Treatment Effects in Bunching Designs: The Impact of Mandatory Overtime Pay on Hours

Leonard Goff*

This version: May 13, 2025

Abstract

This paper studies the identifying power of bunching at kinks when the researcher does not assume a parametric choice model. I find that in a general choice model, identifying the average causal response to the policy switch at a kink amounts to confronting two extrapolation problems, each about the distribution of a counterfactual choice that is observed only in a censored manner. I apply this insight to partially identify the effect of overtime pay regulation on the hours of U.S. workers using administrative payroll data, assuming that each distribution satisfies a weak non-parametric shape constraint in the region where it is not observed. The resulting bounds are informative and indicate a relatively small elasticity of demand for weekly hours, addressing a long-standing question about the causal effects of the overtime mandate.

*Department of Economics, University of Calgary, leonard.goff@ucalgary.ca. I thank my PhD co-advisors Simon Lee and Suresh Naidu, as well as Michael Best, Daniel Hamermesh, and Bernard Salanié for their gracious advice and support throughout this project. I have also benefited from discussions with Christopher Ackerman, Doug Almond, Joshua Angrist, Jushan Bai, Iain Bamford, Marc Bellemare, Sandra Black, Carol Caetano, Brant Callaway, Ivan Canay, Gregory Cox, Junlong Feng, Dalia Ghanem, Bhargav Gopal, Jonas Hjort, Désiré Kédagni, Pat Kline, Wojciech Kopczuk, Bentley MacLeod, Matt Masten, Whitney Newey, Serena Ng, José Luis Montiel Olea, Dilip Ravindran, Emmanuel Saez, Ian Schmutte, Miguel Urquiola, and seminar participants at Berkeley, Columbia, Duke, Lehigh, Microsoft, the University of Calgary, the University of Georgia, the University of North Carolina Greensboro, the US Census Bureau, and Washington State University, as well as audiences at the Canadian Economics Association and Southern Economics Association meetings, and the European Winter Meeting of the Econometric Society. I thank my main data provider as well as Ted To and the Bureau of Labor Statistics. Some additional results are posted as further appendices here.

1 Introduction

A major theme throughout microeconometrics is to separate causal relationships of interest from additional sources of individual heterogeneity in the outcomes observed. When that outcome represents a choice those individuals make—for example when estimating an elasticity of labor supply—a key challenge is to confront the endogeneity introduced by heterogeneity in preferences that may be correlated with prices. Familiar methods for identification leverage random variation (e.g. instrumental variables) or changes over time (e.g. policy reforms), yet many important environments lack opportunities to credibly use such tools. In these settings, new approaches to inferring the responsiveness of agents to incentives are highly valuable.

When a population of decision-makers faces choice sets that exhibit a kink at a common threshold, the popular "bunching design" method uses the cross-sectional distribution of those agents' choices to identify their responsiveness to the incentives that change at that threshold. A generic prediction of optimizing behavior is that the distribution of agents' choices will feature bunching where there are convex kinks in their costs as a function of a second choice variable. Saez (2010) observed that under suitable assumptions, the magnitude of this bunching can be informative about *how* elastic their choices are to the switch in incentives that occurs at a kink. The bunching design has since become a popular research design in a variety of settings, growing from its initial focus on measuring the elasticity of labor supply using the kink in tax liability between tax brackets.[1]

However, the literature has recently emphasized some concerning limits to non-parametric identification in the bunching design. The bunching design approach couples two essential ingredients for identification: i) a choice model; and ii) assumptions about the distribution of heterogeneity in agents' preferences. While i) describes how a given agent's choices would be made given alternative choice sets, ii) captures how different agents would choose differently even if confronted with the same choice set. Blomquist et al. (2021) and Bertanha et al. (2023) show that even if one assumes the restrictive "isoelastic" choice model typical in applications of the bunching design, identification from bunching requires assumptions on the distribution of heterogeneity that cannot be verified directly in the data.

In this paper I find that the upside of confronting this challenge to identification is quite high. In particular, I show that the bunching design remains applicable under weak structural assumptions about choice when the design is used for questions of reduced-form policy evaluation. I do this by recasting the necessary extra assumptions for identification as extrapolation assumptions about two appropriately-defined counterfactual *choices*, in the context of a general non-parametric choice model. This establishes that the essential identifying power of the bunching design does not depend on the isoelastic model from the tax literature that is typically used to motivate the approach.

[1] Kleven and Waseem (2013) pioneered a similar approach to "notches" where the level (rather than the slope) of tax liability jumps discretely at a threshold. See Kleven (2016) and Berthana et al. (2023) for reviews of related methods.

Using the language of potential outcomes, I generalize the parameter of interest beyond the isoelastic model to a local average treatment effect parameter, the "buncher ATE", which captures the mean difference between the two counterfactual choices among observational units that are bunched at the kink. These potential outcomes are directly observed in the data, though not across the full support of their distributions. I propose a new non-parametric assumption to extrapolate from the observed distribution of agents' choices and partially identify the buncher ATE. In particular, I impose a relatively weak shape constraint—*bi-log-concavity*—on the distribution of each potential outcome. Bi-log-concavity nests many previously proposed distributional assumptions for bunching analyses and is testable within the region in which each potential outcome is observed.

My results supplement other partial identification approaches recently proposed for the bunching design. Notably, the bounds I derive for the buncher ATE are substantially narrowed relative to existing approaches by making extrapolation assumptions separately for each of the *two* counterfactuals. By contrast, existing approaches constrain the distribution of a single scalar heterogeneity parameter, a simplification that is afforded by the isoelastic choice model. In the context of that model, Bertanha et al. (2023) and Blomquist et al. (2021) obtain bounds on the elasticity when the researcher is willing to put an explicit limit on how sharply the density of heterogeneous choices can rise or fall. My approach based on bi-log-concavity avoids the need to choose any such tuning parameters, and is applicable in the general choice model. However, I show how an explicit bounding approach can be utilized there as well, which in my empirical application yields similar estimates of the identified set. In the general choice model, I impose assumptions on quantile functions rather than on densities, and the "distance" one is required to extrapolate is equal to the bunching probability, a (dimensionless) quantity known from the data.

I apply the above approach to evaluate a major labor market policy that has proven difficult to assess via other research designs: the "time-and-a-half" overtime pay rule introduced by the U.S. Fair Labor Standards Act (FLSA) of 1938. The time-and-a-half requires a pay premium for long work hours: firms must pay a worker one and a half times their normal hourly wage for any hours worked in excess of 40 within a single week. Although many salaried workers are exempt from it, the time-and-a-half rule applies to a majority of the U.S. workforce, including nearly all of its over 80 million hourly workers. Workers in many industries average multiple overtime hours per week, making overtime the largest form of supplemental pay in the U.S. (Hart, 2004; Bishow, 2009).

In marked contrast to the federal minimum wage (which was also introduced by the 1938 FLSA), only a small literature has studied the effects of the FLSA overtime rule on the labor market. A key reason for this is that the overtime rule has hardly varied: the policy has remained as time-and-a-half after 40 hours in a week, for now more than 80 years. Reforms to overtime policy have been rare and have focused on eligibility, leaving the central parameters of the rule unaffected. This lack of variation has afforded few opportunities to leverage research designs that exploit policy changes to identify causal effects,[2] and remains as the Department of Labor plans a major expansion to eligibility in 2024 (U.S. Department of Labor, 2024).

By leveraging the bunching design, this paper makes use of variation *within* the overtime rule itself. With wages held constant, the policy introduces a sharp discontinuity in the marginal cost to the firm of a worker-hour—a convex "kink" in firms' costs—which provides firms with an incentive to set workers' hours exactly at 40 in a given week. I take the perspective of firms setting workers' hours in an optimizing way, which yields the implication that the mass of workers working 40 hours in a given week will be larger or smaller depending on how responsive firms are to the wage increase imposed by the time-and-a-half rule. I draw on a novel administrative dataset of the exact hours for which workers are paid in a single week, using the bunching observed at 40 hours among hourly workers in these data to assess how the FLSA has affected the hours of U.S. workers.

In the overtime setting, the potential outcomes considered by the buncher ATE correspond to, respectively: i) the number of hours the firm would choose for the worker this week if the worker's normal wage rate applied to all of this week's hours; and ii) the number that the firm would choose if the worker's overtime rate applied to all of this week's hours. The buncher ATE then reflects a local average wage elasticity of hours demand between workers' standard wage and overtime wage rates. Choice from the kinked choice set can be fully characterized by these counterfactuals: firms choose one or the other of them or they choose the location of the kink. The magnitude of bunching at 40 hours then identifies directly a feature of the joint distribution of the potential outcomes, allowing one to make statements about treatment effects purged of selection bias.[3]

However, as noted above, identification hinges crucially on extrapolation assumptions about the marginal distributions of the two potential outcomes. The bi-log-concavity assumption I rely on can be economically motivated in the case of working hours, in addition to being partially testable in the payroll data I use. The resulting bounds for the buncher ATE turn out to be quite informative. While the buncher ATE represents a local reduced-form quantity, I use it to assess the overall average effect of the FLSA by layering on additional (also non-parametric) assumptions.

I also show that the data in the bunching design are informative about counterfactual policies that change the location or "sharpness" of a kink. To do so, I extend a characterization of bunching from Blomquist et al. (2015), and show that when combined with a general continuity equation (Kasy, 2022) it yields bounds on the derivative of bunching and mean hours with respect to policy

[2] See Brown and Hamermesh (2019). A few studies that have used difference-in-differences approaches to estimating effects of U.S. overtime policy on hours: Hamermesh and Trejo (2000) consider the expansion of a daily overtime rule in California to men in 1980, while Johnson (2003) use a supreme court decision on the eligibility of public-sector workers in 1985. Costa (2000) studies the initial phase-in of the FLSA in the years following 1938. See footnotes 3 and 38 for a comparison of results to these papers. Quach (2024) looks at recent reforms to eligibility criteria for exemption from the FLSA, estimating the expansion's effects on hourly/salary classification, employment and earnings.

[3] This echoes Kline and Tartari's 2016 approach to studying labor supply, but in reverse. They use observed marginal distributions of counterfactual choices to identify features of their joint distribution, assuming optimizing behavior.

parameters. I use this to evaluate proposed reforms to the FLSA: e.g. lowering the overtime threshold below 40 hours (e.g. the *Thirty-Two Hour Workweek Act* proposed in the U.S. House of Representatives in 2021), or increasing the premium pay factor from 1.5 to 2.

