ANNUAL REVIEWS Further
Click here to view this article's online features:
- Download figures as PPT slides
- Navigate linked references
- Download citations
- Explore related articles
- Search keywords

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

# Bunching

Henrik Jacobsen Kleven

Department of Economics, London School of Economics, London WC2A 2AE, United Kingdom; email: h.j.kleven@lse.ac.uk

Annu. Rev. Econ. 2016. 8:435–64

First published online as a Review in Advance on September 14, 2016

The *Annual Review of Economics* is online at economics.annualreviews.org

This article's doi: 10.1146/annurev-economics-080315-015234

Copyright © 2016 by Annual Reviews. All rights reserved

JEL codes: H00, J01, C01

## Keywords

bunching estimation, kinks, notches, structural parameters, optimization frictions, reference dependence

## Abstract

Recent years have seen a surge of applied work using bunching approaches, a development that is closely linked to the increased availability of administrative data. These approaches exploit the incentives for bunching created by discontinuities in the slope of choice sets (kinks) or in the level of choice sets (notches) to study the behavior of individuals and firms. Although the bunching approach was originally developed in the context of taxation, it is beginning to find applications in many other areas, such as social security, social insurance, welfare programs, education, regulation, private sector prices, and reference-dependent preferences. This review provides a guide to bunching estimation, discusses its strengths and weaknesses, surveys a range of applications across fields, and considers reasons for the ubiquity of kinks and notches.

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

## 1. INTRODUCTION

Recent years have seen the development of a new empirical approach in economics: the bunching approach. This approach uses bunching around points that feature discontinuities in incentives to elicit behavioral responses and estimate structural parameters. The approach was initially developed to estimate behavioral responses to taxes and transfers, but is now finding applications in other areas and settings. This review provides a guide to bunching estimation, discusses its strengths and weaknesses, draws links to other literatures, and ponders directions for future research.

The literature distinguishes between two conceptually different bunching designs. One type of design is based on kink points—discrete changes in the slope of choice sets—and was developed by Saez (2010) and Chetty et al. (2011). The other type of design is based on notch points—discrete changes in the level of choice sets—and was developed by Kleven & Waseem (2013). In the context of taxes and transfers, the distinction corresponds to whether the discontinuity occurs in the marginal tax rate or in the average tax rate. Kinks and notches offer different empirical advantages and challenges, as discussed below, and they tend to feature in different types of settings. Although kinks are commonly observed in income redistribution policies (such as graduated income tax systems), notches are ubiquitous across a wide range of other tax and nontax settings.

The emergence of the bunching approach is closely linked to another recent development in applied research: the increased use of administrative data. Because of the local nature of bunching responses—moving to specific points from nearby regions—estimating bunching precisely requires large data sets with very little measurement error. We rarely see any bunching in survey data due to small sample sizes and measurement error. With access to big administrative data sets, conversely, simply plotting the raw data can often reveal bunching and provides prima facie evidence of a causal effect of the incentive in question. A key question, however, is what we can learn from such responses in terms of structural and more externally valid parameters.

I argue that two broad lessons have emerged from the bunching literature to date. First, although bunching provides compelling nonparametric evidence of a behavioral response, moving from observed bunching to a structural parameter that can be used to predict the effects of policy changes is difficult. This is particularly true in the context of labor supply—the context for which the bunching approach was initially developed—due to a range of optimization frictions that attenuate bunching and are difficult to observe and model.[1] These frictions include aspects such as hours constraints, search costs, inattention, and uncertainty. Such frictions imply that any evidence of sharp bunching in earnings likely results from tax evasion or tax avoidance rather than real labor supply responses. Indeed, several applications of the bunching approach explicitly consider evasion and avoidance as their main objects of interest. Second, these difficulties of estimating structural elasticities do not invalidate the bunching approach, but they imply that the approach may be better used in different ways than initially intended. This includes studying different outcomes than labor supply (some that are less subject to optimization friction), and it includes using bunching for other purposes than to obtain price elasticities for policy prediction. I provide many examples of such alternative uses of the bunching approach below.

The bunching literature is tied to an earlier literature estimating labor supply in the presence of kinked budget sets, namely the nonlinear budget set approach pioneered by Burtless & Hausman (1978) and Hausman (1981). This literature estimated labor supply using models that predict bunching at kink points even though no bunching was found in the survey data they used, an issue that was debated by Heckman (1983) and Hausman (1983). The way that theory and data were reconciled in those studies was by allowing for measurement error in the data and optimization

[1] As argued below, this limitation is stronger for kinks than for notches, but it does apply to both.

error by households through the modeling of the error term. Although access to administrative data largely resolves the problem of measurement error, it does not reduce the scope for optimization error in shaping bunching, as highlighted above. The study of optimization frictions in the recent bunching literature is closely related to the debate about how to model the error term in the nonlinear budget set approach.

Bunching designs are related to two other research designs often used in empirical work: the regression discontinuity (RD) and the regression kink (RK) designs as laid out, for example, by Imbens & Lemieux (2008) and Card et al. (2015). RD and RK designs essentially exploit notched and kinked incentives, respectively, but in situations in which the assignment variable—the variable that determines whether the agent is above or below the relevant threshold—is not subject to choice or manipulation. Bunching designs consider the opposite case, in which the assignment variable is a direct choice. In this sense, whenever we observe discrete jumps in incentives at specific thresholds, it is potentially possible to use either RD/RK designs or bunching designs, depending on the manipulability of the assignment variable. A complication in practice is that the manipulability of the assignment variable may not always be clearly determined, especially in situations with optimization frictions.

The article proceeds as follows. Section 2 describes the relationship between the bunching literature and the traditional nonlinear budget set approach, Section 3 lays out the theory underlying bunching estimation, Section 4 describes the empirical implementation and challenges of bunching approaches, Section 5 discusses applications across a wide range of topics, and Section 6 concludes.

## 2. TRADITIONAL NONLINEAR BUDGET SET APPROACH

The econometric study of nonlinear budget sets was initially developed by Burtless & Hausman (1978) and Hausman (1981), who considered, respectively, labor supply responses to the negative income tax experiments and those to the federal income tax in the United States. They started from the observation that income tax and transfer systems create piecewise linear budget sets with two types of kink points. A convex kink point is created in which the marginal tax rate discretely increases (such as at bracket cutoffs in graduated income taxes), and a nonconvex kink point is created in which the marginal tax rate discretely falls (such as at points at which means-tested transfers are fully exhausted and no longer taxed away at the margin). The first type of kink should produce bunching, whereas the second type of kink should produce a hole in the distribution of earnings. They parametrically estimated labor supply models in which workers locate either in the interior of a linear budget segment or at a convex kink point. This approach became very dominant during the 1980s and was applied to a wide range of government policies, such as income taxes, welfare programs, social insurance, and social security. A review of this literature is provided by Moffitt (1990).

An advantage of the approach was its clear link between theory and empirics, but there was an elephant in the room: Although the models underlying the estimations implied bunching at convex kink points, no bunching was found in the survey data used.[2] Nor were any holes observed around nonconvex kink points. Given that the size of bunching and holes at kink points is proportional to

[2]A notable exception is the study by Burtless & Moffitt (1984) of the effect of US Social Security on retirement ages and earnings after retirement. They found strong bunching in retirement ages at 65 (corresponding to a convex kink point in Social Security benefits as a function of retirement age) as well as in postretirement earnings at the exemption threshold above which benefits are taxed away (corresponding to a convex kink point in benefits as a function of earnings). Bunching in earnings at the exemption threshold was also studied by Friedberg (1998, 2000).



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

the compensated elasticity of labor supply, a point later clarified by Saez (1999, 2010), it would seem that the nonlinear budget set approach should produce compensated elasticity estimates of zero. The solution to this bind was to allow the econometric model to have two error terms: One error term would represent unobserved preference heterogeneity; the other error term would represent optimization error capturing the inability of individuals to fine-tune hours worked precisely. The preference error term affects whether the true desired location is at the kink, whereas the optimization error term allows for individuals preferring the kink to be observed away from it.[3] The models were then estimated using maximum likelihood assuming that each of the error terms is normally distributed.

This approach allowed researchers to structurally estimate labor supply models, sometimes finding very large compensated elasticities as in Hausman (1981), using data with no bunching or holes around kink points. The approach would yield an estimate of the variance of the optimization error term: This would be determined by the amount of clustering around kink points—the less clustering, the larger the variance—and represent the degree of optimization error among individuals. As discussed below, this approach is conceptually related to the recent bunching literature, which emphasizes the role of optimization frictions in creating a gap between observed elasticities and true structural elasticities (Chetty et al. 2011, Chetty 2012, Kleven & Waseem 2013) and argues that the latter may be much larger than the former.

Where the two literatures diverge is in terms of empirical identification. In the nonlinear budget set literature, identification was achieved using a parametric model and making distributional assumptions on the two error terms. The presence of kinks and bunching (or their absence) was largely treated as a technical complication in fitting models to the data; the fact that kink points represent quasi-experimental variation in incentives and that bunching can be directly informative of responsiveness was not exploited. The recent literature, conversely, uses bunching directly to elicit behavioral responses and to estimate elasticities. Unlike the earlier literature, the recent bunching literature achieves identification only from what happens locally around the kink rather than from variation within brackets.

## 3. BUNCHING THEORY

### 3.1. Kinks

This section lays out the bunching theory that underlies the empirical designs discussed below. The analysis is framed in terms of earnings responses to taxes, but the conceptual framework—or modified versions of it—has found applications in a range of other settings. It starts by considering kink points created by discontinuities in marginal tax rates, the analysis of which was developed by Saez (2010).

Consider individuals with preferences defined over after-tax income (value of consumption) and before-tax income (cost of effort). The utility function can be written as $u(z - T(z), z/n)$, where $z$ is earnings, $T(z)$ is a tax function, and $n$ is ability. There is heterogeneity in ability captured by a density distribution $f(n)$. Assuming that the ability distribution, preferences, and the tax system are smooth, individual optimization generates an earnings distribution that is also smooth. As a baseline, we consider a linear tax system $T(z) = t \cdot z$ and denote the smooth earnings distribution in this baseline by $h_0(z)$.

[3]The optimization error term could also represent measurement error: Both forms of error attenuate bunching and are observationally equivalent in this context. Another (closely related) approach to deal with the absence of bunching when fitting nonlinear budget set models to the data was to smooth the budget set around the kink (MaCurdy et al. 1990).



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

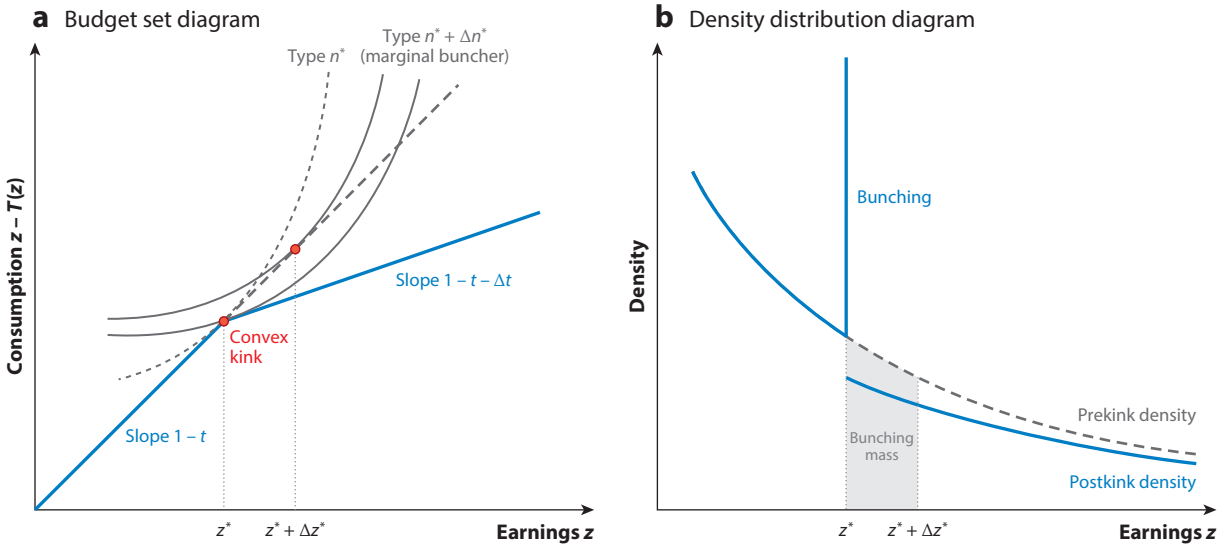


**Figure 1**

Kink analysis, showing the effects of a convex kink—a discrete increase in the marginal tax rate from $t$ to $t + \Delta t$ at the earnings threshold $z^*$—in a (*a*) budget set diagram and (*b*) density diagram. In panel *a*, the individual with ability $n^* + \Delta n^*$ is the marginal bunching individual. This individual chooses $z^* + \Delta z^*$ before the kink is introduced and $z^*$ after the kink is introduced. All workers initially located on the interval $(z^*, z^* + \Delta z^*)$ bunch at the kink, whereas all those initially located above $z^* + \Delta z^*$ reduce earnings within the interior of the upper bracket. As shown in panel *b*, the implications of these responses for the earnings distribution are sharp bunching at $z^*$ (the size of which is equal to the *gray shaded area* just above $z^*$) and a left shift of the distribution in the upper bracket.

