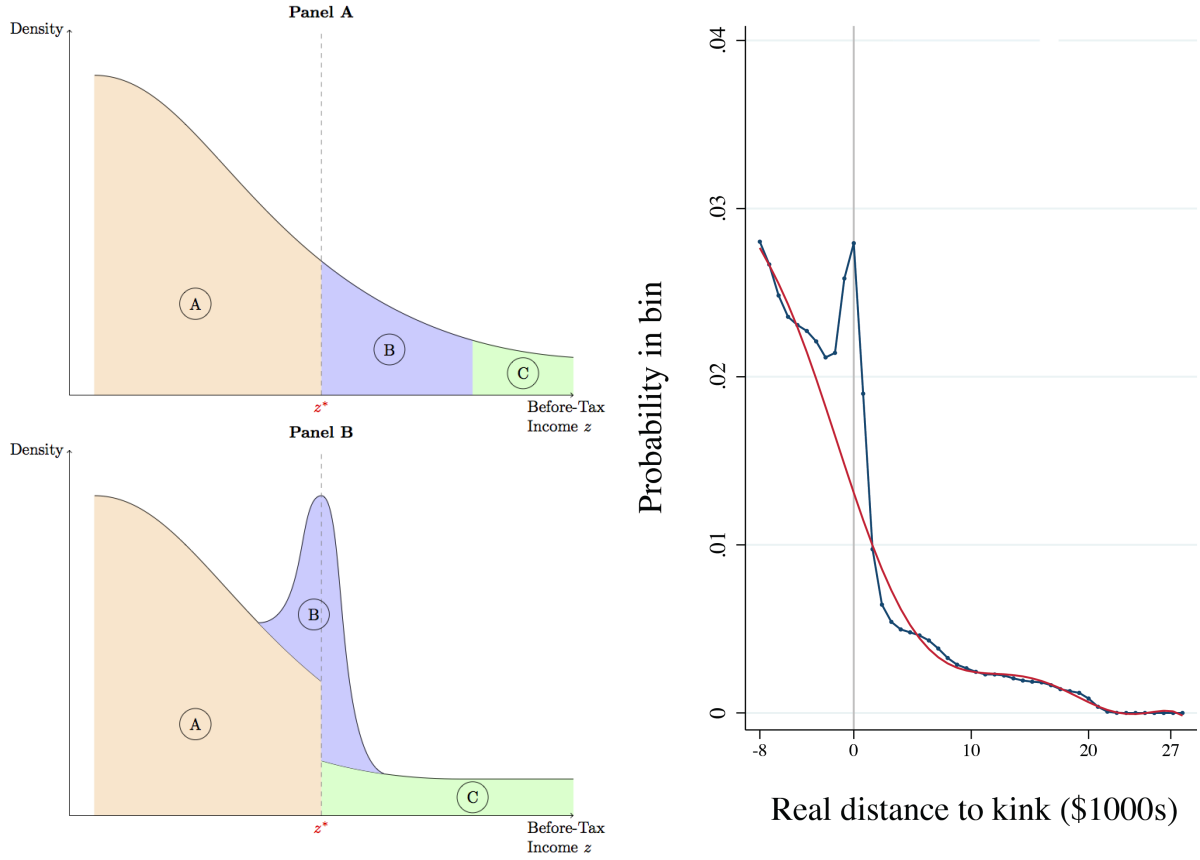
Notes: The figure shows normalized excess mass by age among those with self-employment income (Panel A) and "wage earners" (Panel B). A claimant is considered "self-employed" in year t if s/he has any self-employment income (as measured on tax form 1040 Schedule SE). "Wage earners" are defined as those who do not have self-employment income. The figure shows that at ages 70 and above, bunching dissipates more quickly for the self-employed than for wage earners—indeed bunching is immediately indistinguishable from zero at age 70 among the self-employed. To maximize our power given the relatively small number of self-employed, these results pool the years 1983 to 1999. Our main results pool only 1990 to 1999 since the BRR was constant over these years; when we examine wage earners and the self-employed over only 1990 to 1999, the results are qualitatively comparable but notably lose statistical power among the self-employed. See also notes from Figure 2.

Figure B.4: Probability of claiming OASI in year t+1 among 61-68 year-olds in year t who are not claiming, 1990-1998



Note: The figure shows the probabiliy that an individual claims OASI in year $t+1$, conditional on not claiming OASI in year t, for those ages 61-68 in year t from 1990 to 1998.

Figure B.5: Simulated earnings distribution at age 69



Note: The left-hand side shows how a downward sloping, counterfactual density can lead to apparently asymmetric bunching when bunching is diffuse. The right-hand side shows a simulated distribution of earnings at age 69, given the estimated model parameters. The y-axis shows the simulated density of earnings in each bin, and the x-axis shows the distance to the exempt amount. The figure demonstrates that we simulate more excess mass below the exempt amount than above it, consistent with the empirical distribution of earnings at age 69 (and other ages) shown in Figure 2. See Appendix A.4 for details.