The empirical setting of overtime pay involves confronting two challenges that are not typical of existing bunching-design analyses. Firstly, 40 hours is not an "arbitrary" point and bunching there could arise in part from factors other than it being the location of the kink. I use two strategies to estimate the amount of bunching that would exist at 40 absent the FLSA, and deliver clean estimates of the rule's effect. My preferred approach exploits the fact that when a worker makes use of paid-time-off hours these do not count towards that week's overtime threshold, shifting the location of the kink week-to-week in a plausibly idiosyncratic way. A second feature of the overtime setting is that work hours may not be set unilaterally by one party: in principle either the firm or the worker could choose a given worker's schedule. I provide evidence that week-to-week variation in hours is mostly driven by firms. Even if bargaining weight between workers and firms varies arbitrarily, I show that bunching at 40 hours is informative about labor demand rather than supply.

Empirically, I find that the FLSA overtime rule does in fact reduce hours of work among hourly workers, despite the theoretical possibility that offsetting wage adjustments might eliminate any such effect (Trejo, 1991). My preferred estimate suggests that about one quarter of the bunching observed at 40 among hourly workers is due to the FLSA, and those working at least 40 hours work, on average, about 30 minutes less in a week than they would absent the time-and-a-half rule. Across specifications, I obtain estimates of the local wage elasticity of weekly hours demand near 40 hours in the range -0.04 to -0.19, indicating that firms are fairly resistant to changing hours to avoid overtime payments.

The structure of the paper is as follows. Section 2 lays out a motivating conceptual framework for work hours that relates my bunching approach to existing literature on overtime policy. Section 3 introduces the payroll data I use in the empirical analysis. In Section 4 I develop the generalized bunching-design approach in the context of the overtime application, with Appendix B expanding on some of the supporting formal results and further generalizations. Section 5 applies these results to estimate effect of the FLSA overtime rule on work hours, as well as the effects of proposed reforms to the FLSA. Section 6 discusses the empirical findings from the standpoint of policy objectives, and Section 7 concludes.

2 Conceptual framework for the overtime setting

Although the main identification results of this paper are not specific to the overtime application, I will use it as a running example throughout. To set the stage, this section outlines a framework for reasoning about the determination of weekly hours among hourly workers that motivates the bunching design approach in that context. Readers primarily interested in the econometric contribution of this paper may wish to skip directly to Section 4

Given the time-and-a-half rule, total pay for a given worker in a particular week is a kinked function of the worker's hours that week, as depicted in Figure 1. This is true provided that the worker's hourly wage w is fixed with respect to the choice of hours that week. Indeed, the data (detailed in Section 3) reveal that hours tend to vary considerably between weeks for a given hourly worker, while workers' wages change only infrequently. I propose to view this as a two stage-process. In a first step, workers are hired with an hourly wage set along with an "anticipated" number of weekly hours. Then, with that hourly wage fixed in the short-run, final scheduling of hours is controlled by the firm and varies by week given shocks to the firm's demand for labor.

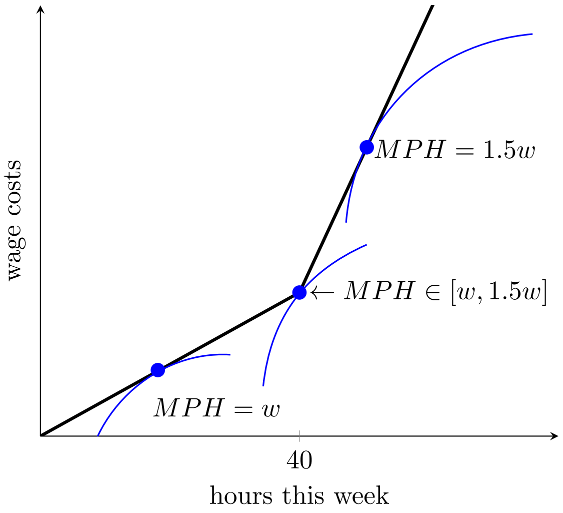


FIGURE 1: With a given worker's straight-time wage fixed at w, labor costs as a function of hours have a convex kink at 40 hours, given the overtime rule. Simple models of week-by-week hours choice (see Section 4.2) yield bunching when for some workers, the marginal product of an hour at 40 is between w and $1.5w$.

Wages and anticipated hours set at hiring

We begin with the hiring stage, which pins down the worker's wage. The hourly rate of pay w that applies to the first 40 of a worker's hours is referred to as their *straight-time wage* or simply *straight wage*. The following provides a benchmark model to endogenize such straight wages. This yields predictions about how wages may themselves be affected by the overtime rule, which will prove useful in our final evaluation of the FLSA. However, the basic bunching design strategy of Section 4 will only require that *some* straight-time wage is agreed upon and fixed in the short-run for each worker, as can be observed directly in the data.

Suppose that firms hire by posting an earnings-hours pair (z, h), where z is total weekly compensation offered to each worker, and h is the number of hours of work per week advertised at the time of hiring. The firm faces a labor supply function $N(z,h)$ determined by workers' preferences over the labor-leisure trade-off,[4] and makes a choice of (z^*, h^*) given this labor supply function and their production technology. For simplicity, workers are here taken to be homogeneous in production, paid hourly, and all covered by the overtime rule.[5]

While labor supply has above been viewed as a function over *total* compensation z and hours, there is always a unique straight wage associated with a particular (z,h) pair, such that h hours of work yields earnings of z, given the FLSA overtime rule:

$$w_s(z,h) := \frac{z}{h + 0.5 \cdot \mathbb{1}(h > 40)(h-40)} \tag{1}$$

We can distinguish the two main views proposed in the literature regarding the effects of overtime policy by supposing that a worker's straight-time wage is set according to Eq. (1), given values z^* and h^* that the firm and worker agree upon at the time of hiring. Trejo (1991) calls these two views the *fixed-job* and the *fixed-wage* models of overtime.

The *fixed-job* view observes that for a generic smooth labor supply function $N(z,h)$ (and smooth revenue production function with respect to hours), the optimal job package (z^*, h^*) for the firm to post will be *the same* as the optimal one absent the FLSA, as the hourly wage rate simply adjusts to fully neutralize the overtime premium.[6] Suppose for the moment that workers in fact work exactly h^* hours each week (abstracting away from any reasons for the firm to ever deviate from h^* in a given week). Then the FLSA would have no effect on earnings, hours or employment, provided that $w_s(z^*, h^*)$ is above any applicable minimum wage (Trejo, 1991).

On the *fixed-wage* view, the firm instead faces an exogenous straight-time wage when determining (z^*, h^*). Versions of this idea are considered in Brechling (1965), Rosen (1968), Ehrenberg (1971), Hamermesh (1993), Hart (2004) and Cahuc and Zylberberg (2014). This can be captured by a discontinuous labor supply function $N(z,h)$ that exhibits perfect competition on the quantity $w_s(z,h)$. I show in Appendix H.1 that in this case h^* and z^* are pinned down by the concavity of production with respect to hours and the scale of fixed costs (e.g. training for each worker) that do not depend on hours. The fixed-wage job makes the clear prediction that the FLSA will cause a reduction in hours, and bunching at 40.[7]

Existing work has investigated whether the fixed-job or fixed-wage model better accords with

[4] This labor supply function can be viewed as an equilibrium object that reflects both worker preferences and the competitive environment for labor. Appendix H.2 embeds $N(z,h)$ in a simple extension of the imperfectly competitive Burdett and Mortensen (1998) search model, and considers how it might react endogenously to the FLSA.

[5] By "covered" I mean workers that are not exempt from the FLSA overtime rule, at firms covered by the FLSA.

[6] In Appendix H.1 I give a closed-form expression for (z^*, h^*) when both labor supply and production are isoelastic: hours and earnings are each increasing in the elasticity of labor supply with respect to earnings, and decreasing in the magnitude of the elasticity of labor supply with respect to pay.

[7] A fixed-wage model tends to predict an overall positive effect on employment given plausible assumptions on labor/capital substitution (Cahuc and Zylberberg, 2014), though total labor-hours will decrease (Hamermesh, 1993).

the observed joint distribution of hourly wages and hours (Trejo, 1991; Barkume, 2010). These papers find that wages do tend to be lower among jobs that have overtime pay provisions and more overtime hours, but by a magnitude smaller than would be predicted by the pure fixed job model. These estimates could be driven by selection however, e.g. of lower-skilled workers into covered jobs with longer hours. In Appendix E.4, I construct a new empirical test of Eq. (1) (at the level of individual paychecks), that is instead based on assuming that the conditional distribution of pay is smooth across 40 hours. I find that roughly one quarter of paychecks around 40 hours reflect the wage/hours relationship predicted by the fixed-job model.

This finding is consistent with a model in which hours remain flexible week-to-week, while straight-wages remain fairly static after being set initially according to Equation (1).[8] In common with the fixed wage model, this two-stage framework allows for the possibility that the overtime rule affects hours, and predicts bunching at 40; however, this is driven by short-run rigidity in straight-wages, rather than by perfect competition as in previous fixed-wage approaches.

Dynamic adjustment to hours by week

After (z^*, h^*) is set, there are many reasons to still expect week-to-week variation in the number of hours that a firm would desire from a given worker. If demand for the firm's products is seasonal or volatile, it may not be worthwhile to hire additional workers only to reduce employment later. Similarly, productivity differences between workers may only become apparent to supervisors after those workers' straight wages have been set, and vary by week.

Throughout Section 4, I maintain a strong version of the assumption that the firm—rather than the worker—chooses the final hours that I observe on a given paycheck. This simplification eases notation and emphasizes the intuition behind my identification strategy. Appendix F presents a generalization in which some fraction of workers choose their hours, along with intermediate cases in which the firm and worker bargain over hours each week. The results there show that if some workers have control of their final hours, the bunching-design strategy will only be informative about effects of the FLSA among workers whose final hours are chosen by the firm.[9]

Available survey evidence suggests that this latter group is the dominant one: a relatively small share of workers report that they choose their own schedules. For example, the 2017-2018 Job Flexibilities and Work Schedules Supplement of the American Time Use Survey asks workers whether they have some input into their schedule, or whether their firm decides it. Only 17% report that they have some input. In a survey of firms, only 10% report that most of their employees have

[8] This dovetails other recent evidence of uniformity and discretion in wage-setting, e.g. nominal wage rigidity (Grigsby et al. 2021), wage standardization (Hjort et al., 2020) and bunching at round numbers (Dube et al., 2020).

[9] The reason is that while the kink draws firms exactly to 40 hours, workers instead face an incentive to avoid it.

control over which shifts they work (Matos et al., 2017).[10]

3 Data and descriptive patterns

The main dataset I use comes from a large national payroll processing company. Administrative hours data at the weekly level has previously been unavailable to overtime researchers, and studies of overtime in the U.S. have typically relied on self-reported integer hours from surveys such as the Current Population Survey. The ability to observe exact number of hours that the worker was paid for in a given week allows me to construct the distribution of hours-of-pay without rounding or other sources of measurement error.

The payroll processing company provided anonymized paychecks for workers from a random sample of their employers, for all pay periods in 2016 and 2017. At the paycheck level, I observe the check date, straight wage, and amount of pay and hours corresponding to itemized pay types, including normal pay, overtime pay, sick pay, holiday pay, and paid time off. The data also include state and industry for each employer and for employees: age, tenure, gender, state of residence, pay frequency and salary if one is specified.