Suppose that a convex kink—a discrete increase in the marginal tax rate from $t$ to $t + \Delta t$—is introduced at the earnings threshold $z^*$. The kinked tax function is given by $T(z) = t \cdot z + \Delta t \cdot (z - z^*) \cdot \mathbf{I}(z > z^*)$, where $\mathbf{I}(\cdot)$ is an indicator function. **Figure 1*a*,*b*** illustrates the effects in a budget set diagram and a density distribution diagram, respectively. Absent the kink, workers locate along the linear budget line with slope $1 - t$ depending on their abilities. As shown in the figure, an individual with ability $n^*$ chooses earnings $z^*$, and an individual with ability $n^* + \Delta n^*$ chooses $z^* + \Delta z^*$. When the kink is introduced, the individual initially located at $z^* + \Delta z^*$ is tangent to the upper part of the budget set at the kink point $z^*$ and therefore moves down to the kink. This is the marginal bunching individual: All workers initially located on the interval $(z^*, z^* + \Delta z^*)$ move to the kink point; all workers initially located above this interval stay in the interior of the upper bracket. This behavior produces excess bunching in the earnings distribution at the kink point, as shown in **Figure 1*b***. It does not produce a hole in the distribution above the kink because those located above the marginal buncher reduce their earnings in response to the higher marginal tax rate and fill up the hole. These interior responses are represented by the left shift of the density distribution above $z^*$. The excess bunching at $z^*$ is precisely offset by the missing mass on $(z^*, \infty)$ in the postkink relative to the prekink distribution.[4]

[4]A clarification of terminology is in order here: I am using the terms prekink (baseline) and postkink, although in many empirical applications there is no such temporal variation in the kink. In a typical application, there is an observed scenario with a kink (postkink) and an unobserved—but potentially estimable—counterfactual scenario without a kink (prekink). Furthermore, the analysis here assumes that the counterfactual scenario without a kink is characterized by the lower-bracket tax rate $t$ throughout (in which case bunchers are coming from above $z^*$) as opposed to the higher-bracket tax rate $t + \Delta t$ throughout (in which case bunchers would be coming from below). To be consistent with this counterfactual benchmark,

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

The key insight of the bunching approach is that the (compensated) earnings elasticity can be inferred from the response by the marginal buncher, $\Delta z^*$, and that this response is proportional to the amount of excess bunching. For the marginal buncher, the earnings response $\Delta z^*$ represents a standard interior response between two tangency points. Hence, assuming that the kink $\Delta t$ is small, we can define an earnings elasticity as

$$e = \frac{\Delta z^*/z^*}{\Delta t/(1-t)}. \tag{1}$$

Given that the kink does not change the tax rate on inframarginal units of income below $z^*$, it does not produce income effects on the (small) bunching segment $(z^*, z^* + \Delta z^*)$. The absence of income effects implies that $e$ represents a compensated elasticity. Large kinks can produce large bunching segments in which case the elasticity $e$ will be a weighted average of the compensated and uncompensated elasticities.[5]

The final step of the approach is to link the earnings response $\Delta z^*$ in the elasticity formula to the amount of bunching, which is the empirical entity that will be estimated. Denoting total bunching by $B$, we have

$$B = \int_{z^*}^{z^*+\Delta z^*} h_0(z)\,\mathrm{d}z \simeq h_0(z^*)\,\Delta z^*, \tag{2}$$

where the approximation assumes that the baseline (counterfactual) density $h_0(z)$ is constant on the bunching segment $(z^*, z^* + \Delta z^*)$. The constant density assumption simplifies the analysis (and is innocuous when the bunching segment is small), but the assumption is in general unnecessary. Empirical implementations of the approach can allow for curvature and use the exact bunching relationship in Equation 2. From Equations 1 and 2, we have a relationship going from the estimable entities $B, h_0(z^*)$ via the earnings response of the marginal buncher $\Delta z^*$ to the compensated elasticity $e$. This is a local elasticity at the earnings level $z^*$.

The preceding analysis assumes homogeneous preferences $u(\cdot)$ and thus a single elasticity $e$ at the earnings level $z^*$. However, it is straightforward to allow for heterogeneity in elasticities. Consider a joint distribution of abilities $n$ and elasticities $e$ given by $\hat{f}(n, e)$, and a joint baseline distribution of earnings $z$ and elasticities $e$ given by $\hat{h}_0(z, e)$. We have $h_0(z) = \int_e \hat{h}_0(z, e)\,\mathrm{d}e$. At each elasticity level $e$, we can characterize earnings responses to a kink exactly as above and denote the response of the marginal buncher by $\Delta z_e^*$. We can then link bunching $B$ to the average earnings response $E[\Delta z_e^*]$ as follows:

$$B = \int_e \int_{z^*}^{z^*+\Delta z_e^*} \hat{h}_0(z, e)\,\mathrm{d}z\mathrm{d}e \simeq h_0(z^*)\,E\left[\Delta z_e^*\right]. \tag{3}$$

---

the counterfactual distribution should be estimated assuming that the bunching mass in the observed distribution comes from above $z^*$. One could alternatively consider a counterfactual in which the higher-bracket tax rate applies throughout and estimate a counterfactual distribution assuming that the bunching mass comes from below. As long as the kink is small, the two procedures will produce the same elasticity estimate. If the kink is large, the two elasticity estimates may be different. For reasons explained in Section 4.1, this subtle distinction between different linear counterfactuals (i.e., whether bunchers are coming from above or below) is unlikely to make much of a difference in most bunching applications.

[5]This can be seen formally as follows. The earnings supply function of an individual in the upper tax bracket $z > z^*$ can be written as $z = z(1-\tilde{t}, Y)$, where $\tilde{t}$ is the marginal tax rate in the upper bracket (i.e., $\tilde{t} = t$ before the kink; $\tilde{t} = t + \Delta t$ after the kink) and $Y \equiv \tilde{t} \cdot z - T(z)$ is virtual income. Denoting by $e^{\mathrm{c}}$ and $e^{\mathrm{u}}$ the compensated and uncompensated elasticities of $z$ with respect to $1-\tilde{t}$, the Slutsky decomposition implies $e^{\mathrm{u}} = e^{\mathrm{c}} + \eta$, where $\eta = (1-\tilde{t})(\partial z/\partial Y)$ is the income effect. Considering a small change in the marginal tax rate from $t$ to $t + \Delta t$ above the threshold $z^*$, and using the Slutsky decomposition, one finds that the earnings reduction $\Delta z$ in the interior of the upper bracket satisfies $(\Delta z/z)/(\Delta t/(1-t)) = (1 - \Delta a/\Delta t) \cdot e^{\mathrm{c}} + \Delta a/\Delta t \cdot e^{\mathrm{u}}$, where $\Delta a \equiv \Delta t\,(z - z^*)/z$ is the change in the average tax rate at the income level $z$. For upper-bracket taxpayers located close to the kink ($z \approx z^*$), we have $\Delta a \approx 0$ such that the right-hand side equals $e^{\mathrm{c}}$. Specifically, the marginal bunching individual (whose response is like an interior response and therefore can be characterized as above) comes from a point $z = z^* + \Delta z^*$ close to the kink, so the earnings response $\Delta z = \Delta z^*$ of this individual is related to $e^{\mathrm{c}}$.

Here the approximation assumes that the counterfactual density $\hat{h}_0(z, e)$ is constant in $z$ on the bunching segment $(z^*, z^* + \Delta z_e^*)$ for all $e$. Replacing $\Delta z^*$ by $E[\Delta z_e^*]$ in Equation 1, we can link bunching to the local average earnings elasticity at $z^*$.

The analysis presented so far is exact only if the kink is sufficiently small. In the presence of large kinks, it is necessary to specify preferences parametrically to obtain exact elasticities (but this introduces potential functional form sensitivity, as discussed below). The typical approach is to specify a quasi-linear, isoelastic utility function

$$u = z - T(z) - \frac{n}{1 + 1/e} \cdot \left(\frac{z}{n}\right)^{1+1/e}, \tag{4}$$

thus ruling out income effects of tax changes on earnings $z$. With this utility function, earnings in the linear tax baseline are given by $z = n(1 - t)^e$.

As explained above, in the presence of the kink the marginal buncher (with ability $n^* + \Delta n^*$) is tangent to the upper part of the kinked budget set at $z^*$ and to the initial linear budget set at $z^* + \Delta z^*$. Hence two tangency conditions must be met for the marginal buncher: Actual earnings with the kink satisfy $z^* = (n^* + \Delta n^*)(1 - t - \Delta t)^e$, and counterfactual earnings without the kink satisfy $z^* + \Delta z^* = (n^* + \Delta n^*)(1 - t)^e$. These two conditions imply

$$\frac{z^* + \Delta z^*}{z^*} = \left(\frac{1 - t}{1 - t - \Delta t}\right)^e, \tag{5}$$

or equivalently

$$e = -\frac{\log(1 + \Delta z^*/z^*)}{\log(1 - \Delta t/(1 - t))}, \tag{6}$$

which is a generalization of Equation 1. When $\Delta t$ is small (so that $\Delta z^*$ is also small), we have $\log(1 + \Delta z^*/z^*) \approx \Delta z^*/z^*$ and $\log(1 - \Delta t/(1 - t)) \approx -\Delta t/(1 - t)$, in which case the exact parametric formula (Equation 6) is approximately equal to the simpler nonparametric version (Equation 1).

Finally, although we have focused on the implications of a convex kink, the conceptual analysis can be extended to a nonconvex kink as created by a discrete fall in the marginal tax rate at a threshold. For example, nonconvex kinks are observed at points where means-tested transfers are fully phased out and therefore no longer taxed away at the margin. This type of kink should produce a hole around the threshold $z^*$, as individuals who would otherwise locate in a range just below the threshold are willing to locate strictly above the threshold, whereas individuals further down do not respond at all. In this case, there will be a marginal responding individual, who is precisely indifferent between a point strictly below and a point strictly above the threshold. No individual locates between these two points, and the width of the hole can be linked to the compensated earnings elasticity. Even though nonconvex kinks are quite common, their analysis has not received much attention in the literature for the simple reason that no research has found any evidence of holes around such kinks. This nonfinding poses a challenge to the framework that I discuss and try to resolve below.

## 3.2. Notches

We now turn to the analysis of notches created by discontinuities in tax liability (i.e., in the average tax rate), the analysis of which was developed by Kleven & Waseem (2013). The basic conceptual framework is the same as above: Preferences are modeled in the same way, there is a smooth distribution of ability $f(n)$, and there is a smooth distribution of earnings $h_0(z)$ in the baseline

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

without notches. As with kinks, I start by considering a homogeneous earnings elasticity and then generalize to allow for heterogeneity.

Starting from a baseline with linear taxation, consider the introduction of a notch—a discrete increase in the average tax rate from $t$ to $t + \Delta t$—at the earnings threshold $z^*$. That is, we consider a tax function given by $T(z) = t \cdot z + \Delta t \cdot z \cdot \mathbf{I}(z > z^*)$. This upward tax notch (average tax increase) is analogous to a convex kink (marginal tax increase), and the case of downward tax notches is discussed below. The notch considered here takes the form of a discontinuity in a proportional tax rate, and thus the threshold represents a discontinuity in both the average and marginal tax rate. Although such proportional tax notches are quite common in practice, an alternative form of notch consists of a tax liability jump without any change in the marginal tax rate. It is straightforward to include such notches in the analysis as well (Kleven & Waseem 2013).

**Figure 2** shows the implications of the notch in a budget set diagram (**Figure 2*a***) and in density distribution diagrams (**Figure 2*b*,*c***). There will be bunching at the notch point by all individuals who had incomes in an interval $(z^*, z^* + \Delta z^*)$ prior to the introduction of the notch. The individual originally located at $z^* + \Delta z^*$ is the marginal bunching individual: This person is exactly indifferent between the notch point $z^*$ and the best interior point $z^{\mathrm{I}}$ after the tax change. Those initially located above $z^* + \Delta z^*$ reduce their earnings in response to the proportional tax change, but stay in the interior of the upper bracket. There is a hole in the postnotch density distribution, as no individual is willing to locate between $z^*$ and $z^{\mathrm{I}}$.

An important difference between kinks and notches is that the latter create a region of strictly dominated choice $(z^*, z^* + \Delta z^{\mathrm{D}})$. In this region, it is possible to increase both consumption and leisure by moving down to the notch point $z^*$, making such earnings choices dominated under any parametric form for preferences. The dominated region $(z^*, z^* + \Delta z^{\mathrm{D}})$ creates a lower bound for the bunching region $(z^*, z^* + \Delta z^*)$. In the case of L-shaped Leontief preferences—such that the compensated earnings elasticity is zero—the bunching region would correspond exactly to the dominated region.