Table B.1: Robustness of normalized bunching to alternative birth month restrictions

	b_{68}	b_{69}	b_{70}	b_{71}	b_{72}
A) Born January-March	3545.4	4036.2	565	881.7	-236.8
	[2681.1, 4409.7]***	[2934.2, 5138.2]***	[42, 1088.1]**	[-14.3, 1777.6]*	[-890.4, 416.8]
B) Born any month	3992.2	3552.3	1203.9	941.4	-231.4
	[3386.8, 4597.7]***	[3092.4, 4012.2]***	[929, 1478.9]***	[453, 1429.8]***	[-510.7, 47.8]

Notes: The table shows excess normalized bunching and its confidence interval at each age from 68 to 72 for two samples: those born January to March (Row A), and those born in any month (Row B). The data are pooled over the period from 1983-1999. The table shows that we continue to estimate significant bunching at age 70 (and in some cases 71) when the sample is restricted to those born in January to March. Limiting the sample only to those born in January yields insignificant results, with little statistical power. *** indicates $p<0.01$;**$p<0.05$; * $p<0.10$.

Table B.2: Robustness to alternative empirical choices

Binsize	Degree	Excluded Bins	b_{68}	b_{69}	b_{70}	b_{71}	b_{72}
			Panel A: Baseline				
\$800	7	4	3442.3	2868.4	657.9	1068.8	70.1
			[2763.5, 4121.2]***	[2763.5, 4121.2]***	[195.5, 1120.4]***	[527.2, 1610.4]***	[-328.5, 468.7]
			Panel B: Robustness to binsize				
\$400	7	8	3107.8	2606.3	462.2	923.0	-55.3
			[2653.6, 3561.9]***	[2090.4, 3122.2]***	[111.1, 813.2]***	[541.5, 1304.4]***	[-391.4, 280.7]
\$1,600	7	2	3047.0	2941.0	601.4	1210.9	241.2
			[2362.4, 3731.7]***	[2458.5, 3423.5]***	[48.2, 1154.5]**	[581.5, 1840.3]***	[-363.6, 846.1]
			Panel C: Robustness to degree				
\$800	6	4	3677.2	3267.3	1310.6	993.7	224.6
			[3117.4, 4237.0]***	[2810.5, 3724.0]***	[853.3, 1767.9]***	[527.8, 1459.6]***	[-162.0, 611.2]
\$800	8	4	3535.1	2948.0	710.8	1084.3	82.4
			[2944.7, 4125.6]***	[2529.1, 3366.9]***	[296.3, 1125.2]***	[554.5, 1614.1]***	[-393.4, 558.2]
			Panel D: Robustness to excluded region				
\$800	7	3	2170.2	2182.4	202.2	191.2	-55.0
			[1605.1, 2735.4]***	[1697.1, 2667.7]***	[-126.9, 531.4]	[-243.9, 626.2]	[-390.8, 280.8]
\$800	7	5	3610.6	2651.3	298.1	1103.3	579.9
			[2672.5, 4548.6]***	[1972.2, 3330.4]***	[-304.8, 901.0]	[229.6, 1977.1]**	[-161.8, 1321.6]

Notes: The table shows the estimated bunching amount at each age from 68 to 72, varying the bin size, degree of the polynomial of the smooth density, or number of excluded bins around the exempt amount. Note that varying the bin size but fixing the number of excluded bins automatically changes the width of the excluded region, so to (approximately) fix the width of the excluded region when changing the bin size, we also change the number of excluded bins. *** indicates $p<0.01$;**$p<0.05$; * $p<0.10$.

Table B.3: Heterogeneity in Estimates of Elasticity and Adjustment Cost across Samples

	(1)	(2)	(3)	(4)
	ε	p-value for ε equality	ϕ	p-value for ϕ equality
Men	0.44	0.39	\$62	0.00
	[0.38, 0.52]***		[14, 167]***	
Women	0.42		\$489	
	[0.32, 0.50]***		[165, 720]***	
High lifetime earnings	0.48	0.05	\$24	0.00
	[0.41, 0.58]***		[2, 90]***	
Low lifetime earnings	0.44		\$538	
	[0.32, 0.51]***		[217, 688]***	
High lifetime earnings variability	0.39	0.25	\$116	0.16
	[0.35, 0.46]***		[37, 315]***	
Low lifetime earnings variability	0.38		\$178	
	[0.33, 0.46]***		[55, 378]***	

Notes: This table implements our "comparative static" method separately in each of several groups shown in each row. "High/low lifetime earnings" refers to the group of individuals with mean real earnings from 1951 (when the data begin) to 1989 that are above/below the median level in our study population. "High/low lifetime earnings variability" refers to the group of individuals for whom the standard deviation of real earnings from 1951 to 1989 is above/below the median level in our study population. Columns 2 and 4 show the p-values for the two-sided test of equality in the estimates between each set of groups (*i.e.* men *vs.* women, high *vs.* low lifetime earnings, and high *vs.* low earnings variability), for ε and ϕ, respectively. We pool data from two policy changes: (a) around the 1989/1990 transition analyzed in Table 2, and (b) around the age 69/70 transition analyzed in Table 4. We pool the transitions because this gives us the maximum power to detect differences across groups. The results are generally comparable when we investigate each transition separately. See also notes from Tables 2 and 4.