3.1 Sample description

I construct a final sample for analysis based on two desiderata: a) the ability to observe hours within a single week; and b) a focus on workers who are non-exempt from the FLSA overtime rule. For a) it is necessary to drop paychecks from workers who are not paid on a weekly basis (roughly half of the workers in the sample). Otherwise, it would not be possible to observe hours in a single week: the time period in which hours are regulated by the FLSA. To achieve b) I keep paychecks only from hourly workers, since nearly all workers who are paid hourly are subject to the overtime rule. I also drop any workers who have no variation in hours, as those workers are likely salaried workers for whom salary information was simply missing, and hours data are uninformative. As a final check for being non-exempt from the FLSA, I also drop observations from workers who never receive overtime pay during the study period.

The final sample includes 630,217 paychecks for 12,488 workers across 566 firms. Appendix E.1 provides further details of the sample construction, and compares its regional and industry distribution to that of a representative sample of workers.

Table 1 shows how the sample compares to survey data that is constructed to be representative of the U.S. labor force. Column (1) reports means from the final sample used in estimation, while (2)

[10] One rationalization of these observations is that if the worker and firm fail to agree on a worker's hours, the worker's outside option may be unemployment while the firm's is just one less worker (Stole and Zwiebel, 1996).

	(1) Estimation sample	(2) Initial sample	(3) CPS	(4) NCS
Tenure (years)	3.21	2.81	6.34	.
Age (years)	37.15	35.89	39.58	.
Female	0.23	0.46	0.50	.
Weekly hours	38.92	27.28	36.31	35.70
Gets overtime	1.00	0.37	0.17	0.52
Straight-time wage	16.16	22.17	18.09	23.31
Weekly overtime hours	3.56	0.94	.	1.04
Number of workers in sample	12,488	149,459	63,404	228,773

TABLE 1: Comparison of the sample with representative surveys. Columns 1 and 2 average across periods within worker from the administrative payroll sample, and then present means across workers. Column 2 presents means of worker-level data from the Current Population Survey and Column 3 averages representative job-level data from the National Compensation Survey.

reports means before sampling restrictions. Column (3) reports means from the Current Population Survey (CPS) for the same years 2016–2017, among individuals reporting hourly employment. The "gets overtime" variable for the CPS sample indicates that the worker usually receives overtime, tips, or commissions. Column (4) reports means for 2016–2017 from the National Compensation Survey (NCS), a representative establishment-level dataset accessed on a restricted basis from the Bureau of Labor Statistics. The NCS reports typical overtime worked at the quarterly level for each job in an establishment (drawn from firm administrative data when possible).[11]

The sample I use is more male, earns lower straight-time wages, and works more overtime than a typical hourly worker in the U.S. Column (2) in Table 1 reveals that my sampling restrictions can explain why the estimation sample tilts male and has higher overtime hours than the workforce as a whole. The initial sample is fairly representative on both counts, while conditioning on workers paid weekly oversamples industries that have more men, longer hours, and lower pay. Appendix E compares the industry and regional distributions of the estimation sample to the CPS.

3.2 Hours and wages in the sample

I turn now to the main variables to be used in the analysis. Figure 2 reports the distribution of hours of work in the final sample of paychecks. The graphs indicate a large mass of individuals who were paid for exactly 40 hours that week, amounting to about 11.6% of the sample.[12] Appendix Figure 3 shows that overtime pay is present in virtually all weekly paychecks that report more than 40 hours,

[11] The hourly wage variable for the CPS may mix straight-time and overtime rates, and is only present in outgoing rotation groups. The tenure variable comes from the 2018 Job Tenure Supplement. The NCS does not distinguish between hourly and salaried workers, reporting an average hourly rate that includes salaried workers, who tend to be paid more. This likely explains the higher value than the CPS and payroll samples.

[12] The second largest mass occurs at 32 hours, and is explained by paid time off as discussed in Section 5.

in line with the presumption that workers in the final sample are not FLSA-exempt.

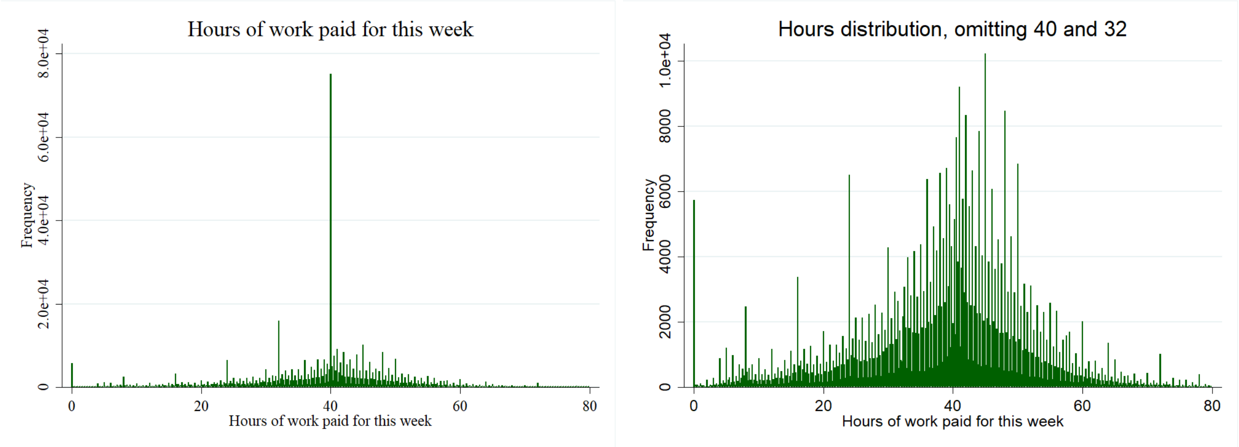


FIGURE 2: Empirical densities of hours worked pooling all paychecks in final estimation sample. Sample is restricted to hourly workers receiving overtime pay at some point (to ensure nearly all are non-exempt from FLSA, see text), and workers having hours variation. The right panel omits the points 40 and 32 to improve visibility elsewhere. Bins have a width of 1/8, below the granularity at which most firms record hours.

Table 2 documents that while the hours paid in 70% of all pay checks in the final estimation sample differ from those of the last paycheck by at least one hour, just 4% of all paychecks record a different straight-time wage than the previous paycheck for the same worker. Among the roughly 22,500 wage change events, the average change is about a 45 cent raise per hour, and when hours change the magnitude is about 7 hours on average and roughly symmetric around zero.[13]

	Mean	Std. dev.	N
Indicator for hours changed from last period	0.84	0.37	630,217
Indicator for hours changed by at least 1 hour	0.70	0.46	630,217
Indicator for wage changed from last period	0.04	0.19	630,217
Indicator for wage changed, if hours changed	0.04	0.19	529,791
Absolute value of hours difference, if hours changed	6.83	8.23	529,791
Difference in wage, if wage changed	0.45	26.46	22,501

TABLE 2: Changes in hours or straight wages between a worker's consecutive paychecks.

[13] Appendix E reports some further details from the data. Figure 5 shows the distribution of between-paycheck hours changes. Table 1 documents the prevalence of overtime pay by industry. Table 4 regresses hours, overtime, and bunching on worker and firm characteristics, showing that bunching and overtime hours are predicted by recent hiring at the firm. Table 5 shows that about 63% of variation in total hours can be explained by worker and employer-by-date fixed effects. Figure 8 considers the joint distribution of wages and hours and reproduces Bick et al.'s (2022) finding that mean wages increase with hours until just beyond 40, before declining.

4 Empirical strategy: a generalized kink bunching design

Let us now turn to the firm choosing the hours of a given worker in a particular week, with costs a fixed kinked function of hours as depicted in Figure 1. This section shows that under weak assumptions, firms facing such a kink will make choices that can be completely characterized by choices they *would* make under two counterfactual linear cost schedules that differ with respect to wage. I relate the observable bunching at 40 hours to a treatment effect defined from these two counterfactuals, which I then use to estimate the impact of the FLSA on hours.

The identification results in this section hold in an even more general setting in which a decision-maker faces a choice set with a possibly multivariate kink and has "nearly" convex preferences. I present the general version of this model in Appendix B. Throughout this section I refer to a worker i in week t as a *unit*: an observation of h_{it} for unit it is thus the hours recorded on a single paycheck.

4.1 A general choice model

Let us start from the conceptual framework introduced in Section 2. In choosing the hours h_{it} of worker i in week t, worker i's employer faces a kinked cost schedule, given the worker's straight-time wage w_{it} (which may depend on t). If the firm chooses less than 40 hours, it will pay $w = w_{it}$ for each hour, and if the firm chooses $h > 40$ it will pay $40w$ for the first 40 hours and $1.5w(h-40)$ for the remaining hours, giving the convex shape to Figure 1. We can write the kinked pay schedule for unit it as a function of hours this week h, as:

$$B_{it}(h) = w_{it}h + .5w_{it}\mathbb{1}(h > 40)(h-40) = \max\{B_{0it}(h), B_{1it}(h)\}$$

where $B_{0it}(h) = w_{it}h$ and $B_{1it}(h) = 1.5w_{it}h - 20w_{it}$. The kinked pay schedule $B_{it}(h)$ is equal to $B_{0it}(h)$ for values $h \leq 40$ and $B_{it}(h)$ is equal to $B_{1it}(h)$ for values $h \geq 40$. The functions B_0 and B_1 recover the two segments in Figure 1 when restricted to these domains respectively (see Appendix Figure 1). The following definition is generalized in Appendix B:

Definition (potential outcomes). *Let h_{0it} denote the hours of work that of unit it would be paid for if instead of $B_{it}(h)$, the pay schedule for week t's hours were $B_{0it}(h)$. Similarly, let h_{1it} denote the hours of pay that would occur for unit it if the pay schedule were $B_{1it}(h)$.*

The potential outcomes h_0 and h_1 thus imagine what would happen if instead of the kinked piecewise pay schedule $B_k(h)$, one of $B_0(h)$ or $B_1(h)$ applied globally for all values of h.

Let h_{it} denote the actual hours for which unit it is paid. Our first assumption is that actual hours and potential outcomes reflect choices made by the firm:

Assumption CHOICE. *Each of h_{0it}, h_{1it} and h_{it} reflect choices the firm would make under the pay schedules $B_{0it}(h)$, $B_{1it}(h)$, and $B_{it}(h)$ respectively.*

CHOICE reflects the assumption that hours are perfectly manipulable by firms. Note that if firm preferences over a unit's hours are quasi-linear with respect to costs (e.g. if they maximize weekly profits), the term $-20w_{it}$ appearing in B_{1it} plays no role in firm choices. As such, I will often refer to h_{1it} as choice made under linear pay at the overtime rate $1.5w_{it}$, keeping in mind that the exact definition for B_1 given above is necessary for the interpretation if preferences are not quasi-linear.