As with kinks, the fundamental idea is that the earnings response $\Delta z^*$ of the marginal buncher (which can be uncovered from bunching) is related to the compensated elasticity $e$. In the case of notches, the relationship between the two can be characterized using the indifference condition between the notch point $z^*$ and the interior location $z^{\mathrm{I}}$ for the marginal buncher, as opposed to the tangency condition at $z^*$ used in the case of kinks.

Based on the preference specification (Equation 4), utility at the notch point $z^*$ for the marginal buncher is given by

$$u^* = (1-t)\, z^* - \frac{n^* + \Delta n^*}{1 + 1/e} \left( \frac{z^*}{n^* + \Delta n^*} \right)^{1+1/e}. \tag{7}$$

Using the first-order condition $z^{\mathrm{I}} = (n^* + \Delta n^*)(1 - t - \Delta t)^e$, one can write utility at the interior point $z^{\mathrm{I}}$ as

$$u^{\mathrm{I}} = \left( \frac{1}{1+e} \right) (n^* + \Delta n^*)(1 - t - \Delta t)^{1+e}. \tag{8}$$

From the condition $u^* = u^{\mathrm{I}}$ and using the relationship $n^* + \Delta n^* = \frac{z^* + \Delta z^*}{(1-t)^e}$, we can rearrange terms so as to obtain

$$\frac{1}{1 + \Delta z^*/z^*} - \frac{1}{1 + 1/e} \left[ \frac{1}{1 + \Delta z^*/z^*} \right]^{1+1/e} - \frac{1}{1+e} \left[ 1 - \frac{\Delta t}{1-t} \right]^{1+e} = 0. \tag{9}$$

This condition, which is the analog of Equations 5 and 6 for kinks, characterizes the relationship between the percentage earnings response $\Delta z^*/z^*$, the percentage change in the average net-of-tax rate $\Delta t/(1-t)$, and the compensated elasticity $e$. As the earnings response is estimated from



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

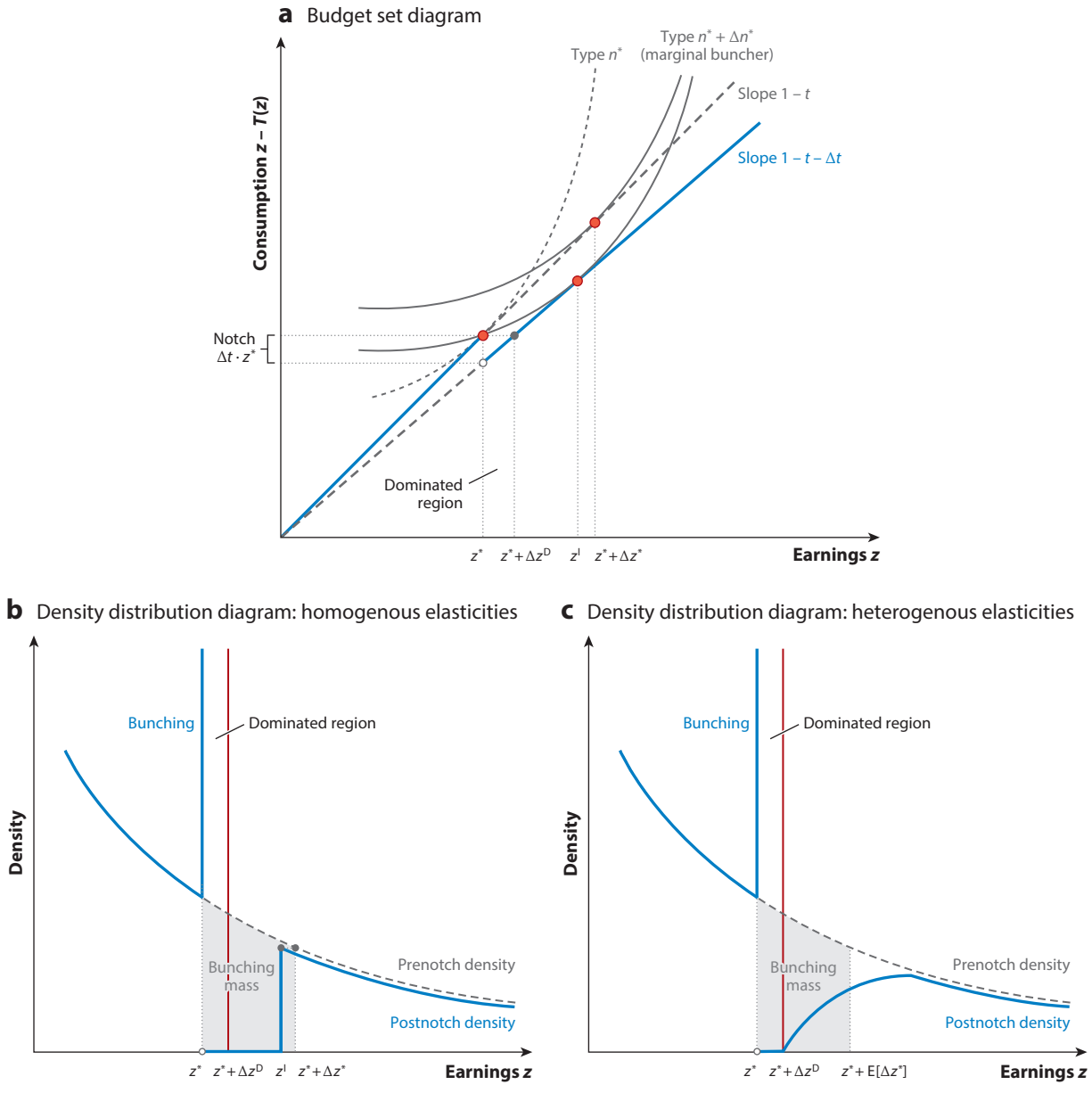


**Figure 2**

Notch analysis, showing the effects of a notch—a discrete increase in the average tax rate from $t$ to $t + \Delta t$ at the earnings threshold $z^*$—in (*a*) a budget set diagram and (*b*,*c*) density distribution diagrams. In panel *a*, the individual with ability $n^* + \Delta n^*$ is the marginal bunching individual. This individual chooses $z^* + \Delta z^*$ before the notch is introduced, and is indifferent between the threshold $z^*$ and the best interior location $z^I$ after the notch. All workers initially located on $(z^*, z^* + \Delta z^*)$ bunch at the notch, whereas all those initially located above $z^* + \Delta z^*$ reduce earnings marginally within the interior of the upper bracket. Panel *a* is drawn for a specific elasticity $e > 0$, and panel *b* shows the corresponding postnotch density distribution, which features sharp bunching at $z^*$ (the size of which is equal to the *gray shaded area* just above $z^*$) and an empty hole between $z^*$ and $z^I$. Panel *c* shows the density distribution with heterogeneous elasticities on $(0, \bar{e})$, which is completely empty only in the strictly dominated region.

bunching, the condition should be viewed as defining the elasticity $e$ as an implicit function of the observed values of $\Delta z^*/z^*$ and $\Delta t/(1-t)$.

The elasticity formula (Equation 9) confirms the argument above that the strictly dominated range is a lower bound for the earnings response in this frictionless model. As the elasticity $e$ converges to zero, Equation 9 implies

$$\lim_{e \to 0} \Delta z^* = \frac{\Delta t \cdot z^*}{1 - t - \Delta t} \equiv \Delta z^{\mathrm{D}}, \tag{10}$$

where $\Delta z^{\mathrm{D}}$ is defined such that the earnings level $z^* + \Delta z^{\mathrm{D}}$ ensures the same consumption as the notch point $z^*$. The fact that, absent any optimization frictions, notches should create a completely empty range of the earnings distribution under any elasticity is very useful for the empirical estimation of structural elasticities in settings in which optimization frictions are present.

I have so far assumed a single elasticity $e$ at the earnings level $z^*$, but it is conceptually straightforward to allow for heterogeneity as in the analysis of kinks. With a joint distribution $\hat{f}(n, e)$, the preceding analysis characterizes earnings responses $\Delta z_e^*$ at a particular value of $e$. Aggregating across all elasticity levels [assuming that $e$ is smoothly distributed on $(0, \bar{e})$] would give the type of earnings distribution illustrated in **Figure 2*c***. Here the hole does not have a sharp upper edge; instead, the distribution gradually converges toward the counterfactual. As in the case of kinks, we can link bunching $B$ to the average earnings response $E[\Delta z_e^*]$ based on the relationship given in Equation 3. Using Equation 9, one can then estimate the elasticity at the average response $E[\Delta z_e^*]$. Because of the nonlinearity of Equation 9, the elasticity at the average response is generally different from the average elasticity, creating a form of aggregation bias. However, such bias can be bounded and is typically very small, as discussed below.

The analysis can be extended to the case of downward tax notches, in which the average tax rate falls discretely above a threshold (see Kleven et al. 2014). In this case, the theoretical predictions are a mirror image of those described above. In response to a reduction in the average tax rate by $\Delta t$ above the threshold $z^*$, there will be bunching just above $z^*$ and a hole below $z^*$. A marginal bunching individual at ability $n^* - \Delta n^*$ is indifferent between the notch point $z^*$ and the prenotch location $z^* - \Delta z^*$.[6] This indifference condition can be shown to imply

$$\left(1 - \frac{\Delta z^*}{z^*}\right) + e\left(1 - \frac{\Delta z^*}{z^*}\right)^{-1/e} - (1+e)\left(1 + \frac{\Delta t}{1-t}\right) = 0, \tag{11}$$

which is the analog of the condition in Equation 9 for the case of upward tax notches. It is then possible to link bunching $B$ to the earnings response $\Delta z^*$ (using Equation 2 or 3) and the earnings response to the elasticity $e$ (using Equation 11). A conceptual difference between upward and downward tax notches is that the latter do not create a strictly dominated region because moving from below the notch to above the notch (and thus obtaining larger consumption) is associated with less leisure. Because of the absence of a dominated region, as the elasticity $e$ converges to zero, Equation 11 implies that the earnings response $\Delta z^*$ also converges to zero. Besides this difference, upward and downward notches work in similar ways, as opposed to convex and nonconvex kinks, which work in very different ways.

The preceding analysis of notches relies on a functional form for utility. Although the earnings response $\Delta z^*$ (estimated from bunching at the notch) can be nonparametrically identified, the underlying structural elasticity $e$ that could be used for out-of-sample prediction cannot. It is of interest to develop a reduced-form approach without such parametric reliance, similar to

[6] With a downward tax notch, the best interior location $z^{\mathrm{I}}$ is identical to the prenotch location $z^* - \Delta z^*$ because the marginal buncher faces no change in the budget set in the no-bunching scenario.



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

the Saez elasticity (Equation 1) for kinks. As discussed by Kleven & Waseem (2013), a reduced-form approach is less straightforward for notches than for kinks because the behavioral response is driven by a jump in the average tax rate rather than in the marginal tax rate of direct relevance to the structural parameter of interest. They propose a reduced-form approximation in which the earnings response $\Delta z^*$ is related to the change in the implicit marginal tax rate between $z^*$ and $z^* + \Delta z^*$ created by the notch. With the implicit marginal tax rate defined as $t^* \equiv [T(z^* + \Delta z^*) - T(z^*)]/\Delta z^* \approx t + \Delta t \cdot z^*/\Delta z^*$, the reduced-form elasticity is given by

$$e_R \equiv \frac{\Delta z^*/z^*}{\Delta t^*/(1-t^*)} \approx \frac{(\Delta z^*/z^*)^2}{\Delta t/(1-t)}. \quad (12)$$

As shown by Kleven & Waseem (2013), this simple quadratic formula represents an upper bound on the true structural elasticity $e$ under a weak assumption on preferences.

## 3.3. Extensions

In this section, I discuss a number of extensions of the baseline framework presented above: optimization frictions, reference dependence, dynamics, and extensive margin responses.

**3.3.1. Optimization frictions.** In the frictionless model considered above, bunching depends on a structural elasticity $e$ and a vector of observable parameters $\mathbf{x}$, i.e., $B = B(e, \mathbf{x})$. This allows us to go from an estimate of bunching to an estimate of the structural elasticity. However, in practice, agents may face optimization frictions such as adjustment costs and attention costs that prevent them from bunching at kinks and notches (Chetty et al. 2011, Chetty 2012, Kleven & Waseem 2013).[7] In this case, we can write bunching as $B = B(e, \phi, \mathbf{x})$, where $\phi$ is a parameter, or a vector of parameters, characterizing the adjustment costs. In this case, a single observation of bunching will be consistent with a set of $(e, \phi)$ combinations. The observed elasticity obtained from bunching is generally different from the true structural elasticity $e$, with the gap between the two determined by the unobserved friction $\phi$.

I now describe two conceptual approaches to uncover the structural parameter $e$ in the presence of optimization frictions, first a nonparametric approach based on notches and then a parametric approach that can be applied to both kinks and notches.[8] In general, the fundamental problem described above is that we have only one empirical moment $B$ and two unobserved parameters $(e, \phi)$, or more than two unobserved parameters if $\phi$ is a vector. Hence the general solution is to obtain additional empirical moments that depend on the same parameters $(e, \phi)$.

In the case of notches, Kleven & Waseem (2013) develop an approach in which the additional empirical moment is the hole in the distribution above the threshold, or specifically the hole in the strictly dominated region just above the threshold. Because the dominated region should be empty in a frictionless world under any preferences, the observed density mass in this region can be used to measure attenuation bias from frictions. To see how the approach works, denote by $a(z, e, \phi)$ the share of individuals at earnings level $z$ and elasticity level $e$ who are unresponsive due

[7] Another implication of optimization friction is that agents who do respond may not be able to target the threshold precisely, so that excess bunching manifests itself as diffuse excess mass rather than a point mass. As long as excess mass is not too diffuse to be visible, this can easily be incorporated in empirical applications by allowing for a bunching interval rather than a bunching point.

[8] Other things equal, bunching at notches should be less affected by optimization frictions than bunching at kinks because notches create much stronger incentives to bunch and are therefore more likely to overcome adjustment costs.



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

to adjustment costs $\phi$. We then have

$$B = \int_e \int_{z^*}^{z^* + \Delta z_e^*} (1 - a(z, e, \phi)) \hat{h}_0(z, e) \, dz de \simeq h_0(z^*)(1 - a^*(\phi)) E\left[\Delta z_e^*\right], \quad (13)$$

where the approximation assumes a locally constant counterfactual density (as above) and a locally constant share of unresponsive individuals, $a(z, e, \phi) = a^*(\phi)$ for $(z^*, z^* + \Delta z_e^*)$ and all $e$. In this expression, $E[\Delta z_e^*]$ is the frictionless response governed by the parameter $e$. Given estimates of $B$ and $h_0(z^*)$, the frictionless response can be identified using an estimate of the share of nonoptimizers $a^*(\phi)$, which can be obtained from the observed density mass in the strictly dominated region.

The approach is based on the assumption that the share of nonoptimizers is constant in a region above the notch, but it does not rely on specific parametric assumptions on the structure and distribution of adjustment costs $\phi$.[9] The local fraction of nonoptimizers $a^*(\phi)$ represents a sufficient statistic (together with bunching) for the structural elasticity $e$. However, a limitation of the approach is that not all notches are associated with strictly dominated regions. In the labor-leisure context considered here, upward tax notches create dominated regions, but downward tax notches do not. More generally, notches in other decision contexts than labor-leisure do not always create strictly dominated regions (e.g., Almunia & Lopez-Rodriguez 2015, Best et al. 2015b, Best & Kleven 2016). In such cases, one can implement a more parametric version of the approach by ruling out extreme preferences, allowing for the recovery of $a^*(\phi)$ from a very narrow range above the threshold (e.g., Best et al. 2015b).

Kinks do not allow for this type of approach. A single kink offers only one empirical moment for estimation—the size of bunching—which is consistent with many $(e, \phi)$ combinations.[10] To separately estimate the structural parameter $e$, one must obtain at least one more bunching observation that depends on the same underlying parameters. In principle, this is possible if we observe variation in the size of the kink that is orthogonal to $e$ and $\phi$. In practice, there are two potential sources of such variation: (*a*) differently sized kinks located at different earnings thresholds and (*b*) changes in the size of a kink at a given earnings threshold over time. With such additional variation, we can make progress by assuming that the additional bunching moment(s) is generated by the same underlying elasticity and friction. Versions of approaches (*a*) and (*b*) have been developed by Chetty et al. (2010, 2011) and Gelber et al. (2014).

These approaches require us to specify the adjustment cost parameters in $\phi$. The simplest possible case is one with a fixed cost of adjusting earnings equal to $\bar{\phi}$, corresponding to the specification in Gelber et al. (2014). In this case, if we observe bunching in two different kink

[9]Kleven & Waseem (2013) argue that, in general, $a^*(\phi)$ obtained from the dominated region is a lower bound on friction over the entire bunching segment $(z^*, z^* + \Delta z_e^*)$. This is because, given the adjustment cost $\phi$, the fraction of nonresponders $a(z, e, \phi)$ is naturally increasing in earnings as the utility gain of moving to the notch is falling in earnings. However, the bias depends on the distribution of adjustment costs $\phi$. In a scenario in which a fraction of agents have zero adjustment costs (optimizers) and a fraction have prohibitively high adjustment costs (nonoptimizers), $a^*(\phi)$ accurately captures the fraction of nonoptimizers and yields unbiased estimates.

[10]In theory, there is another potentially useful moment besides the size of the bunch, namely the width or the diffuseness of the bunch. If there is a lot of optimization friction that makes it hard to target the kink point precisely, we would expect excess bunching to be very diffuse around the threshold. Hence, in a situation in which bunching around the kink is small (by itself, suggesting either a small elasticity $e$ or a large friction $\phi$) and at the same time sharp (suggesting a small friction $\phi$), it becomes harder to argue that the true structural elasticity is large. Still, it is conceivable that a subset of agents are completely unaware of the kink or completely unable to adjust their behavior with any amount of precision (while another subset of informed and flexible agents can bunch sharply), allowing for the observation of small and sharp bunching in a situation in which the true structural elasticity is large.



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

scenarios, we have

$$B_1 = B_1\left(e, \bar{\phi}, \mathbf{x}_1\right), \tag{14}$$

$$B_2 = B_2\left(e, \bar{\phi}, \mathbf{x}_2\right), \tag{15}$$

where $\mathbf{x}_i$ is a vector of observable tax variables and the counterfactual density for kink $i$. These are two equations in two unknowns that can be solved for the adjustment cost and the structural elasticity $e$. The key parametric assumptions are that the friction takes the form of a fixed cost and that the fixed cost and elasticity are the same at the two kinks. If we generalize the adjustment cost to include both a fixed cost element $\bar{\phi}$ and a variable cost element governed by another parameter $\gamma$, then we would need three bunching moments to estimate the model or alternatively have to calibrate one of the friction parameters. Chetty et al. (2010) consider a more involved model in which the friction results from the cost of searching for a job with earnings at the kink. In this model, the search cost depends on two parameters (a scale parameter $\phi$ and an elasticity parameter $\gamma$), and search effort affects the variance of the distribution job offers (assumed to be normal with a baseline variance under zero search given by $\sigma^2$). Hence, the model has three friction parameters $\phi$, $\gamma$, and $\sigma^2$ and a structural elasticity $e$, which would require four bunching moments to be fully identified. As they have only two bunching moments, they calibrate two friction parameters $\phi$ and $\sigma^2$ and solve for $\gamma$ and $e$ based on a system like Equations 14 and 15.

Finally, the kink approaches described here—relying on parametric assumptions on the structure and distribution of optimization frictions—are related to the nonlinear budget set approach discussed in Section 2, which was based on parametrically estimating a distribution of optimization errors using bunching (or rather its absence) at kink points.[11]

**3.3.2. Reference points.** An issue that has received relatively little attention in the literature is that kinks and notches may represent reference points. This would be the case if the threshold were a natural focal point for reasons other than the financial incentive (for example, because it is a salient round number) or if the creation of a statutory threshold makes it a focal point. When governments legislate that public policies change at specific thresholds, they are potentially creating reference points in addition to financial incentives. A concrete example is when social security benefits change at statutory retirement ages, conceivably creating focal points for retirement by workers and their employers. Opposite optimization frictions, such reference point effects, amplify bunching and make the observed elasticity overstate the structural price elasticity.

Theories of reference dependence can be cast in the language of kinks and notches. The most influential theory of reference dependence is prospect theory by Kahneman & Tversky (1979), which posits that utility is defined over differences from a reference point and features a kink—a discontinuity in the first derivative—at the reference point. This kink is assumed to be convex, a feature known as loss aversion, implying that individuals should bunch at the reference point. An alternative theory of reference dependence is one in which utility features a notch—a discontinuity in the level—at the reference point. This would produce bunching on one side and a hole on the other side of the reference point.[12] These two models may be appropriate in different settings and could potentially be distinguished empirically. The notch-based theory is arguably more natural

[11] The identifying variation is different in the two approaches, however. Although the nonlinear budget set approach achieved identification based on cross-sectional variation in labor supply and taxes within brackets, the kink approach is based on differences in bunching observed under different tax parameters.

[12] As shown by Allen et al. (2016), the joint presence of bunching and a hole on each side of the reference point can also be reconciled with prospect theory due to another one of its features: diminishing sensitivity—a discontinuity in the second



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

in settings in which reference points represent goals that agents strive to meet (e.g., Pope & Simonsohn 2011, Allen et al. 2016), as opposed to settings in which reference points represent expectations or the status quo (e.g., Koszegi & Rabin 2006).

Allowing for reference dependence could resolve two unexplained findings in the bunching literature. First, although we observe excess bunching around convex kink points (at least those that are large and salient), no research has found evidence of holes around nonconvex kink points (e.g., Saez 2010, Kleven & Waseem 2012, Einav et al. 2015b). A potential explanation is that, whereas a convex kink point represents a desirable location and hence may come to serve as a reference point, a nonconvex kink point is an undesirable location and therefore not a natural reference point. In other words, a convex kink tells agents where to be, and a nonconvex kink tells them only where not to be. Second, several studies find asymmetric bunching around convex kinks—bunching below the threshold, but not above—similar to the prediction for notches (e.g., Devereux et al. 2014, Gelber et al. 2014, Seim 2015). Such a pattern can be reconciled with the models of reference-dependent preferences described above.

These arguments imply that observed bunching may confound the financial incentive effect with a reference point effect as well as optimization frictions. That is, observed bunching is given by $B = B(e, \phi, r, x)$, where $r$ captures the reference point effect. In such cases, the solution to identifying the incentive effect would have to be in the same spirit as those discussed for frictions: It is necessary to obtain additional empirical moments to separately identify $e$, $\phi$, and $r$. At least three moments that depend on the same underlying parameters are required. Although this may sound difficult to do in practice, in some settings it is feasible. An example is a situation in which reference points are round numbers, in which case there will be round-number bunching, as documented by Kleven & Waseem (2013) for reported taxable income and by Best & Kleven (2016) for house prices. When the kink or notch of interest is located at a round number, extra moments that depend on the same reference point effect can be obtained from similar round numbers that are not notches. Netting out round-number bunching in this way, we are left with two unknown parameters, $e$ and $\phi$, that can be estimated using the approaches described above.

**3.3.3. Dynamics.** The bunching theory presented above is based on a static labor supply model. Three basic points have been made on the implications of extending the bunching analysis to a dynamic setting. First, bunching responses to a within-period tax kink in a multiperiod setting relates to the Frisch elasticity (including intertemporal substitution) rather than to the static compensated elasticity (Saez 2010). This argument assumes that bunching decisions made in a given period do not affect the likelihood of bunching at kinks or notches in future periods. Second, although the Frisch elasticity is larger than the compensated elasticity in the standard life-cycle labor supply model (MaCurdy 1981, Blundell & MaCurdy 1999), this is not necessarily true in a model in which current earnings affect future wages through career concerns, learning by doing, etc. In models with career effects of work effort, the response to temporary tax changes can be relatively small (Saez 2010, Best & Kleven 2013b), which may be a reason—in addition to optimization frictions discussed above—for observing small bunching at kinks and notches. Third, the presence of career effects reduces, but does not eliminate, the presence of a strictly dominated range above notches (Kleven & Waseem 2013). This assumes that the relationship between current earnings and future wages is continuous. In this case, when crossing the notch point, current net-of-tax earnings fall discretely, while future net-of-tax earnings increase only marginally by continuity

derivative of utility—according to which agents are risk averse in gains and risk loving in losses. Without diminishing sensitivity, prospect theory predicts bunching at the reference point, but no hole.



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

of the career effect. Given consumption smoothing behavior, this implies lower consumption in all periods along with lower leisure in the current period, so this is still strictly dominated.

Extending the bunching approach to dynamic settings is still in its infancy. As discussed below, a few recent bunching studies explicitly model dynamics and structurally estimate elasticities, but in different contexts than labor supply. Although the static model may be a reasonable approximation in a number of contexts—such as annual labor supply decisions by low-skilled workers—in other contexts dynamics is central to the decision problem (Best et al. 2015b; Einav et al. 2015a,b; Manoli & Weber 2015).

**3.3.4. Extensive margin.** Bunching at kinks or notches represents intensive margin responses to price incentives, and above we focus on how to relate such responses to a structural parameter of interest. A difference between kinks and notches is that the latter, by introducing a discrete jump in tax liability, may also create extensive margin responses above the threshold. Such responses will shift down the distribution throughout the upper bracket. The approach laid out above relies on an estimate of the earnings response $\Delta z^*$ by the marginal buncher using the relationships in Equations 2 and 3, which require an estimate of the counterfactual density $h_0(z)$ around the threshold $z^*$. More precisely, to estimate the intensive margin response conditional on participation, we need an estimate of the counterfactual density absent intensive margin responses to the notch, as opposed to the counterfactual density absent any response to the notch. Kleven & Waseem (2013) label the two concepts the partial and full counterfactuals, respectively.

Kleven & Waseem (2013) make a simple point of relevance to the empirical estimation of $h_0(z)$: Extensive margin responses converge to zero just above the notch. Intuitively, if in the absence of the notch individuals prefer earnings slightly above $z^*$, then in the presence of the notch they are better off moving to $z^*$ (which is almost as good as the prenotch situation) than moving to $z = 0$. In other words, the partial counterfactual of relevance to the estimation is smooth around the notch.