My second assumption is that each unit's firm optimizes some vector **x** of choice variables that pin down that unit's hours. As a leading case, we may think of hours of work as a single component of firms' choice vector **x** (Appendix B.3 gives some examples of this). Firm preferences are taken to be convex in **x** and the unit's wage costs z:

Assumption CONVEX. *Firm choices for unit it maximize some $\pi_{it}(z, \mathbf{x})$, where π_{it} is strictly quasiconcave in $(z, \mathbf{x})$ and decreasing in z. Hours are a continuous function of* **x** *for each unit.*

Relative to existing literature, Assumption CONVEX is most closely related to Blomquist et al. (2015), who consider a nonparametric choice model in which workers facing an income tax kink determine their earnings by choosing two quantities (hours and effort).[14] However, the way that I accommodate multiple margins of choice differs from that of Blomquist et al. (2015). Those authors define an effective utility function in terms of consumption and earnings alone (analogous to z and h in my setting) by concentrating out all but the observed choice variable, and then assuming quasi-concavity of this concentrated utility function. CONVEX instead assumes convexity of preferences defined directly over the primitive margins of choice. This assumption can be evaluated on choice-theoretic grounds alone, requiring no assumptions on how h depends on **x** beyond continuity.

For the sake of brevity, I have above stated a version of CONVEX that is a bit stronger than necessary for the identification results below. Appendix B relaxes CONVEX to allow for "double-peaked" preferences with one peak located exactly at the kink (this is relevant if firms have a special preference for a 40 hour work week). The appendix also shows that bunching still has some identifying power without any convexity of preferences. Note that the assumption that firms rather than workers choose hours enters in the claim that π is decreasing (rather than increasing) in z, but Appendix F relaxes this to allow some workers to set their hours.

Observables in the bunching design

The starting point for our analysis of identification in the bunching design is the following mapping between actual hours h_{it} and the counterfactual hours choices h_{0it} and h_{1it}. Appendix Lemma B.1

[14] Blomquist et al. (2017) also discusses a nonparametric choice model for the bunching design, but takes the choice variable to be an observable scalar.

shows that Assumptions CHOICE and CONVEX imply that:

$$h_{it} = \begin{cases} h_{0it} & \text{if } \; h_{0it} < 40 \\ 40 & \text{if } \; h_{1it} \leq 40 \leq h_{0it} \\ h_{1it} & \text{if } \; h_{1it} > 40 \end{cases} \qquad (2)$$

That is, a worker will work h_{0it} hours when the counterfactual choice h_{0it} is less than 40, and h_{1it} hours when h_{1it} is greater than 40. They will be found at the corner solution of 40 if and only if the two counterfactual outcomes fall on either side, "straddling" the kink.[15] Figure 3 depicts the implications of Eq. (2) for what is therefore observable by the researcher in the bunching design: censored distributions of h_0 and of h_1, and a point-mass of $\mathcal{B} = P(h_{1it} \leq 40 \leq h_{0it})$ at the kink.

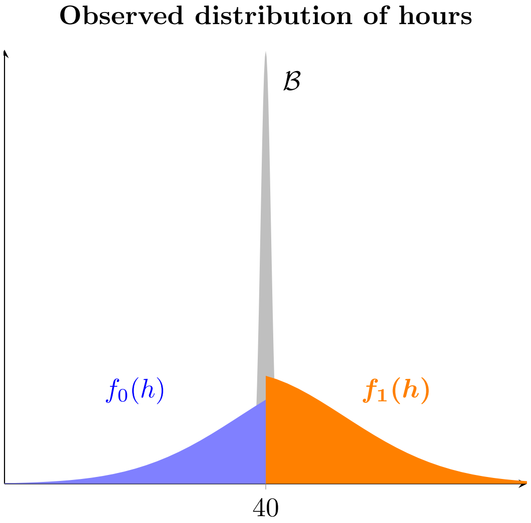


FIGURE 3: Observables in the bunching design, given Equation (2). To the left of the kink at 40, the researcher observes the density $f_0(h)$ of the counterfactual h_{0it}, up to values $h = 40$. To the right of the kink, the researcher observes the density $f_1(h)$ of h_{1it} for values $h > 40$. At the kink, one observes a point-mass of size $\mathcal{B} := P(h_{it} = 40) = P(h_{1it} \leq 40 \leq h_{0it})$.

Equation (2) represents a central departure from most previous approaches to the bunching design, which characterize bunching in terms of the counterfactual h_0 only.[16] I show below that such is a simplification afforded by the benchmark isoelastic utility model, but in a generic choice model, both h_0 and h_1 are necessary to pin down actual choices h_{it}. Appendix B shows that Eq. (2) also holds in settings with possibly non piecewise-linear kinked choice sets of the form: $z \geq \max\{B_0(\mathbf{x}), B_1(\mathbf{x})\}$ where B_0 and B_1 are weakly convex in the full vector **x**, and z any "cost"

[15] "Straddling" can only occur in one direction, with $h_{1it} \leq k \leq h_{0it}$. The other direction: $h_{0it} \leq k \leq h_{1it}$ with at least one inequality strict, is ruled out by the weak axiom of revealed preference (see Appendix B).

[16] Blomquist et al. (2015) also derive an expression for $\mathcal{B}$ in terms of agents' choices given all intermediate slopes between those occurring on either side of the kink. I discuss this and offer a generalization in Appendix Lemma B.2.

decision-makers dislike.

Intuition for Equation (2) in the overtime setting

As an illustration of Equation (2), suppose that firms balance the cost $B_{it}(h)$ against the value of h hours of the worker's labor, in order to maximize that week's profits. Then $h_{0it} = MPH_{it}^{-1}(w_{it})$ and $h_{1it} = MPH_{it}^{-1}(1.5w_{it})$, where denotes $MPH_{it}(h)$ is the marginal product of an hour of labor for unit it, as a function of that unit's hours h. Assuming that production is strictly concave (in line with Assumption CONVEX), the function $MPH_{it}(h)$ will be strictly decreasing in h.

Figure 1 depicts this visually. Consider for example a worker with a straight-wage of \$10 an hour. If there exists a value $h < 40$ such that the worker's MPH is equal to \$10, then the firm will choose this point of tangency. This happens if and only if the marginal product of an hour at 40 hours this week is less than \$10. If instead, the marginal product of an hour is still greater than \$15 at $h = 40$, the firm will choose the value $h > 40$ such that MPH equals \$15. The third possibility is that the MPH at $h = 40$ is *between* the straight and overtime rates \$10 and \$15. In this case, the firm will choose the corner solution $h = 40$, contributing to bunching at the kink.

Appendix B.3 provides some examples that use the full generality of Assumption CONVEX, in which firms simultaneously consider *multiple* margins of choice aside from a given unit's hours. For example, the firm may attempt to mitigate the added cost of overtime by reducing bonuses when a worker works many overtime hours. Eq. (2) remains valid even when such additional margins of choice are unmodeled and unobserved by the econometrician, varying possibly by unit.

4.2 Special case: the benchmark isoelastic model

This section introduces the canonical approach in the bunching-design literature (Saez, 2010; Chetty et al., 2011; Kleven, 2016; Blomquist et al., 2021), which specializes to a particularly simple case of the general model from the last section that I refer to as the "isoelastic model". Although it serves as an important benchmark, I show in this section that the isoelastic model can be rejected on economic grounds in the overtime setting, when confronted with the data. This underscores the need for results valid in the general choice model, which I develop in Section 4.3.

The isoelastic model strengthens Assumption CONVEX to suppose that $\mathbf{x} = h$ and that decision-makers' utility has a constant elasticity, with preferences identical between units up to a scalar heterogeneity parameter. By assuming that firms consider *only* hours h as a margin of choice, the isoelastic model amounts to a model of revenue production in which firm profits from unit it are:

$$\pi_{it}(z, h) = a_{it} \cdot \frac{h^{1+\frac{1}{\epsilon}}}{1+\frac{1}{\epsilon}} - z \tag{3}$$

where $\epsilon < 0$ is common across units, and z represents wage costs for worker i in week t. Eq. (3) is analogous to the isoelastic, quasilinear labor *supply* model used in the context of tax kinks.

Under a linear pay schedule $z = wh$, the profit maximizing number of hours is $(w/a_{it})^{\epsilon}$, so ϵ yields the elasticity of hours demand with respect to a linear wage. Letting $\eta_{it} = a_{it}/w_{it}$ denote the ratio of a unit's current productivity factor a_{it} to their straight wage, we have:

$$h_{0it} = MPH_{it}^{-1}(w_{it}) = \eta_{it}^{-\epsilon} \qquad \text{and} \qquad h_{1it} = MPH_{it}^{-1}(1.5w_{it}) = 1.5^{\epsilon} cdot \eta_{it}^{-\epsilon},$$

By Eq. (2), actual hours h_{it} are thus ranked across units in order of η_{it}, and the value of η_{it} determines whether a worker works overtime in a given week. If η_{it} is continuously distributed with support overlapping the interval $[40^{-1/\epsilon}, 1.5 \cdot 40^{-1/\epsilon}]$, then the observed distribution of h_{it} will feature a point mass at 40—"bunching"—and a density elsewhere.

Identification in the isoelastic model

In the context of the isoelastic model, a natural starting place for evaluating the FLSA would be to estimate the parameter ϵ. The classic bunching-design method pioneered by Saez (2010) identifies ϵ by relating it to the observable bunching probability:

$$\mathcal{B} := P(h_{it} = 40) = \int_{40}^{1.5^{|\epsilon|} \cdot 40} f_0(h) \cdot dh \tag{4}$$

where f_0 is the density of h_0. If the function f_0 were known, the value of ϵ could be pinned down by simply solving Eq. (4) for $|\epsilon|$. However, f_0 is not globally identified from the data: from Figure 3 we can see that f_0 is only identified to the left of the kink, while the density of h_1 is identified to the right of the kink. Since $h_{1it} = 1.5^{\epsilon} \cdot h_{0it}$, it is convenient in the isoelastic model to analyze observables after applying a log transformation to hours: the quantity $\delta = \ln h_{0it} - \ln h_{1it} = |\epsilon| \cdot \ln 1.5$ is homogeneous across all units it, and the density of $\ln h_{1it}$ is thus a simple leftward shift of the density of $\ln h_{0it}$, by δ, as shown in Figure 4.

Standard approaches in the bunching design make parametric assumptions that interpolate f_0 through the missing region of Figure 4 to point-identify ϵ.[17] The approach of Saez (2010) assumes for example that the density of h_0 is linear through the missing region $[40, 40 \cdot e^{\delta}]$ of Figure 4. The popular method of Chetty et al. (2011) instead fits a global polynomial, using the distribution of hours outside the missing region to impute the density of h_0 within it. Neither approach is

[17] Bertanha et al. (2023) note that given a full parametric distribution for f_0, the entire model could be estimated by maximum likelihood. This approach would enforce (4) automatically while enjoying the efficiency properties of MLE.