To see this formally, consider a model in which individuals choose earnings conditional on participation ($z > 0$) as characterized above, and then make a discrete choice between $z > 0$ and $z = 0$ facing a fixed cost of participation $q$ that is smoothly distributed in the population. Utility from participation is given by $u(z - T(z), z/n) - q$, and utility from nonparticipation is given by $u_0$. This implies that an individual participates if and only if $q \leq u(z - T(z), z/n) - u_0 \equiv \bar{q}$. When a notch is introduced at $z^*$, this creates both intensive and extensive responses by those with $z > z^*$. Consider individuals initially located at $z = z^* + \epsilon$, where $\epsilon > 0$ is sufficiently small that the threshold $z^*$ is preferred to the initial choice, conditional on participating. Such individuals respond at the extensive margin only if they were initially very close to the indifference point between participation and nonparticipation, specifically if $q \in (\bar{q} - \Delta\bar{q}, \bar{q})$, where

$$\Delta\bar{q} = u((z^* + \epsilon)(1 - t), (z^* + \epsilon)/n) - u(z^*(1 - t), z^*/n). \tag{16}$$

This implies $\lim_{\epsilon \to 0} \Delta\bar{q} = 0$ such that there are no extensive responses very close to the threshold.

Best & Kleven (2013a) generalize this argument to a matching frictions model with bargaining. In such frameworks, if a match between two parties (for example, a worker and an employer) would have occurred at $z^* + \epsilon \approx z^*$ absent the notch, then it is better for the worker to accept a pay cut of $\epsilon$ than to break the match (either dropping out of the labor market or keep searching for another employer). Kopczuk & Munroe (2015) note that, in a matching frictions model, part of the extensive response may reflect that matches above the threshold break down in favor of searching for another match below the notch (as opposed to dropping out of the market completely). They go on to argue that such extensive responses may be present "close to the notch" (making the hole bigger than the bunch), but it remains the case in their setting that these responses must converge

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

to zero just above the notch. The next section discusses how these insights impact the empirical implementation of the bunching approach in settings in which extensive responses are present.

## 4. BUNCHING ESTIMATION

### 4.1. Counterfactual Distribution and Bunching

The conceptual approach laid out in the previous section relies on an estimate of the counterfactual distribution, i.e., what the distribution would have looked like in the absence of kinks or notches. I start by describing what has become the standard approach to obtaining such counterfactuals, developed by Chetty et al. (2011) in the context of kinks and extended by Kleven & Waseem (2013) to notches, and then discuss potential issues with the approach and possible refinements. For the sake of concreteness, I describe the estimation procedure in the language of earnings responses to taxes (even though a broader set of applications is considered below), focusing on the case of kinked/notched tax increases in which bunchers are coming from above an earnings threshold.

The standard approach is to fit a flexible polynomial to the observed distribution, excluding data in a range around the threshold $z^*$, and extrapolate the fitted distribution to the threshold. Grouping individuals into earnings bins indexed by $j$, researchers have estimated the counterfactual distribution using a regression of the following form:

$$c_j = \sum_{i=0}^{p} \beta_i \cdot (z_j)^i + \sum_{i=z_-}^{z_+} \gamma_i \cdot \mathbf{1}[z_j = i] + \nu_j, \tag{17}$$

where $c_j$ is the number of individuals in bin $j$, $z_j$ is the earnings level in bin $j$, $[z_-, z_+]$ is the excluded range, and $p$ is the order of the polynomial.[13] In the case of kinks, the excluded range should be the area featuring excess bunching and will typically be a narrow symmetric range around the threshold (i.e., $[z_-, z_+] = [z^* - \Delta, z^* + \Delta]$).[14] In the case of notches, the excluded range should span the entire area affected by bunching responses (i.e., the area featuring either bunching or a hole) and will generally be a wider asymmetric range around the threshold. We come back to the determination of this range below.

The counterfactual bin counts are obtained as the predicted values from Equation 17 omitting the contribution of the dummies in the excluded range [i.e., $\hat{c}_j = \sum_{i=0}^{p} \hat{\beta}_i \cdot (z_j)^i$]. Excess bunching is then estimated as the difference between the observed and counterfactual bin counts in the bunching range (i.e., in the excluded range in the case of kinks and in the low-tax side of the excluded range in the case of notches). In the frictions methodology of Kleven & Waseem (2013), the fraction of nonoptimizers is obtained as the cumulated observed bin counts in the dominated range as a fraction of the cumulated counterfactual bin counts in that range. Following Chetty et al. (2011), standard errors are typically calculated using a bootstrap procedure in which a large number of earnings distributions are generated by random resampling of the residuals in Equation 17.

I now consider a number of technical issues and extensions of the baseline specification (Equation 17). First, the approach relies on specifying the excluded range $[z_-, z_+]$. In the case of kinks, this can typically be determined visually, but in the case of notches, the diffuseness of

[13] The estimation procedure initially developed by Saez (2010) can be viewed essentially as a special case of Equation 17, where $p = 0$, i.e., where the counterfactual distribution is assumed to be uniform in a narrow range around the bunch.

[14] In the frictionless model presented above, bunching is a mass point at $z^*$, in which case we have $\Delta = 0$. In reality, bunching will always be somewhat diffuse due to optimization error and randomness in earnings, which is accounted for by $\Delta > 0$.



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

the hole above the notch may make it difficult to visually determine the upper bound $z_+$. To deal with this issue, Kleven & Waseem (2013) develop an approach based on the condition that, absent extensive margin responses, excess mass below the notch must be equal to missing mass above the notch. Hence, they estimate Equation 17 together with $z_+$ through an iterative procedure that ensures that bunching mass equals missing mass.[15]

Second, Equation 17 does not account for the left shift in the observed distribution in the interior of the upper bracket, as illustrated in **Figure 1*b*** for kinks and **Figure 2*b*,*c*** for notches. That is, it assumes that the observed bin counts above $z_+$ correspond to counterfactual bin counts even though these observations may be distorted by intensive margin responses to the higher marginal tax rate within the upper bracket. In general, such effects are larger for kinks than for notches because the interior marginal tax rate change is typically larger for kinks. Although it is possible to account for such effects when estimating the counterfactual, there are two reasons why this can be ignored in many practical applications: (*a*) As shown in **Figure 1*b*** for kinks, the interior shift in the distribution corresponds to the bunching response $\Delta z^*$ (the shift is smaller in the case of notches), which tends to be a very small number due to the local nature of bunching responses. (*b*) The implications of shifting a distribution to the left depends on its slope and will have a significant effect on bin counts around the threshold only if the distribution is sufficiently steep. Because of points (*a*) and (*b*), unless bunching responses are very large and the density is steep, the interior shift will have little impact on observed bin counts in a range above the threshold.[16,17]

Third, notches may create extensive margin responses above the threshold. Such responses shift down the observed distribution in the upper bracket, making missing mass above $z^*$ larger than excess bunching at $z^*$. In this case, estimating the counterfactual using bins above $z^+$ does not represent the full counterfactual stripped of all behavioral responses to the notch. As clarified by Kleven & Waseem (2013) and discussed above, correctly estimating the intensive margin response relies on an estimate of the partial counterfactual stripped of intensive responses only. This partial counterfactual should ensure that excess mass at the notch equals missing mass above the notch (even though this is not a feature of the full counterfactual) and be smooth at the threshold $z^*$, both of which are guaranteed by the approach described above. The smoothness of the partial counterfactual at $z^*$ follows from the insight that extensive responses converge to zero as we approach the threshold, an implication of a broad class of models. However, although the approach based on Equation 17—using observations above $z^+$ that are affected by extensive responses—is not conceptually wrong, in practice it is harder to obtain a robust estimate of the

[15] Diamond & Persson (2016) extend this method to estimate both the lower and upper bounds of the excluded range by making a parametric assumption on the counterfactual distribution (log-concavity) in addition to imposing the condition that bunching mass equals missing mass.

[16] Following Chetty et al. (2011), the standard approach to dealing with this issue has been to assume that the observed distribution is a downward shift (as opposed to a left shift) of the counterfactual within the estimation range above $z^*$ [as opposed to the full upper bracket $(z^*, \infty)$], thus estimating Equation 17 based on an upward adjustment of $c_j$ above $z^*$ that ensures that the counterfactual and observed distributions integrate to the same number in the estimation range. This approach may introduce bias, especially in relatively flat distributions in which interior responses do not affect bin counts (except at the very top of the distribution, away from the threshold being analyzed). It would be feasible to implement a conceptually more satisfying approach that does not have this potential bias, but for the reasons stated above, it will matter very little in most applications.

[17] These arguments have implications for the discussion in Section 3.1 regarding the distinction between different counterfactual scenarios in which bunchers come from either above the threshold (if the counterfactual features the lower-bracket tax rate throughout) or below the threshold (if the counterfactual features the higher-bracket tax rate throughout). The arguments provided here imply that, in most applications, the choice of which counterfactual benchmark to consider will have very little impact on the estimation.



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

counterfactual distribution if extensive responses are strong (as discussed in Kopczuk & Munroe 2015 and Best & Kleven 2016). In such cases, an alternative to the baseline specification is one in which the counterfactual and bunching are estimated using only data below the notch (i.e., below $z_-$).

Fourth, in some settings, there may be round-number bunching, a possibility discussed above in the context of reference points. In this case, the smooth counterfactual obtained from Equation 17 is imprecise and—if the kink or notch is itself located at a round number—biased. To solve this problem, Kleven & Waseem (2013) extend the baseline specification with round-number fixed effects, identified off of round numbers that are not kinks or notches. These round-number fixed effects should be flexible enough to account for the anatomy of rounding in the data (i.e., the fact that some round numbers are rounder than others and therefore associated with stronger bunching).

Finally, the estimation approach described above is a minimalist approach, requiring only a single cross section of data, and it may not be compelling in all contexts. A requirement for the approach to work is that the (intensive margin) distortions created by kinks and notches are very local, so that the extrapolation of the fitted distribution is done over a relatively small range. In situations in which the approach is not compelling, more sophisticated alternatives exist that require richer data or richer variation. It is worth considering two such alternatives.

First, if there is time variation in the size of the kink or notch, it is possible to identify the behavioral response based on the difference in bunching before versus after the tax change. Similarly, if there is cross-sectional variation in the size of the kink or notch across agents, the behavioral response can be identified from the difference in bunching across agents. Such difference-in-bunching strategies use the observed distribution under one tax regime to obtain the counterfactual distribution for the other tax regime. Brown (2013) and Best et al. (2015a) provide examples of this kind of approach. Note that a difference-in-bunching strategy rules out using the additional bunching observation to separately estimate the structural elasticity and optimization frictions, as described in the previous section. That is, the presence of an extra bunching observation can be used either to improve the estimation of the counterfactual or to estimate an additional parameter, but not both.

Second, with panel data it may be possible to use information on individual choices over time to improve the estimation of the counterfactual. Best et al. (2015b) develop a panel approach in a setting in which notches (in interest rates) apply to a state variable (mortgage debt), the evolution of which is observed in real time, and in which the notches apply at a point in time, namely at the time of remortgaging. In this case, mortgage debt is observed just before the remortgaging decision and can be used to obtain an individual-level counterfactual. Although their setting lends itself very naturally to such a panel approach, it may be possible to leverage panel data in other settings to improve the estimation.

## 4.2. Identification Assumptions and Issues

The research design described above allows researchers to go from bunching at kinks or notches to estimates of intensive margin elasticities. The approach relies on a set of identification assumptions, which I summarize here.

**4.2.1. Smoothness.** The main assumption is smoothness of the counterfactual distribution. There are two potential threats to this assumption: (*a*) If other policies change at the same threshold, then the distribution may not be smooth absent the specific policy being analyzed. In such cases, bunching represents a reduced-form response to a package of policies rather than to a



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

specific policy, making it difficult to uncover structural parameters. (*b*) If the threshold serves as a reference point, either because it is natural focal point for reasons unrelated to the policy or the policy itself makes it a reference point, then bunching confounds the incentive effect with a reference-point effect. When threats (*a*) and (*b*) are present, the general solution is to obtain additional bunching observations—for example, from variation in the size of the discontinuity over time or in the cross section—that depend on the same confounding effect.

**4.2.2. Shape of counterfactual.** Besides smoothness, the approach relies on an estimate of the specific shape of the counterfactual distribution. The approach does not require global knowledge of this shape: Only the local properties around the threshold are required. Behavioral responses are typically very local in the case of kinks, but less so in the case of notches. That is, while notches provide more powerful variation than kinks, they require one to extrapolate over a larger range when estimating the counterfactual based on Equation 17 (or to obtain the counterfactual using a different approach, as described above). Sensitivity analyses with respect to the order of the polynomial $p$ and the excluded range $[z_-, z_+]$ can reveal the robustness of the estimates.

**4.2.3. Model.** Estimating the smooth counterfactual density around the threshold is sufficient to identify reduced-form responses, but the identification of structural parameters makes it necessary to specify a model. This point is not specific to the bunching approach: Translating any reduced-form estimate—for example, based on quasi-experiments or randomized experiments—into structural parameters that may have external validity beyond the experiment requires modeling assumptions. Sections 3.1 and 3.2 consider a simple labor supply model that gives a direct mapping between the reduced-form bunching response and a structural parameter of interest, the compensated elasticity of labor supply. This is a static, frictionless model with no uncertainty and quasi-linear, isoelastic preferences. Those sections discuss the implications of relaxing some of these assumptions (such as introducing optimization frictions), but changing any of these model features will in general change the mapping between bunching and structural parameters. Recent work has begun to develop more structural bunching approaches that allow for dynamics and uncertainty, in particular Einav et al. (2015a,b) on spending responses to a health insurance kink in US Medicare and Best et al. (2015b) on debt responses to mortgage interest notches in the United Kingdom. These studies are discussed further in Section 5.

**4.2.4. Aggregation bias.** In the presence of heterogeneity in the elasticity $e$, bunching identifies the average behavioral response across different $e$ types. Hence, when the elasticity is calculated based on a bunching estimate and one of the (nonlinear) elasticity formulas derived above, the resulting estimate represents the elasticity at the average response as opposed to the average elasticity, creating potential aggregation bias. It is possible to bound such aggregation bias in the case of notches because we observe the possible range of responses. The minimum response corresponds to the dominated range (which may be zero for some notches) and represents an elasticity of $e_{\min} = 0$. The maximum response corresponds (roughly) to the point of convergence between the actual and counterfactual distributions, represented by an elasticity of $e_{\max} \gg 0$. The largest possible variance in responses (and thus the largest possible aggregation bias) occurs when the estimated average response is generated from these minimum and maximum responses, with appropriately chosen population weights. An upper bound on aggregation bias can therefore be obtained by calculating the average elasticity based on $e_{\min}$ and $e_{\max}$ using the same population weights. Best et al. (2015b) conduct such an exercise and find that the potential aggregation bias is very small.

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

## 5. BUNCHING APPLICATIONS

### 5.1. Tax Policy and Tax Enforcement

I start by discussing a set of tax applications, for which the bunching approach was initially developed.[18] The first cohort of bunching papers produced two basic insights that have guided subsequent research in the area.

The first insight is that, in contexts in which bunching is likely to require real earnings responses, the observed amount of bunching is very small (or zero) in elasticity terms. This applies to wage earners whose earnings are subject to third-party reporting and therefore difficult to manipulate through evasion or avoidance activities. Indeed, Saez (2010) finds zero bunching for wage earners at the large kink points created by the US income tax schedule and earned income tax credit (EITC). Chetty et al. (2011) find visually clear bunching by wage earners at a large kink point in Denmark, but the corresponding elasticity was only 0.01 for all wage earners and 0.02 for married women. Similarly, Bastani & Selin (2014) find no bunching by wage earners at a large kink in Sweden. Additionally, bunching estimates of real earnings elasticities tend to be small even when they are based on notch points that create much stronger incentives than kink points. Even though the absolute amount of bunching is generally much larger and sharper at notch points, it is still modest in elasticity terms, as documented by Kleven & Waseem (2013) for formal wage earners in Pakistan and by Kleven et al. (2014) for high-income foreign employees in Denmark.

The second insight is that, in contexts in which evasion and avoidance responses are feasible, observed bunching can be large in elasticity terms. Saez (2010), Chetty et al. (2011), Kleven & Waseem (2013), and Bastani & Selin (2014) find much larger bunching for self-employed individuals than for wage earners, consistent with the larger scope for evasion and avoidance among the self-employed (and also with smaller adjustment costs in labor supply). Kleven et al. (2011) estimate the evasion channel using the difference in bunching before and after randomized tax audits. They focus on two bunching contexts that feature easily available evasion and avoidance opportunities, namely bunching by self-employed individuals at the top income tax kink in Denmark and bunching by stockholders at a kink point in a separate stock income tax. Although they do find evidence of evasion-driven bunching, most of the observed bunching can be explained by legal tax avoidance. In particular, bunching in the stock income tax implies a huge elasticity of 2.2, almost all of which can be attributed to avoidance such as intertemporal shifting of dividends in closely held corporations. Le Maire & Schjerning (2013) analyze the importance of intertemporal income shifting for bunching by self-employed individuals in Denmark. Extending the static bunching model presented in Section 3.1 to allow for intertemporal shifting, they estimate that more than half of the elasticity for the self-employed represents income shifting.

These insights have led to two developments in the literature. One is to directly study the role of optimization frictions such as adjustment and information costs in shaping behavioral

[18] Although the modern bunching approach originates with Saez (2010), the very first tax-bunching paper is arguably by Slemrod (1985), who exploits the fact that US tax authorities use tax tables to ascertain tax liability in $50 income bins, implying that actual tax liability is a step function of taxable income rather than the continuous function implied by the statutory tax code. In other words, there are (tiny) notches at every $50 multiple of taxable income, creating an incentive to reduce income (most likely through underreporting) from just above to just below one of these income thresholds. Indeed, Slemrod (1985) finds that there is a disproportionate amount of taxpayers located within the top $10 of these $50 brackets.

responses to tax-transfer incentives; the other is to study evasion and avoidance responses to tax and enforcement policies. I now discuss these two developments in turn.

**5.1.1. Optimization frictions.** A general insight from the literature is that bunching is larger (in elasticity terms) when the kink or notch is larger, when it is more salient, and when it is stable over time. This provides prima facie evidence that bunching is shaped by optimization frictions. Chetty et al. (2011) study such frictions qualitatively, arguing that the anatomy of bunching in Denmark is consistent with the predictions of a model in which workers face hours constraints within jobs along with search costs of switching jobs. The two key predictions of such a model are a size prediction (larger kinks generate larger elasticities) and a scope prediction (kinks that affect a larger number of workers generate larger elasticities), with the latter reflecting that firms or unions tailor wage-hours packages to match aggregate worker preferences and thus create more aggregate bunching at common kinks.

Moving beyond these qualitative insights, some papers quantify the impact of optimization frictions and estimate the underlying structural parameters that govern behavior without frictions. This research has pursued the different approaches described in Section 3.3, including the notch approach by Kleven & Waseem (2013) and the parametric kink approaches by Chetty et al. (2010), Chetty (2012), and Gelber et al. (2014). This body of work argues that structural elasticities may be larger than observed bunching-based elasticities by an order of magnitude. For example, using the mass of individuals observed in dominated regions above notches, Kleven & Waseem (2013) estimate that about 90% of workers do not adjust labor supply due to some form of optimization friction. Using the same approach in a different context—value-added tax notches in the United Kingdom—Liu & Lockwood (2015) find something similar: Almost 90% of firms do not adjust turnover due to optimization frictions. These findings imply that, if not for frictions, bunching at notches would be 10 times larger than observed.[19]

Motivated by the idea that bunching is inversely related to optimization frictions such as inattention and misperception, Chetty et al. (2013) use variation in bunching across neighborhoods to proxy for differences in knowledge about tax incentives. Specifically, they study earnings responses to the EITC in the United States using bunching by self-employed individuals at the first kink of the EITC as their measure of knowledge about incentives. Because responding to a policy requires knowing about the policy, they identify the impact of the EITC on wage earnings by comparing neighborhoods that differ in the degree of bunching by the self-employed ("knowledge") at the first EITC kink. To avoid confounding effects from omitted variables that vary across low-bunching and high-bunching neighborhoods, they leverage their measure of knowledge against another source of variation, namely that EITC eligibility depends on the presence of children. Based on an event study approach, they consider changes in earnings around the birth of the first child in high- versus low-knowledge neighborhoods. Their findings suggest substantial intensive margin earnings responses to EITC incentives, conditional on knowledge. The bunching design by Chetty et al. (2013) is conceptually different from those discussed above: Rather than using

[19]A different way to gauge the impact of frictions in attenuating bunching is to compare the kink-based elasticity of 0.01 by Chetty et al. (2011) to the reform-based elasticity of 0.2 by Kleven & Schultz (2014), both of them based on Danish register data and the same measure of third-party reported earnings. Using a large tax reform in Denmark in the late 1980s, the difference-in-differences estimate by Kleven & Schultz (2014) identifies the elasticity from variation across tax brackets over time and is arguably less sensitive to the types of adjustment and attention costs that may affect bunching. As discussed by Kleven & Schultz (2014), despite some differences in time period and sample, it is difficult to explain the large difference in estimates by anything other than optimization frictions.

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

bunching to directly elicit a behavioral elasticity, they essentially use it as an instrument for the perceived EITC incentives.

**5.1.2. Evasion and enforcement.** The study of tax evasion using bunching can be divided into three categories. The first category uses bunching created by discontinuities in tax rates—in general conflating real and evasion/avoidance responses—and then leverages variation in enforcement parameters to elicit evasion responses based on differences in bunching. The first paper following such a strategy was Kleven et al. (2011) discussed above, studying the effect of randomized tax audits on bunching around tax kinks. Another paper in this category is by Fack & Landais (2016), who study the relationship between bunching at kink points and the introduction of third-party reporting on deductions for charitable contributions in France.

The second category exploits discontinuities directly in tax enforcement. Almunia & Lopez-Rodriguez (2015) study an enforcement notch created by a large taxpayers unit in Spain. Above a threshold for firm turnover, tax authorities devote larger resources to enforcement (more audits and better audits), which creates bunching below the threshold and missing mass above. Furthermore, variation in bunching across sectors that differ with respect to the prevalence of third-party information trails is used to gauge the interaction between audit effects and information trails. Dwenger et al. (2016) study a randomized field experiment in which they introduce notched audit probabilities as well as standard uniform audit probabilities. Although randomization is in principle sufficient for identification without the additional quasi-experimental variation from notches, in practice field experiment studies have struggled to find significant effects of randomized (uniform) audit threats. Dwenger et al. (2016) find that notched audit probabilities create effects that are larger and more statistically significant than uniform audit probabilities.

The third category exploits discontinuous changes in tax bases that vary with respect to the scope for evasion. Best et al. (2015a) develop a bunching approach based on minimum tax schemes in which agents are taxed on either base $z_1$ or base $z_2$ (at the tax rates $\tau_1$ or $\tau_2$, respectively) depending on which tax liability is larger. That is, tax liability is given by $T = \max\{\tau_1 z_1, \tau_2 z_2\}$, which implies a discrete jump in both the tax rate and tax base where $z_1/z_2$ crosses the threshold $\tau_2/\tau_1$. Because tax liability is continuous under such a scheme, the threshold $\tau_2/\tau_1$ represents a kink point rather than a notch point. Versions of such minimum tax schemes are ubiquitous around the world. Best et al. (2015a) specifically consider a minimum tax scheme in Pakistan in which corporations are taxed either on profits or on turnover (with a lower rate applying to the broader turnover base). They show that this kink creates potentially strong compliance incentives, but only weak real incentives at the firm level, enabling them to use bunching at the minimum tax kink to elicit evasion responses to switches between profit and turnover taxes. They estimate that turnover taxes reduce evasion by up to 60–70% of corporate income as compared to profit taxes.[20] The methodology and qualitative findings by Best et al. (2015a) are replicated for Hungary by Mosberger (2015).

## 5.2. Other Policies and Prices

Although the bunching approach was originally developed in the context of taxation and has mostly been applied there, versions of the approach are beginning to find applications in many

[20] In a similar spirit, Bachas & Soto (2016) use corporate tax notches in Costa Rica to estimate a corporate evasion rate equal to 70% of profits, with most of the effect coming through over-reporting of costs rather than through under-reporting of turnover.



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

other settings that feature kinks or notches. The settings that have been studied include pensions (Brown 2013, Manoli & Weber 2015), social insurance (Persson 2014; Einav et al. 2015a,b; Le Barbanchon 2016), welfare programs (Yelowitz 1995, Camacho & Conover 2011), education (Dee et al. 2011, Brehm et al. 2015, Diamond & Persson 2016), labor regulation (Garicano et al. 2013, Gourio & Roys 2014), minimum wages (Harasztosi & Lindner 2015), fuel economy policy (Sallee & Slemrod 2012, Ito & Sallee 2015), electricity prices (Ito 2014), cellular service prices (Grubb & Osborne 2015), and mortgage interest rates (Best et al. 2015b, DeFusco & Paciorek 2016).

Two general points are worth noting about this range of applications. First, many of these applications are based on notches rather than on kinks. Indeed, once we move beyond the study of progressive tax-transfer schedules, notches seem to be more commonly observed than kinks. Second, a number of the applications exploit discontinuities in private sector incentive schemes, as opposed to discontinuities created by government policies. That the bunching approach may allow for causal identification based on observational (equilibrium) variation in privately set prices is a potentially important advantage for the applicability of the approach.

Rather than discussing all the applications listed above, I focus on two that develop the bunching approach in a more structural and dynamic direction. The first application is provided by Einav et al. (2015a,b), who study drug demand responses to prices using a kink point created by the so-called donut hole in US Medicare. Instead of applying the simple Saez (2010) estimator, they specify and structurally estimate a dynamic model that allows for uncertainty and frictions (in the form of lumpiness in spending). Making parametric assumptions, they estimate the model so as to fit the reduced-form bunching pattern along with other moments of the data. Based on the structural estimation, they are able to make out-of-sample predictions of the spending responses to alternative health insurance designs. Einav et al. (2015a) compare this structural bunching approach to the simpler Saez approach in terms of their out-of-sample predictions and show that the two approaches—both of them consistent with the reduced-form bunching pattern in the data—can produce very different out-of-sample predictions due to their different assumptions about frictions and uncertainty.