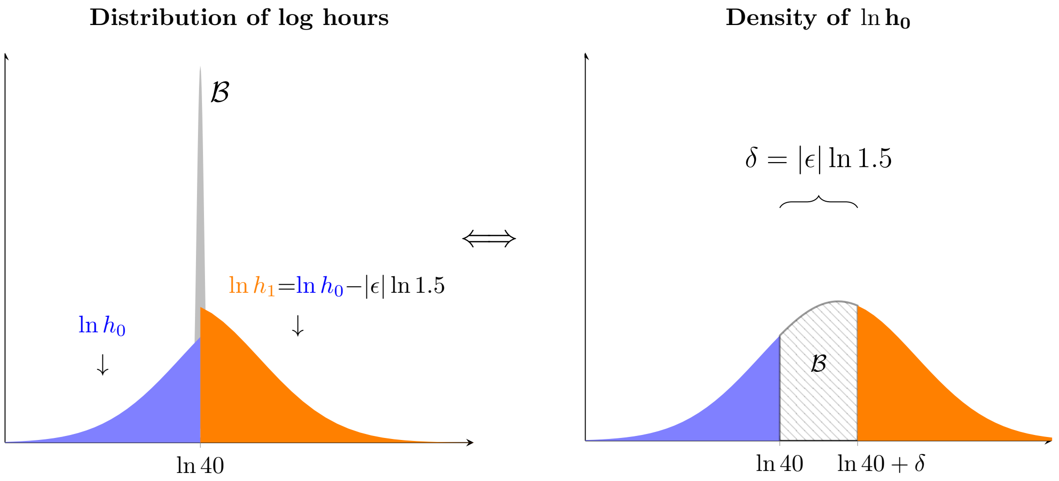


FIGURE 4: The left panel depicts the distribution of observed log hours $\ln h_{it}$ in the isoelastic model, while the right panel depicts the underlying full density of $\ln h_{0it}$. Specializing from the general setting of Figure 3, we have in the isoelastic model that $\tilde{f}_1(h) = \tilde{f}_0(h + |\epsilon| \cdot \ln 1.5)$, where $\tilde{f}_d$ is the density of $\ln h_d$. Thus, the full density of $\tilde{f}_0$ is related to the observed distribution by "sliding" the observed distribution for $h > 40$ to the right by the unknown distance $\delta = |\epsilon| \ln 1.5$, leaving a missing region in which f_0 is unobserved. The total area in the missing region from $\ln 40$ to $\ln 40 + \delta$ must equal the observed bunching mass $\mathcal{B}$.

particularly suitable in the overtime context. [18]

If in the other extreme the researcher is unwilling to assume anything about the density of h_0 in the missing region of Figure 4, then the data are compatible with any finite $\epsilon < 0$ as emphasized by Blomquist et al. (2021) and Bertanha et al. (2023). In particular, given (4), an arbitrarily small $|\epsilon|$ could be rationalized by a density that spikes sufficiently high just to the right of 40, while an arbitrarily large $|\epsilon|$ can be reconciled with the data by supposing that the density of h_0 drops quickly to some very small level throughout the missing region.

Rejection of the isoelastic model

Compared with the isoelastic model, the general choice model from Section 4.1 allows for a wide range of underlying choice models that might drive a firm's hours response to the FLSA. This robustness over structural models turns out to be important in the overtime context.

Table 3 reports estimates of the parameter ϵ in the isoelastic model when various shape constraints are assumed about the distribution of $\ln h_0$. The first row of Table 3 imposes a linear density across the missing region. The second assumes that the density of h_0 is monotonic across the missing region. The third imposes the non-parametric shape constraint of *bi-log-concavity* (BLC) on

[18] The linear method of Saez (2010) implies monotonicity of the density in the missing region, which is unlikely to hold given that 40 appears to be near the mode of the h_0 latent hours distribution. Meanwhile, the method of Chetty et al. (2011) ignores the "shift" by δ in the right panel of Figure 4. Both approaches ultimately rely on parametric assumptions, and sufficient conditions for each are outlined in Appendix I.2.

the CDF of h_0. This is the same restriction that will be later imposed for h_0 and h_1 in the context of the general choice model—and a detailed discussion of BLC is given in Section 4.3. BLC nests the linear density assumption and leads to partial, rather than point, identification of ϵ.[19]

Distributional assumption	**LB of CI for ϵ**	**Implied $\frac{MPH(40)}{MPH(10)}$**	**Model rejected?**
Linear density of log hours	-0.183	0.05%	Yes
Monotonic density of log hours	-0.207	0.12%	Yes
Bi-log-concave CDF of log hours	-0.198	0.09%	Yes

TABLE 3: Testing the plausibility of the isoelastic model in the overtime context. The isoelastic model is considered rejected if the confidence interval for ϵ does not overlap the "reasonable" range of values $(-\infty, -0.6]$, which corresponds to workers maintaining at least 10% of their 10-hour marginal productivity when they get to 40 hours for the week. Estimates are drawn from Appendix Figure 2 ($p = 0$ column).

If the iso-elastic model holds, a given value of ϵ implies an elasticity of revenue production with respect to hours of work, governing for any $\epsilon < 0$ how quickly the marginal product of an hour of labor declines with h (e.g. due to worker fatigue). I find that across all three methods reported in Table 3, the range of values for ϵ compatible with the data are all economically implausible. For concreteness, let us define a "plausible" production function as one in which the marginal product of an hour of labor after a worker has worked 40 hours in a week is at least 10% of the marginal product of an hour of labor after they have worked just 10 hours that week,[20] but is no greater than $MPH(40)$. Via Eq. (3), this translates into $\frac{MPH_{it}(40)}{MPH_{it}(10)} = 4^{1/\epsilon} \in [0.1, 1]$ or $\epsilon \in (-\infty, -0.6]$. The second column of Table 3 reports the lower bound of a confidence interval for ϵ, drawing from estimates reported in Appendix Figure 2. These "best-case" values of ϵ never fall in $(-\infty, -0.6]$.

In short, the observed bunching at 40 hours is too small to be reconciled with a model in which a single ϵ parameterizes the decline of hourly productivity with hours—the production function is too concave to be realistic.[21] This motivates a model like the one presented in Section 4.1, in which we can interpret the estimand of the bunching design as a *reduced-form* averaged elasticity of the demand for hours, where the margins of choice available to the firm are not assumed to consist of hours alone. As described through some examples in Appendix B.3, this elasticity may reflect adjustment by firms along additional margins that can attenuate the hours response, and thus reduce

[19] Appendix E.3 (in online material) assumes BLC (nesting linearity as in Saez (2010)) for the distribution of h_0, rather than for $\ln h_0$. When using the un-logged hours distribution, it is no longer redundant to impose the above restrictions on the distribution of h_1 in addition to h_0. Assuming that h_0 and h_1 are both BLC suggests that $\epsilon \in [-.179, -.168]$. The width of these bounds is about 4 times smaller than if BLC is assumed for h_0 only.

[20] This is a very conservative figure. Although general evidence is lacking, Pencavel (2015) finds hourly productivity to be relatively constant until 49 hours in a week (so $\frac{MPH(40)}{MPH(10)} \approx 1$), e.g. for munitions workers during World War I.

[21] The estimates in Table 3 attribute all of the bunching observed at 40 to the FLSA: attributing just a portion of the bunching at 40 to the FLSA (as I do in Section 5.1) would only further reduce the magnitude of ϵ. Industry-specific bounds on ϵ range from -0.26 to -0.06, suggesting the isoelastic model is also rejected within each industry.

the magnitude of bunching.

4.3 Identifying treatment effects in the general choice model

In this section I turn to identification in the general choice model of Section 4.1. Without a single preference parameter like ϵ that characterizes responsiveness to incentives for all units, we face the following question: what quantity might be identifiable from the data without the restrictive isoelastic model, but still help us to evaluate the effect of the FLSA on hours?

Let us refer to the difference $\Delta_{it} := h_{0it} - h_{1it}$ between h_0 and h_1 as unit it's *treatment effect*. Recall that h_0 and h_1 are interpreted as potential outcomes, indicating what *would* have happened had the firm faced either of two counterfactual pay schedules instead of the kink in a given week. Δ_{it} thus represents the causal effect of a one-time 50% increase in worker i's wage on their hours in week t. As this is the difference between the hours that unit's firm would choose if the worker were paid at their straight-time rate versus at their higher overtime rate for all hours in that week, we would expect that Δ_{it} tend to be positive. However, it is not required by the general choice model that $\Delta_{it} \geq 0$ for all units it (see Appendix B for a discussion).

In the isoelastic model $\Delta_{it} = h_{0it} \cdot (1 - 1.5^{\epsilon})$, representing a special case in which treatment effects are homogeneous across units after a log transformation of the outcome: $\ln h_{0it} - \ln h_{1it} = |\epsilon| \cdot \ln 1.5$. In general, we can expect Δ_{it} to vary much more flexibly across units, and a reasonable parameter of interest becomes a summary statistic of Δ_{it} of some kind. In particular, Eq. (2) suggests that bunching is informative about the distribution of Δ_{it} among units "near" the kink. To see this, let $k = 40$ denote the location of the kink, and write the bunching probability as:

$$\mathcal{B} = P(h_{1it} \leq k \leq h_{0it}) = P(h_{0it} \in [k, k + \Delta_{it}]) = P(h_{1it} \in [k - \Delta_{it}, k]), \tag{5}$$

i.e. units bunch when their h_0 potential outcome lies to the right of the kink, but within that unit's individual treatment effect of it. Note that by Eq. (2) we can also write bunching in terms of the marginal distributions of h_0 and h_1:[22]

$$\mathcal{B} = F_1(k) - F_0(k) \tag{6}$$

where F_0 and F_1 denote the cumulative distribution functions of each potential outcome.

[22] To obtain this expression, write $1 = P(h \leq k) + P(h > k) = \{P(h_0 < k) + \mathcal{B}\} + P(h_1 > k) = P(h_0 \leq k) + \mathcal{B} + 1 - P(h_1 \leq k)$ where the first equality uses Eq. (2) and the second assumes continuity of the CDF of h_0.

4.3.1 Parameter of interest: the buncher ATE

I focus my identification analysis on the average treatment effect among units who locate at exactly 40 hours, a parameter I call the "buncher ATE". In the overtime setting some additional care is needed in defining this parameter, to allow for the possibility that a mass of units would still work exactly 40 hours, even absent the FLSA. Let us indicate such "counterfactual bunchers" by an (unobserved) binary variable $K^*_{it} = 1$, and define the buncher ATE to be:

$$\Delta^*_k = \mathbb{E}[\Delta_{it}|h_{it} = k, K^*_{it} = 0],$$

That is, Δ^*_k is the average value of Δ_{it} among bunchers who bunch in response to the FLSA kink. In evaluating the FLSA, I suppose that all counterfactual bunchers have a zero treatment effect, such that $h_{0it} = h_{1it} = k$. Since $\Delta_{it} = 0$ for these units by assumption, we can move back and forth between Δ^*_k and $\mathbb{E}[\Delta_{it}|h_{it} = k]$, provided the counterfactual bunching mass $p := P(K^*_{it} = 1)$ is known. In this section, I treat p as given, and present two strategies to estimate it empirically in Section 5.1. To simplify notation, the discussion of the buncher ATE in this section will largely focus on the case of $p = 0$, so that Δ^*_k simplifies to $\mathbb{E}[\Delta_{it}|h_{it} = k]$.

Comparison with literature: While the buncher ATE captures a reduced form labor demand response in levels (i.e. measured as a difference in hours), it can be related directly to the elasticity of labor demand by first applying a log transformation to hours. In the isoelastic model, for example, $\mathbb{E}[\ln h_{0it} - \ln h_{1it}|h_{it} = k] = \epsilon \cdot \ln(1.5)$. More generally, we have that $\delta^*_k := \mathbb{E}[\ln h_{0it} - \ln h_{1it}|h_{it} = k] = \ln(1.5) \cdot \mathbb{E}[\bar{\epsilon}_{it}|h_{it} = k]$, with ϵ replaced by a weighted "arc" elasticity of demand averaged among the bunchers $\mathbb{E}[\bar{\epsilon}_{it}|h_{it} = k]$, and integrated over hypothetical intermediate overtime rates between 1 and 1.5. To see this, let $h_{it}(\rho)$ be the hours that unit it would work if their employer faced a linear pay schedule at rate $\rho \cdot w_{it}$. In this notation, $h_{0it} = h_{it}(1)$ and $h_{0it} = h_{it}(1.5)$. Assuming differentiability of $h_{it}(\rho)$ in ρ for each unit, we have by the fundamental theorem of calculus that:

$$\delta^*_k = \mathbb{E}\left[\int_1^{1.5} \frac{d}{d\rho} \ln h_{it}(\rho) \cdot d\rho \middle| h_{it} = k\right] = \ln(1.5) \cdot \mathbb{E}[\bar{\epsilon}_{it}|h_{it} = k] \tag{7}$$

where $\bar{\epsilon}_{it} := \left(\int_1^{1.5} \frac{\frac{d}{d\rho}\ln h_{it}(\rho)}{\frac{d}{d\rho}\ln \rho} \cdot \frac{1}{\rho} \cdot d\rho\right) / \left(\int_1^{1.5} \frac{1}{\rho} \cdot d\rho\right)$ is a weighted average elasticity of hours demand with respect to a linear wage rate, integrated over the range of wages $[w_{it}, 1.5w_{it}]$.[23]

This notation also allows us to compare the buncher ATE to a result of Blomquist et al. (2015)

[23] Intuitively, the weighting in proportion to $1/\rho$ "undoes" the fact that level differences in $\ln h_{it}$ translate into larger *elasticities* when ρ is large. In the limit of a "small" kink this weighting will have little effect. The analogous expression for δ^*_k in the case of a tax kink with tax rates $\tau_1 > \tau_0$ would involve a factor of $\ln\left(\frac{1-\tau_0}{1-\tau_1}\right)$ rather than $\ln 1.5$.

that considers what bunching reveals in the context of non-parametric utility. They show that the bunching probability can be written as $\mathcal{B} = \int_1^{1.5} \epsilon(\rho) \cdot \frac{k}{\rho} \cdot f_\rho(k) \cdot d\rho$, where $f_\rho(h)$ reflects the density of the counterfactual $h_{it}(\rho)$ and $\epsilon(\rho)$ is an average compensated elasticity among those with $h_{it}(\rho) =$.[24]. However $\int_1^{1.5} k/\rho \cdot f_\rho(k) \cdot d\rho$ is not identified, and this expression therefore does not pin down a convex average of elasticities as Eq. (7) does.

By inspecting the data close to the kink on either side, we can see that $f_1(k)$ and $f_{1.5}(k)$ are identified (as $f_0(k)$ and $f_1(k)$ in the notation of Figure 3). But $f_\rho(k)$ is *only* identified for the isolated values of $\rho = 1$ and $\rho = 1.5$, so there is no principled way of "extrapolating" from the distributions of h_0 and h_1 to capture the magnitude of $f_\rho(k)$ for intermediate $\rho \in (1, 1.5)$. Blomquist et al. (2015) therefore take a pessimistic view of the prospect of identifying an averaged elasticity from bunching. Eq. (7), by contrast, reveals that we can avoid needing to make any assumptions about the distribution of $h_{it}(\rho)$ for intermediate $\rho \in (1, 1.5)$ by focusing on the buncher ATE (whether in logs or levels) and extrapolating from the observed distribution across *hours* h. For hours, we can present evidence to support the extrapolation assumption made using data away from the kink. We cannot however marshal such evidence in favor of extrapolation assumptions over ρ. This is the fundamental difference between my approach and the more negative conclusions of Blomquist et al. (2015) and the later published version Blomquist et al. (2021).

4.3.2 Reducing the buncher ATE to a pair of extrapolation problems

To simplify the discussion, let us for the moment continue supposing that $p = 0$, so that $\Delta_k^* = \mathbb{E}[\Delta_{it}|h_{it} = k]$. Our goal is to invert (5) in some way to learn about the buncher ATE from the observable bunching probability $\mathcal{B}$. In Figure 4, we've seen the intuition for this exercise in the context of the isoelastic model, in which there is only a scalar notion of heterogeneity and $h_{1it} = h_{0it} \cdot 1.5^\epsilon$. The key implication of the isoelastic model that aids in identification is *rank invariance* between h_0 and h_1. Rank invariance (Chernozhukov and Hansen 2005) says that $F_0(h_{0it}) = F_1(h_{1it})$ for all units, i.e. increasing each unit's wage by 50% does not change any unit's rank in the hours distribution (for example, a worker at the median of the h_0 distribution also has a median value of h_1). Rank invariance is satisfied by models in which there is perfect positive co-dependence between the potential outcomes (e.g. the left panel of Figure 5).

Rank invariance is useful because it allows us to translate statements about Δ_{it} into statements about the *marginal* distributions of h_{0it} and h_{1it}. In particular, under rank invariance the buncher ATE is equal to the quantile treatment effect $Q_0(u) - Q_1(u)$ averaged across all u between $F_0(k)$

[24] See Theorem 4 of Blomquist et al. (2015) and Theorem 8 Blomquist et al. (2021)

and $F_1(k) = F_0(k) + \mathcal{B}$, where Q_d is the quantile function of h_{dit}, i.e.:

$$\Delta_k^* = \frac{1}{\mathcal{B}} \int_{F_0(k)}^{F_1(k)} [Q_0(u) - Q_1(u)] du, \tag{8}$$

so long as $F_0(y)$ and $F_1(y)$ are continuous and strictly increasing. I focus on partial identification of the buncher ATE, for which it is sufficient to place point-wise bounds on the quantile functions $Q_0(u)$ and $Q_1(u)$ throughout the range $u \in [F_0(k), F_1(k)]$ as depicted in Figure 6.

While rank invariance already relaxes the isoelastic model used thus far in the literature, a still weaker assumption proves sufficient for Eq. (8) to hold:

Assumption RANK. *For some positive values* Δ_0^* *and* Δ_1^*, $h_{0it} \in [k, k + \Delta_0^*]$ *iff* $h_{1it} \in [k - \Delta_1^*, k]$.

Unlike (strict) rank invariance, Assumption RANK allows ranks to be reshuffled by treatment among bunchers and among the group of units that locate on each side of the kink.[25] For example, suppose that a 50% increase in the wage of worker i would result in their hours being reduced from $h_{0it} = 50$ to $h_{1it} = 45$. If another worker j's hours are instead reduced from $h_{0jt} = 48$ to $h_{1jt} = 46$ under a 50% wage increase, workers i and j will switch ranks, without violating RANK. Note that RANK is also compatible with the existence of counterfactual bunchers $p > 0$.

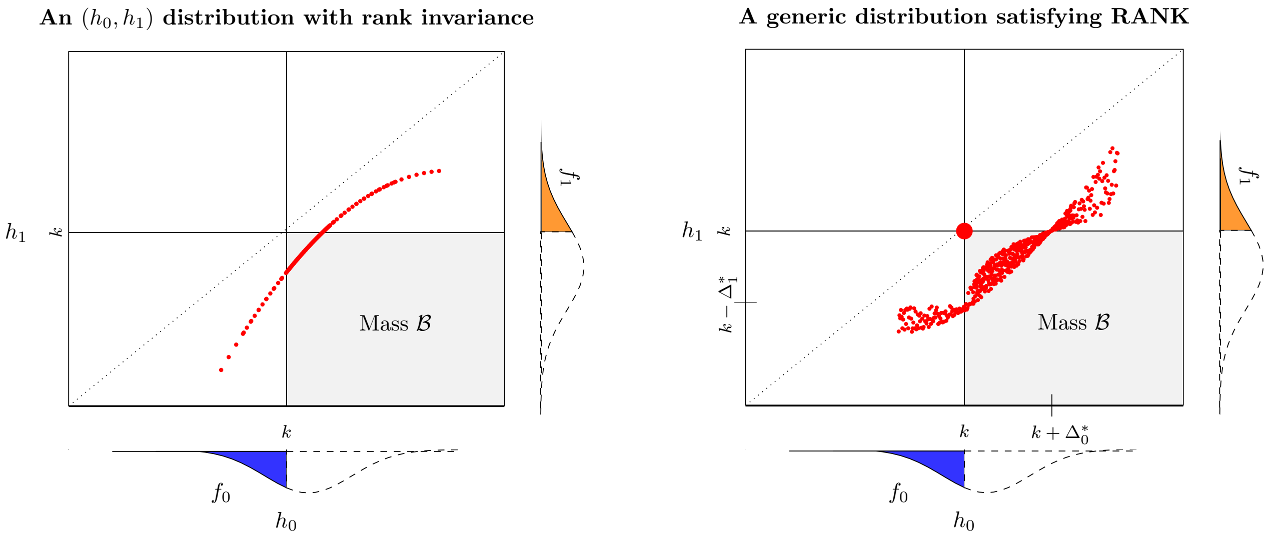
The right panel of Figure 5 shows an example of a distribution satisfying RANK, which requires the support of (h_0, h_1) to narrow to a point as it crosses $h_0 = k$ or $h_1 = k$. If this is not perfectly satisfied, Appendix B.5 demonstrates how the RHS of Equation (8) will then yield a lower bound on the true buncher ATE (and can also still be interpreted as an averaged quantile treatment effect). Appendix Figure 9 generalizes RANK to case in which some workers choose their hours, resulting in mass also appearing in the north-west quadrant of Figure 5.