Why is the out-of-sample prediction sensitive to the underlying model of Einav et al. (2015a,b)? The reason is closely related to the discrepancy between observed and structural elasticities—or between micro and macro elasticities—analyzed in the earlier bunching literature (e.g., Chetty et al. 2010, Chetty 2012, Kleven & Waseem 2013, Gelber et al. 2014). Although these analyses were framed in terms of different elasticity concepts, they could alternatively have been framed in terms of the sensitivity of out-of-sample predictions to modeling assumptions. For example, Chetty et al. (2010) convert a Saez-style elasticity of 0.01 into a structural elasticity of 0.34 by specifying a parametric model with frictions (hours constraints and search costs) and estimating a system such as that in Equations 14 and 15 using two bunching moments. For Einav et al. (2015a,b), the frictions are lumpiness and uncertainty, which make their structural elasticity much larger than what would otherwise be implied by the modest amount of observed bunching. At a higher level, these findings hark back to the nonlinear budget set approach in which researchers obtained large structural elasticities despite the absence of bunching by allowing for optimization error or by smoothening the kink.[21]

[21] Also related to these ideas, the early working paper versions of Saez's bunching paper (Saez 1999, 2002) show that the absence of bunching at US tax kinks can be reconciled with an elasticity of 0.5 by allowing for either uncertainty in income (which effectively smoothens the kink) or by hours constraints in which workers choose between a few discrete options (lumpiness). In the language used above, this corresponds to saying that the observed elasticity is zero, whereas the structural elasticity when allowing for randomness or lumpiness may be as high as 0.5 (or higher if allowing for both frictions simultaneously), or similarly that the out-of-sample prediction is sensitive to the assumed model.

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

The second application is provided by Best et al. (2015b), who study the response of household debt and intertemporal consumption allocation to interest rates using mortgage notches in the United Kingdom. Based on bunching at interest rate notches, they estimate both reduced-form mortgage demand elasticities (à la Saez-Kleven-Waseem) and the underlying structural elasticities of intertemporal substitution. The structural estimation requires them to specify a dynamic model and make a set of parametric assumptions. They show that the estimation is very robust to a wide range of assumptions about uncertainty, risk aversion, discount factors, present bias, and beliefs about the future. A key reason that their notch-based findings are robust (whereas the aforementioned kink-based findings are sensitive) is that they directly estimate the amount of friction from the observed density mass in dominated regions above notches, as in Kleven & Waseem (2013). That is, when Best et al. (2015b) change modeling assumptions, the friction is always accounted for using the Kleven-Waseem friction adjustment. Hence, they are not moving between elasticity concepts with and without friction when comparing different models. Even in their setting, if the actual amount of friction were much larger than what is captured by the Kleven-Waseem adjustment, then the true structural elasticity would be much larger as well. In this sense, bunching-based elasticities are always sensitive to what is assumed about optimization frictions.

### 5.3. Behavioral Kinks and Notches: Reference Points and Norms

The applications described above consider settings featuring discontinuities in extrinsic incentives such as taxes, enforcement, and social insurance. Section 3.3 discusses reference dependence as a potential confounder when mapping a reduced-form bunching pattern into a structural price elasticity, for example, when a kink or a notch is located at a salient round number and therefore subject to round-number bunching (Kleven & Waseem 2013, Best & Kleven 2016). Besides posing an identification problem in some settings, reference dependence is an interesting behavioral phenomenon, and the bunching approach offers a way to study it outside the lab. In particular, when we observe sharp bunching at points where extrinsic incentives are smooth, this indicates some form of reference dependence (i.e., an intrinsic or psychological discontinuity). As described above, theories of reference dependence can be understood using the language of kinks and notches and potentially analyzed using the techniques laid out here.

Several recent papers use bunching to study reference dependence and norms. For example, Allen et al. (2016) provide evidence of bunching just below round numbers (and missing mass just above) in marathon finishing times such as 3:00 and 4:00 hours. Absent any discontinuities in extrinsic rewards at those finishing times, they interpret this as evidence of reference-dependent preferences. Based on calibration exercises, they show that the reduced-form bunching pattern is consistent with a model of prospect theory (loss aversion). As explained in Section 3.3, the reason why prospect theory can produce a combination of bunching and holes around reference points—despite modeling these points as kinks—has to do with the theory's assumption of diminishing sensitivity, a discontinuity in the second derivative of utility. Alternatively, if reference dependence is modeled simply as a notch in utility, the reduced-form pattern can be explained without having to invoke the second derivative.

The evidence on marathon runners is consistent with evidence on round-number bunching in a wide range of other settings. Round-number effects have been documented in the contexts of baseball batting averages and students' test scores (Pope & Simonsohn 2011), odometer mileage in used-car purchases (Lacetera et al. 2012, Busse et al. 2013), reported taxable income (Kleven & Waseem 2013), and house prices (Pope et al. 2015, Best & Kleven 2016). The exact reason why round numbers serve as reference points may vary across settings—including goal setting and



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

limited attention—but in general their modeling requires some form of intrinsic or behavioral discontinuity in the objective that agents are maximizing. However, once the behavioral discontinuity is there, this may give rise to confounding extrinsic incentives in equilibrium. For example, in Lacetera et al. (2012), the preference for used cars with odometer values just below 10,000 mile thresholds affects car prices just below those thresholds and therefore creates an extrinsic incentive as well.

Related to the last point, the convention in statistical testing of defining threshold significance levels such as 5% and 10%—arbitrary round-number reference points—creates extrinsic incentives for researchers to pick marginally significant specifications over insignificant ones. Studying the distributions of test statistics, Gerber & Malhotra (2008), Simonsohn et al. (2014), and Brodeur et al. (2016) document the existence of such distortions. Although these authors do not draw a link between their ideas and the bunching approach, there is a close relationship between the two. For example, Brodeur et al. (2016) show evidence of excess mass just below 5% in the distribution of $p$ values along with missing mass between 10% and 25%. They find no evidence of either excess mass or missing mass in the 5–10% range, which may be the result of two offsetting forces: The desire to move below the 5% notch creates missing mass between 5% and 10%, whereas the desire to move below the 10% notch creates excess mass between 5% and 10%.

Finally, I describe a number of contexts in which bunching at reference points goes beyond round-number effects. First, Rees-Jones (2014) argues that loss aversion affects tax avoidance behavior when filing tax returns in the United States. Specifically, when taxpayers file returns at the end of the tax year, they receive refunds (gains) or incur back payments (losses) depending on whether the balance due is negative or positive given realized taxable income and tax withholding during the year. Although the marginal tax rate does not change around a zero balance due, the observed distribution of balances features excess bunching at zero. This is consistent with a loss aversion model in which taxpayers frame back payments as losses that are associated with a discretely larger marginal disutility of lost income. In this setting, a zero balance due serves as a reference point for tax filing behavior.

Second, considering a German church tax that relies on voluntary compliance—underpayments are never penalized and overpayments are not reimbursed—Dwenger et al. (2016) document sharp bunching at the point of exact compliance. Absent any extrinsic incentive to pay taxes, they argue that this empirical pattern requires either a discontinuity in intrinsic motivation at the point of exact compliance, what they label "duty to comply," or the presence of attention/salience effects of exact compliance. Leveraging a randomized field experiment, they show that making the point of exact compliance more salient does not affect the observed bunching pattern, so it may be driven by duty to comply. In any case, this is an example in which the letter of the law, even though it provides no explicit incentive, comes to serve as a reference point for the behavior of intrinsically motivated taxpayers.

Third, Bertrand et al. (2015) advance the idea that gender identity norms prescribe that men should make more money than women within households (i.e., a "male-breadwinner norm"). Consistent with such a norm, they document a sharp cliff at the 50% threshold in the distribution of female income shares in US households. The implication of such a cliff seems to be the existence of a notch in household preferences at a 50% female income share—a breadwinner notch—which would naturally produce excess bunching below 50%. However, the distributions presented by Bertrand et al. (2015) do not feature bunching, nor do they take the logical step of analyzing the breadwinner norm as a notched incentive. Based on these ideas, Kleven et al. (2016) analyze the breadwinner notch using Danish administrative data. They document excess bunching below 50% and missing mass above 50% in the distribution of female income shares, and convert this reduced-form pattern into a money metric of the male-breadwinner norm. This exercise consists of

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

representing the breadwinner notch as an implicit tax on female earnings above the 50% reference point, and then inverting the bunching approach laid out in Section 3: Instead of estimating an elasticity parameter based on observed bunching and a known tax incentive, the idea is to estimate the implicit tax notch based on observed bunching and an elasticity parameter that has to be calibrated or estimated using another source of variation. The idea of using bunching at reference points to estimate money metrics of their importance is potentially feasible in other settings.

## 6. A BROADER PERSPECTIVE

In this article, I have reviewed the bunching approach to empirical research, discussed its main identification assumptions and challenges, and considered a range of applications in public finance and other fields. The approach is still relatively young, with its recent popularity closely linked to the explosion of work using administrative data.[22] Due to the ubiquity of kinks and notches across a wide set of contexts and the increasing availability of administrative data, the approach is likely to find numerous applications and evolve in new directions in the coming years. It is possible that the context in which the approach was first conceived—estimating the real labor supply elasticity—is not the ideal application in light of the challenges discussed here. The presence of optimization frictions and reference dependence makes the observed elasticity very context specific, and without knowing the right model of these behavioral aspects, it is difficult to link the observed elasticity to a stable, structural parameter.[23] I have considered a set of other decision environments—in particular, environments in which the decision variable can be continuously adjusted without much friction—in which the approach may be better suited to uncover structural parameters and sufficient statistics for welfare analysis.

Up to this point, I have focused on bunching as an empirical strategy, taking as given that kinks and notches exist in the real world. A conceptually different set of questions relates to whether kinks or notches are socially desirable and why they are used so frequently. In standard mechanism design models, kinks may be a feature of an optimal incentive scheme—and if not, they generally represent a reasonable approximation of the fully optimal, continuously differentiable scheme—whereas notches are typically ruled out as a property of optimal incentive schemes (Mirrlees 1971). Given that notches are ubiquitous in practice, it is natural to ask if those who implement them make bad decisions or if our models miss relevant features of actual decision environments. The key assumptions that make notches suboptimal is that the underlying fundamentals are continuous, that agents are sophisticated optimizers, and that the available policy instruments are sufficiently flexible. If we break any of those assumptions, notches may become optimal.

The role of constrained policy instruments for the desirability of notches is discussed by Slemrod (2010). Such constraints are (implicitly) the reason that notches can be optimal in the first theoretical treatment of the problem by Blinder & Rosen (1985), a comparison between a notched tax incentive and a linear tax incentive. Gillitzer et al. (2016) explicitly ground the desirability of tax notches in restrictions on the set of policy instruments in their analysis of optimal line drawing in taxation. Although natural restrictions on the policy space may be sufficient reason for notches in some contexts, in other context notches are clearly dominated by feasible continuous

[22] It is striking that, among the roughly 80 papers cited in this review, approximately 60 of them are dated 2010 or later.

[23] As discussed, this limitation is stronger for kinks than for notches. This is the case both because notches create stronger incentive changes (and are thus more likely to overcome optimization frictions) and because notches offer less parametric ways of controlling for any remaining attenuation bias from friction.



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

policies. Examples in which the case for notches is not compelling include the various tax contexts considered by Sallee & Slemrod (2012), Kleven & Waseem (2013), and Best & Kleven (2016).[24]

Another possible reason for notches is that individuals are not the sophisticated optimizers assumed in our canonical models, and specifically that they find discrete categories simpler and more intuitive than a continuum. Although the idea of continuous variables may be second nature to economists and other mathematically inclined people (see, e.g., Dawkins 2011), to most people discrete classification comes more naturally.[25] The discussion of behavioral notches or reference points in the previous section suggests that there is a demand for notches: Even when discrete categories or notches are not imposed by policy makers, individuals tend to create them by dividing continuous space into discrete categories that separate high and low, success and failure, gains and losses, etc. Even scientists do this with their arbitrarily chosen thresholds that separate statistically significant and insignificant (e.g., Brodeur et al. 2016) when they could just consider the *p* value as a continuous measure of statistical precision. Whenever such thresholds are created, there is a notched incentive that could be analyzed using the techniques laid out here.

## DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENT

I thank Miguel Almunia, Raj Chetty, Stefano DellaVigna, Chuck Manski, Emmanuel Saez, and Joel Slemrod for helpful comments and discussions.