Remark: Assumption RANK (like CONVEX) does *not* require that $h_{0it} \geq h_{1it}$ with probability one. While this is true in the examples of Figure 5 above, Appendix Figure 3 depicts an example of a joint distribution satisfying RANK in which some units it have negative treatment effects.

4.3.3 Bounds on the buncher ATE via bi-log-concavity

Given Eq. (8), I obtain bounds on the buncher ATE by assuming that both h_0 and h_1 have *bi-log-concave* distributions. Bi-log-concavity is a nonparametric shape constraint that generalizes log-concavity, a property of many familiar parametric distributions:

[25] When $p = 0$ Assumption RANK is equivalent to an instance of the *rank-similarity* assumption of Chernozhukov and Hansen (2005), in which the conditioning variable is which of the three cases of Eq. (2) holds for the unit. Specifically: $U_d|(h < k) \sim Unif[0, F_0(k)]$, $U_d|(h = k) \sim Unif[F_0(k), F_1(k)]$, and $U_d|(h > k) \sim Unif[F_1(k), 1]$



An (h_0, h_1) distribution with rank invariance

A generic distribution satisfying RANK

FIGURE 5: The joint distribution of (h_{0it}, h_{1it}) (in red), comparing an example satisfying rank invariance (left) to a case satisfying Assumption RANK (right). RANK allows the support of the joint distribution to "fan-out" from perfect co-dependence of h_0 and h_1, except when either outcome is equal to k. The large dot in the right panel indicates a possible mass p of counterfactual bunchers. The observable data identifies the shaded portions of each outcome's marginal distribution (depicted along the bottom and right edges), as well as the total mass $\mathcal{B}$ in the (shaded) south-east quadrant.

Definition (BLC). *A distribution function F is is bi-log-concave (BLC) if both $\ln F$ and $\ln(1-F)$ are concave functions.*

If F is BLC then it admits a strictly positive density f that is itself differentiable with locally bounded derivative: $\frac{-f(h)^2}{1-F(h)} \leq f'(h) \leq \frac{f(h)^2}{F(h)}$ (Dümbgen et al., 2017). Intuitively, this rules out cases in which the density of h_0 or h_1 ever spikes or falls *too* quickly on the interior of its support, leading to non-identification of the type discussed in Section 4.2.[26] Note that for a given value $f(h)$, BLC constrains $f'(h)$ more the closer h is to the median of distribution F.

The assumption that h_0 and h_1 admit BLC distributions can be justified in three primary ways. First, it weakens parametric distributions distributional assumed by previous bunching design studies. BLC nests as a special case distributions with log-concave densities, such as the linear counterfactual density assumption used by Saez (2010), and more generally polynomial densities (when they have real roots) used by Chetty et al. 2011.[27] Secondly, the BLC property is partially testable in the bunching design, in the sense that $F_0(y)$ is observable for all $h < k$ and $F_1(h)$ is observable for all $h > k$. Appendix Figure 4 shows that the observable portions of F_0 and F_1 indeed satisfy BLC. Identification then simply requires us to believe that BLC *also* holds in the unobserved

[26] Bertanha et al. (2023) propose bounds in the isoelastic model by specifying a Lipschitz constant on the density of $\ln \eta_{it}$. This yields global rather than local bounds on f', based on a tuning parameter value that must be chosen.

[27] However, taking seriously the idea that h_0 and h_1 are polynomials allows for perfect extrapolation of their densities and hence point identification of the buncher ATE. In the context of the isoelastic model, I show in online supplemental material that assuming f_0 to belong to any parametric family of analytic functions affords point identification of ϵ.

portions of F_0 and F_1.

Finally, BLC has intuitive meaning in the context of working hours. Hours are BLC if and only if the hazard rate of working time and the hazard rates of non-work time are both increasing. These properties can in turn be motivated economically. In Appendix C I show how BLC arises naturally as a property of work hours when variation in hours stems from stochastic shocks to worker productivity over time, that accumulate within the week and satisfy a Markov property.

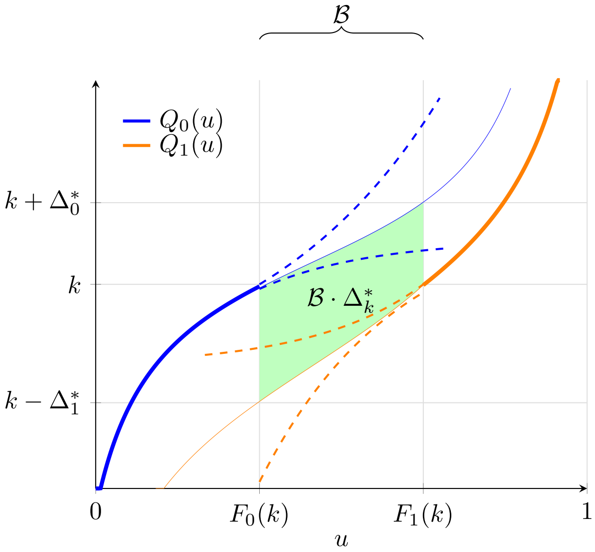


FIGURE 6: Extrapolating the quantile functions for h_0 and h_1 (blue and orange, respectively) to place bounds on the buncher ATE (case depicted has no counterfactual bunchers). The observed portions of each quantile function are depicted by thick curves, while the unobserved portions are indicated by thinner curves. The dashed curves represent upper and lower bounds for this unobserved portion coming from an assumption of bi-log-concavity. The buncher ATE is equal to the area shaded in green, divided by the bunching probability $\mathcal{B}$.[28]The quantities Δ_0^* and Δ_1^* are defined in Assumption RANK below.

We are now ready to state the main identification result, whose logic is summarized by Figure 6. Given the general choice model, RANK converts identification of the buncher ATE into a pair of extrapolation problems, each of which are approached by assuming the corresponding marginal potential outcome distribution is BLC. Let $F(h) := P(h_{it} \le h)$ be the CDF of observed hours.

Theorem 1 (bi-log-concavity bounds on the buncher ATE). *Assume CHOICE, CONVEX, RANK and that h_{0it} and h_{1it} have bi-log-concave distributions conditional on $K_{it}^* = 0$. Then:*

1. *$F(h)$, $F_0(h)$ and $F_1(h)$ are continuously differentiable for $h \neq k$. $F_0(k) = \lim_{h\uparrow k} F(h) + p$, $F_1(k) = F(k)$, $f_0(k) = \lim_{h\uparrow k} f(h)$ and $f_1(k) = \lim_{h\downarrow k} f(h)$, where if $p > 0$ we define the density of h_{dit} at $y = k$ to be $f_d(k) = \lim_{h\to k} f_d(h)$, for each $d \in \{0, 1\}$.*

2. *The buncher ATE Δ_k^* lies in the interval $\left[\Delta_k^L, \Delta_k^U\right]$, where:*

$$\Delta_k^L := g(F_0(k) - p, f_0(k), \mathcal{B} - p) + g\left(1 - F_1(k), f_1(k), \mathcal{B} - p\right)$$

$$\Delta_k^U := -g(1 - F_0(k), f_0(k), p - \mathcal{B}) - g\left(F_1(k) - p, f_1(k), p - \mathcal{B}\right)$$

with $g(a,b,x) = \frac{a}{b}\left[\left(1+\frac{a}{x}\right)\ln\left(1+\frac{x}{a}\right) - 1\right]$. The bounds Δ_k^L and Δ_k^U are sharp.

Proof. See Appendix A. □

Combining Items 1 and 2 of Theorem 1, it follows that the bounds Δ_k^L and Δ_k^U on the buncher ATE are identified, given the CDF $F(h)$ of hours and p.[29] Inspection of the expressions appearing in Theorem 1 reveals that Δ_k^U is always weakly larger than Δ_k^L, and the difference between the two grows the larger the net bunching probability $\mathcal{B} - p$. Some algebra also shows that when net bunching $\mathcal{B} - p$ is strictly positive $\Delta_k^L > 0$, so that the buncher ATE can be bounded away from zero.

Remark 1: The proof of Theorem 1 describes how the BLC assumption can be relaxed relative to its statement above, requiring only that h_{0it} be BLC on the interval $[k, k + \Delta_0^*]$ while h_{1it} is BLC on the interval $[k - \Delta_1^*, k]$ (both conditional on $K_{it}^* = 0$). The constants Δ_0^* and Δ_1^* are defined in Assumption RANK, and the notion of BLC holding on an interval is defined in the proof.

Remark 2: The bounds that will be presented on the buncher ATE presented in Theorem 1 can be easily translated into bounds on the buncher ATE in logs: $\delta_k^* \in [\Delta_k^L/k, \Delta_k^U/k]$, assuming that the distribtion of $\ln h_d$ (rather than h_d) is BLC for each $d \in \{0, 1\}$.

Testability and Sharpness: The proof of Theorem 1 establishes that the bounds Δ_k^L, Δ_k^U are sharp by constructing, for any value $\Delta \in [\Delta_k^L, \Delta_k^U]$, a joint distribution (h_0, h_1) that satisfies the assumptions of Theorem 1, and for which $\Delta_k^* = \Delta$. This distribution is compatible with the data provided that $F(h)$ is BLC on $(-\infty, k)$ and (k, ∞). Since $\Delta_k^U \geq \Delta_k^L$ always, the identified set $[\Delta_k^L, \Delta_k^U]$ is never empty and Theorem 1 cannot be used to falsify the general choice model along with BLC. If one accepts that any positive value of Δ_k^* represents a "plausible" reduced-form elasticity of hours demand then, this model cannot be rejected in the sense the isoelastic model with BLC was in Table 3.

[28] It is worth noting that BLC of h_1 and h_0 implies bounds on the treatment effect $Q_1(u) - Q_0(u)$ at *any* quantile u. But these bounds widen quickly as one moves away from the kink. When $f_0(k) \approx f_1(k)$, the narrowest bounds for a single rank u are obtained for a "median" buncher roughly halfway between $F_0(k)$ and $F_1(k)$. However, averaging over a larger group is more useful for meaningful ex-post evaluation of the FLSA (Sec. 4.4), and reduces the sensitivity to departures from RANK (see Figure 3). In the other extreme, one could drop RANK entirely and bound $\mathbb{E}[h_{0it} - h_{it}]$ directly via BLC of h_0 alone, but the bounds are *very* wide. The buncher ATE balances this trade-off.

[29] Since the bounds depend only on the density around k and the total amount mass to its left/right, point masses elsewhere in the distributions of h_0 and h_1 do not effect on the bounds provided that they are well-separated from k.