## LITERATURE CITED

Allen E, Dechow P, Pope D, Wu G. 2016. Reference-dependent preferences: evidence from marathon runners. *Manag. Sci.* In press

Almunia M, Lopez-Rodriguez D. 2015. *Under the radar: the effects of monitoring firms on tax compliance*. Work. Pap., Univ. Warwick

Bachas P, Soto M. 2016. *Not(ch) your average tax system: corporate taxation under weak enforcement*. Work. Pap., Univ. Calif., Berkeley

Bastani S, Selin H. 2014. Bunching and non-bunching at kink points of the Swedish tax schedule. *J. Public Econ.* 109:36–49

Bertrand M, Kamenica E, Pan J. 2015. Gender identity and relative income within households. *Q. J. Econ.* 130:571–614

Best M, Brockmeyer A, Kleven H, Spinnewijn J, Waseem M. 2015a. Production vs revenue efficiency with limited tax capacity: theory and evidence from Pakistan. *J. Polit. Econ.* 123:1311–55

[24] In fact, in the two contexts analyzed by Kleven & Waseem (2013) and Best & Kleven (2016)—income tax notches in Pakistan and transaction tax notches in the United Kingdom, respectively—policy makers have subsequently replaced the notches by continuous kinked tax schedules (apparently using the analyses in these papers as underpinnings for the policy reforms).

[25] In his essay "The Tyranny of the Discontinuous Mind," Dawkins (2011) discusses (criticizes) the widespread use of discrete classifications in a variety of settings, ranging from relatively small topics (such as poverty lines, exam results, and voting age) to very big topics (such as whether a human embryo is counted as a person, and the discrete classification between racial/ethnic groups and between human and nonhuman species), arguing that all of these things are part of a continuous process and should be considered as such.

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

Best M, Cloyne J, Ilzetzki E, Kleven H. 2015b. *Interest rates, debt and intertemporal allocation: evidence from notched mortgage contracts in the UK*. Work. Pap., London School Econ.

Best M, Kleven H. 2013a. *Housing market responses to transaction taxes: evidence from notches and stimulus in the UK*. Work. Pap., London School Econ.

Best M, Kleven H. 2013b. *Optimal income taxation with career effects of work effort*. Work. Pap., London School Econ.

Best M, Kleven H. 2016. *Housing market responses to transaction taxes: evidence from notches and stimulus in the UK*. Work. Pap., London School Econ.

Blinder AS, Rosen HS. 1985. Notches. *Am. Econ. Rev.* 75:736–47

Blundell R, MaCurdy T. 1999. Labor supply: a review of alternative approaches. In *Handbook of Labor Economics*, Vol. 3, ed. O Ashenfelter, D Card, pp. 1559–695. Amsterdam: North-Holland

Brehm M, Imberman S, Lovenheim M. 2015. *Achievement effects of individual performance incentives in a teacher merit pay tournament*. NBER Work. Pap. 21598

Brodeur A, Le M, Sangnier M, Zylberberg Y. 2016. Star Wars: The empirics strike back. *Am. Econ. J. Appl. Econ.* 8(1):1–32

Brown KM. 2013. The link between pensions and retirement timing: lessons from California teachers. *J. Public Econ.* 98:1–14

Burtless G, Hausman JA. 1978. The effect of taxation on labor supply: evaluating the Gary negative income tax experiment. *J. Polit. Econ.* 86:1103–30

Burtless G, Moffitt R. 1984. The effect of social security benefits on the labor supply of the aged. In *Retirement and Economic Behavior*, ed. HJ Aaron, G Burtless, pp. 135–74. Washington, DC: Brookings Inst.

Busse M, Lacetera N, Pope D, Silva-Risso J, Sydnor J. 2013. Estimating the effect of salience in wholesale and retail car markets. *Am. Econ. Rev.* 103:575–79

Camacho A, Conover E. 2011. Manipulation of social program eligibility. *Am. Econ. J. Econ. Policy* 3(2):41–65

Card D, Lee DS, Pei Z, Weber A. 2015. Inference on causal effects in a generalized regression kink design. *Econometrica* 83:2453–83

Chetty R. 2012. Bounds on elasticities with optimization frictions: a synthesis of micro and macro evidence on labor supply. *Econometrica* 80:969–1018

Chetty R, Friedman JN, Olsen T, Pistaferri L. 2010. *Adjustment costs, firm responses, and labor supply elasticities: evidence from Danish tax records*. NBER Work. Pap. 15617

Chetty R, Friedman JN, Olsen T, Pistaferri L. 2011. Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: evidence from Danish tax records. *Q. J. Econ.* 126:749–804

Chetty R, Friedman JN, Saez E. 2013. Using differences in knowledge across neighborhoods to uncover the impacts of the EITC on earnings. *Am. Econ. Rev.* 103:2683–721

Dawkins R. 2011. The tyranny of the discontinuous mind. *New Statesman*, Dec. 19. **http://www.newstatesman.com/blogs/the-staggers/2011/12/issue-essay-line-dawkins**

Dee T, Jacob B, Rockoff J, McCrary J. 2011. *Rules and discretion in the evaluation of students and schools: the case of the New York Regents Examinations*. Work. Pap., Columbia Univ., New York

DeFusco A, Paciorek A. 2016. The interest rate elasticity of mortgage demand: evidence from bunching at the conforming loan limit. *Am. Econ. J. Econ. Policy*. In press

Devereux MP, Liu L, Loretz S. 2014. The elasticity of corporate taxable income: new evidence from UK tax records. *Am. Econ. J. Econ. Policy* 6(2):19–53

Diamond R, Persson P. 2016. *The long-term consequences of teacher discretion in grading of high-stakes tests*. Work. Pap., Stanford Univ., Stanford, CA

Dwenger N, Kleven H, Rasul I, Rincke J. 2016. Extrinsic and intrinsic motivations for tax compliance: evidence from a field experiment in Germany. *Am. Econ. J. Econ. Policy*. 8:203–32

Einav L, Finkelstein A, Schrimpf P. 2015a. *Bunching at the kink: implications for spending responses to health insurance contracts*. Work. Pap., Mass. Inst. Technol., Cambridge, MA

Einav L, Finkelstein A, Schrimpf P. 2015b. The response of drug expenditure to non-linear contract design: evidence from Medicare Part D. *Q. J. Econ.* 130:841–99

Fack G, Landais C. 2016. The effect of tax enforcement on tax elasticities: evidence from charitable contributions in France. *J. Public Econ.* 133:23–40

Friedberg L. 1998. The Social Security earnings test and labor supply of older men. *Tax Policy Econ.* 12:121–50
Friedberg L. 2000. The labor supply effects of the Social Security earnings test. *Rev. Econ. Stat.* 82:48–63
Garicano L, Lelarge C, Van Reenen J. 2013. *Firm size distortions and the productivity distribution: evidence from France*. Discuss. Pap. 7241, IZA, Bonn, Ger.
Gelber A, Jones D, Sacks DW. 2014. *Earnings adjustment frictions: evidence from the Social Security earnings test*. Work. Pap., Univ. Calif., Berkeley
Gerber A, Malhotra N. 2008. Publication bias in empirical sociological research. *Sociol. Methods Res.* 37:3–30
Gillitzer C, Kleven H, Slemrod J. 2016. A characteristics approach to optimal taxation: line drawing and tax-driven product innovation. *Scand. J. Econ.* In press
Gourio F, Roys N. 2014. Size-dependent regulations, firm size distribution, and reallocation. *Quant. Econ.* 5:377–416
Grubb M, Osborne M. 2015. Cellular service demand: biased beliefs, learning, and bill shock. *Am. Econ. Rev.* 105:234–71
Harasztosi P, Lindner A. 2015. *Who pays for the minimum wage?* Work. Pap., Univ. Coll. London
Hausman JA. 1981. Labor supply. In *How Taxes Affect Economic Behavior*, ed. HJ Aaron, JA Pechman, pp. 27–72. Washington, DC: Brookings Inst.
Hausman JA. 1983. Stochastic problems in the simulation of labor supply. In *Behavioral Simulations in Tax Policy Analysis*, ed. M Feldstein, pp. 47–69. Chicago, IL: Univ. Chicago Press
Heckman J. 1983. Comment. In *Behavioral Simulations in Tax Policy Analysis*, ed. M Feldstein, pp. 70–82. Chicago, IL: Univ. Chicago Press
Imbens GW, Lemieux T. 2008. Regression discontinuity designs: a guide to practice. *J. Econ.* 142:615–35
Ito K. 2014. Do consumers respond to marginal or average price? Evidence from nonlinear electricity pricing. *Am. Econ. Rev.* 104:537–63
Ito K, Sallee J. 2015. *The economics of attribute-based regulation: theory and evidence from fuel-economy standards*. NBER Work. Pap. 20500
Kahneman D, Tversky A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47:263–92
Kleven H, Knudsen M, Kreiner C, Pedersen S, Saez E. 2011. Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica* 79:651–92
Kleven H, Landais C, Saez E, Schultz E. 2014. Migration and wage effects of taxing top earners: evidence from the foreigners' tax scheme in Denmark. *Q. J. Econ.* 129:333–78
Kleven H, Landais C, Søgaard J. 2016. *The breadwinner notch: a bunching approach to estimating reference-dependent preferences*. Unpublished manuscript, London School Econ.
Kleven H, Schultz E. 2014. Estimating taxable income responses using Danish tax reforms. *Am. Econ. J. Econ. Policy* 6(4):271–301
Kleven H, Waseem M. 2012. *Behavioral responses to notches: evidence from Pakistani tax records*. Work. Pap., London School Econ.
Kleven H, Waseem M. 2013. Using notches to uncover optimization frictions and structural elasticities: theory and evidence from Pakistan. *Q. J. Econ.* 128:669–723
Kopczuk W, Munroe D. 2015. Mansion tax: the effect of transfer taxes on the residential real estate market. *Am. Econ. J. Econ. Policy* 7:214–57
Koszegi B, Rabin M. 2006. A model of reference-dependent preferences. *Q. J. Econ.* 121:1133–65
Lacetera N, Pope D, Sydnor J. 2012. Heuristic thinking and limited attention in the car market. *Am. Econ. Rev.* 102:2206–36
Le Barbanchon T. 2016. *Optimal partial unemployment insurance: evidence from bunching in the U.S.* Work. Pap., Bocconi Univ., Milan
Le Maire D, Schjerning B. 2013. Tax bunching, income shifting and self-employment. *J. Public Econ.* 107:1–18
Liu L, Lockwood B. 2015. *VAT notches*. Work. Pap., Univ. Warwick
MaCurdy T. 1981. An empirical model of labor supply in a life-cycle setting. *J. Polit. Econ.* 89:1059–85
MaCurdy T, Green D, Paarsch H. 1990. Assessing empirical approaches for analyzing taxes and labor supply. *J. Hum. Resour.* 25:415–90
Manoli D, Weber A. 2015. *Nonparametric evidence on the effects of financial incentives on retirement decisions.* Work. Pap., Univ. Mannheim

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

Mirrlees JA. 1971. An exploration in the theory of optimal income taxation. *Rev. Econ. Stud.* 38:175–208
Moffitt R. 1990. The econometrics of kinked budget constraints. *J. Econ. Perspect.* 4(2):119–39
Mosberger P. 2015. *Tax optimization responses to the minimum tax scheme: bunching evidence*. Work. Pap., Central Eur. Univ., Budapest
Persson P. 2014. *Social insurance and the marriage market*. Work. Pap., Stanford Univ., Stanford, CA
Pope D, Pope J, Sydnor J. 2015. Focal points and bargaining in housing markets. *Games Econ. Behav.* 93:89–107
Pope D, Simonsohn U. 2011. Round numbers as goals: evidence from baseball, SAT takers, and the lab. *Psychol. Sci.* 22:71–79
Rees-Jones A. 2014. *Loss aversion motivates tax sheltering: evidence from U.S. tax returns.* Work. Pap., Wharton School Univ. Penn., Philadelphia
Saez E. 1999. *Do taxpayers bunch at kink points?* NBER Work. Pap. 7366
Saez E. 2002. *Do taxpayers bunch at kink points?* Work. Pap., Univ. Calif., Berkeley
Saez E. 2010. Do taxpayers bunch at kink points? *Am. Econ. J. Econ. Policy* 2:180–212
Sallee JM, Slemrod J. 2012. Car notches: strategic automaker responses to fuel economy policy. *J. Public Econ.* 96:981–99
Seim D. 2015. *Behavioral responses to wealth taxes: evidence from Sweden*. Work. Pap., Stockholm Univ.
Simonsohn U, Nelson L, Simmons J. 2014. P-curve: a key to the file-drawer. *J. Exp. Psychol. Gen.* 143:534–47
Slemrod J. 1985. An empirical test for tax evasion. *Rev. Econ. Stat.* 67:232–38
Slemrod J. 2010. *Buenas notches: lines and notches in tax system design*. Work. Pap., Univ. Michigan, Ann Arbor
Yelowitz AS. 1995. The Medicaid notch, labor supply, and welfare participation: evidence from eligibility expansions. *Q. J. Econ.* 110:909–39

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

Annual Review of Economics
Volume 8, 2016

# Contents



Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.

**Indexes**



**Errata**

An online log of corrections to *Annual Review of Economics* articles may be found at http://www.annualreviews.org/errata/economics

Annu. Rev. Econ. 2016.8:435-464. Downloaded from www.annualreviews.org
Access provided by London School of Economics and Political Science on 11/28/16. For personal use only.