Comparison of Theorem 1 to existing results. The existing bunching design literature does contain a few results that are suggestive that bunching is informative about a local average response, when responsiveness to incentives varies by unit. For instance, Saez (2010) and Kleven (2016) consider a "small-kink" approximation that $\mathbb{E}[\Delta_{it}|h_{0it} = k] \approx \mathcal{B}/f_0(k)$. The result requires f_0 to be constant throughout the region $[k, k + \Delta_{it}]$ conditional on each value of Δ_{it}, an assumption that is hard to justify except in the limit that the distribution of Δ_{it} concentrates around zero (Appendix Proposition I.4 and Lemma SMALL make the above claims precise). A kink that produces only tiny responses is unlikely to provide a good approximation in a context like overtime, in which treatment corresponds to a 50% increase in the hourly cost of labor. Nevertheless, even in a "small-kink" setting, Theorem 1 offers a refinement to this approximation: a second-order expansion of $\ln(1 + \frac{x}{a})$ shows that when $\mathcal{B}$ is small, the bounds Δ_k^L and Δ_k^U converge around $\frac{\mathcal{B}-p}{2f_0(k)} + \frac{\mathcal{B}-p}{2f_1(k)}$.

A second existing result is the one of Blomquist et al. (2015) discussed in Section 4.3, which relates the bunching probability to a certain weighted average of compensated elasticities in a non-parametric labor supply model. These authors discuss how such an average could be identified from bunching if for example the density of choices at an income tax kink is assumed to be linear across counterfactual tax rates (or could be bounded if e.g. those densities are assumed to vary monotonically with the tax rate). However the data cannot provide evidence of such linearity, since it identifies this density only for two particular tax rates. By contrast, Theorem 1 requires assumptions only about the distributions of the two counterfactuals that are in fact observed (h_0 and h_1), making the extrapolation problem one of moving "beyond" the kink in h rather than considering alternative counterfactuals at intermediate slopes that are not observed.

4.3.4 Alternative: bounds on the buncher ATE via polynomial extrapolation

BLC imposes no assumptions about the smoothness of the distribution of h_0 and h_1 aside from implying that each have differentiable densities (Dümbgen et al., 2017). Instead, Theorem 1 makes use of the fact that BLC is sufficient to extrapolate upper and lower bounds on each $Q_d(u)$ throughout the bunching region $u \in [F_0(k), F_1(k)]$. An alternative approach to obtaining such bounds would be to impose further smoothness by bounding the magnitude of some derivative of Q_d for each $d \in \{0, 1\}$, while relaxing BLC. This approach similarly yields bounds on the buncher ATE.

Let us assume for example that for each $d \in \{0, 1\}$, Q_d is $m + 1$ times differentiable and that $\sup\{|Q_d^{(m+1)}(u)| : u \in [F_0(k), F_1(k)]\} \leq M$ for some constant M, where $Q_d^{(m)}$ is the m^{th} derivative of Q_d. In words, this assumes that the magnitude of the $(m + 1)^{th}$-order derivative of $Q_d(u)$ is no greater than M throughout the bunching region, for either d. In spirit, this is similar to a suggestion of Bertanha et al. (2023) to bound the derivative of the *density* of h_0, and a related suggestion by Blomquist et al. (2021), in the context of the isoelastic model (see Appendix C.1).

Appendix B.6 derives bounds on the buncher ATE based on the above assumption, which have a width of $M\frac{4\mathcal{B}^{m+1}}{(m+2)!}$ and are symmetric about a central value that extrapolates each Q_d across the bunching region $[F_0(k), F_1(k)]$ using a m^{th} order Taylor approximation. If these quantile functions are assumed to be *analytic* functions of u in this region (with an interval of convergence about k that includes the whole bunching region), then the Taylor series converges and the buncher ATE is in principle point identified by taking $m \to \infty$,[30] with no need for a specified M. However analyticity is a strong identification assumption, and in estimation one must stop at some finite m, introducing extrapolation error.

Appendix B.6 relates the derivatives of Q_d to corresponding powers of the density, which suggests that estimation error may grow with the power m chosen. Indeed, Appendix Table 3 shows that sampling error quickly dominates the effect of partial-identification in determining the width of final confidence intervals for the buncher ATE, even with large M. While this suggests that the usefulness of the polynomial extrapolation method may be somewhat limited in practice, it does provide a test of the robustness of the main results to violations of BLC.

4.4 Estimating policy relevant parameters

The buncher ATE yields the answer to a particular causal question, among a well-defined subgroup of the population. Namely: how would hours among workers bunched at 40 hours by the overtime rule be affected by a counterfactual change from linear pay at their straight-time wage to linear pay at their overtime rate? This section discusses how we may then use this quantity to both evaluate the overall ex-post effect of the FLSA on hours, as well as forecast the impacts of proposed changes to the FLSA. This requires additional assumptions which I continue to approach from a partial identification perspective. These assumptions remain weaker than those required by the isoelastic model, in which the buncher ATE recovers the structural elasticity parameter ϵ.

4.4.1 From the buncher ATE to the ex-post hours effect of the FLSA

To consider the overall ex-post hours effect of the FLSA among covered workers, I proceed in two steps. I first relate the buncher ATE to the overall average effect of introducing the overtime kink, holding fixed the distributions of counterfactual hours h_{0it} and h_{1it}. Then, I allow straight-time wages to themselves be affected by the FLSA, using the buncher ATE again to bound the additional effect of these wage changes on hours.

To motivate this strategy, let us first define the parameter of interest to be the difference in

[30] Pollinger (2023) shows that analyticity is in fact enough to identify both intensive and extensive-margin responses in a "locally" isoelastic model, and points out that any finite mixture of analytic distribution functions is also analytic. Note that strengthening analyticity to suppose that the quantile functions are furthermore *polynomial* would offer a simple justification of the popular estimation approach, following Chetty et al. (2011), of fitting polynomials to the observed distribution (see Appendix I.2 in the online supplemental material).

average weekly hours among hourly workers, with and without the FLSA. Letting h^*_{it} indicate the hours unit it would work absent the FLSA, consider the parameter $\theta := \mathbb{E}[h_{it}] - \mathbb{E}^*[h^*_{it}]$, where the second expectation $\mathbb{E}^*$ is over units of workers that would exist in the no-FLSA counterfactual and be covered were it introduced.[31] Defining θ in this way allows us to remain agnostic as to whether the FLSA changes employment, and hence the population of workers it applies to. However, I assume that the hours among any workers who enter or exit employment due to the FLSA are not systematically different from those who would exist without it, so that we may rewrite θ as $\theta = \mathbb{E}[h_{it} - h^*_{it}]$, averaging over individual-level causal effects in the population that does exist given the FLSA.

Next, decompose θ as:

$$\theta = \mathbb{E}[h_{it}(w_{it}, \mathbf{h}_{-i,t}) - h_{0it}(w^*_{it}, \mathbf{h}^*_{-i,t})] = \mathbb{E}[\underbrace{h_{it}(w_{it}, \mathbf{h}_{-i,t}) - h_{0it}(w_{it}, \mathbf{h}_{-i,t})}_{\text{“effect of the kink”}}]$$
$$+ \mathbb{E}[\underbrace{h_{0it}(w_{it}, \mathbf{h}_{-i,t}) - h_{0it}(w^*_{it}, \mathbf{h}_{-i,t})}_{\text{“wage effects”}}] + \mathbb{E}[\underbrace{h_{0it}(w^*_{it}, \mathbf{h}_{-i,t}) - h_{0it}(w^*_{it}, \mathbf{h}^*_{-i,t})}_{\text{“interdependencies”}}], \qquad (9)$$

where the notation makes explicit the dependence of h and h_0 on the worker's straight-time wage w_{it}, and possibly the hours $\mathbf{h}_{-i}$ of other workers in their firm this week. In the notation of the last section: $h_{it} = h_{it}(w_{it}, \mathbf{h}_{-i,t})$, $h_{0it} = h_{0it}(w_{it}, \mathbf{h}_{-i,t})$ and $h_{1it} = h_{1it}(w_{it}, \mathbf{h}_{-i,t})$. I have used that $h^*_{it} = h_{0it}(w^*_{it}, \mathbf{h}^*_{-i,t})$, since pay is linear in hours in the no-FLSA counterfactual.

The first term in Equation (9) reflects the “effect of the kink” quantity $h_{it} - h_{0it}$ examined in Section 4.2, and I view it as the first-order object of interest. The second term reflects that straight-time wages w_{it} may differ from those that workers would face without the FLSA, denoted by w^*_{it}. The third term is zero when firms' choice of hours for their workers decomposes into separate optimization problems for each unit, as in the benchmark model from Section 4.2. More generally, it will capture any interdependencies in hours across units, for instance due to different workers' hours being not linearly separable in production. In online Appendix G I provide evidence that such effects do not play a large role in θ, and I thus treat this term as zero when estimating θ.[32]

Turning first to the “effect of the kink” term, note that with straight-wages and the hours of

[31] The parameter θ is not an average over individual-level treatment effects, but is instead a causal effect on the population distribution of hours. Note that h^*_{it} in this section differs from the “anticipated” hours quantity h^* in Sec. 2.

[32] In particular, I fail to find evidence of contemporaneous hours substitution in response to colleague sick pay, in an event study design. Another piece of evidence comes from obtaining similar “effect of the kink” estimates across small, medium and large firms, which suggests that a firm's capacity to reallocate hours between existing workers does not tend to drive their hours response to the FLSA. See Appendix G. If the third term of Eq. (9) is not zero, my strategy still estimates the average of a unit-level labor demand elasticity in which the hours of a worker's colleagues are fixed.

other units fixed, the kink only has such direct effects on those units working at least $k = 40$ hours:

$$h_{it} - h_{0it} = \begin{cases} 0 & \text{if} \quad h_{it} < k \\ k - h_{0it} & \text{if} \quad h_{it} = k \\ -\Delta_{it} & \text{if} \quad h_{it} > k \end{cases} \tag{10}$$

and thus $\mathbb{E}[h_{it} - h_{0it}] = \mathcal{B} \cdot \mathbb{E}[k - h_{0it}|h_{it} = k] - P(h_{it} > k)\mathbb{E}[\Delta_{it}|h_{it} > k]$. To identify this quantity we must extrapolate from the buncher ATE to obtain an estimate of $\mathbb{E}[\Delta_{it}|h_{it} > k]$, the average effect for units who work overtime. To do this, I assume that the Δ_{it} of units working more than 40 hours are at least as large on average as those who work exactly 40, but that the reduced-form *elasticity* of their response is no greater than that of the bunchers. The logic is as follows: assuming a constant percentage change between h_0 and h_1 over units would imply responses that grow in proportion to h_1, eventually becoming implausibly large. On the other hand, it would be an underestimate to assume high-hours workers, say at 60 hours, have the same effect in levels $h_0 - h_1$ as those closer to 40. Finally, I use bi-log-concavity of h_0 to put bounds on the average effect of the kink among bunchers $\mathcal{B} \cdot \mathbb{E}[k - h_{0it}|h_{it} = k]$. Details are provided in Appendix J.9.

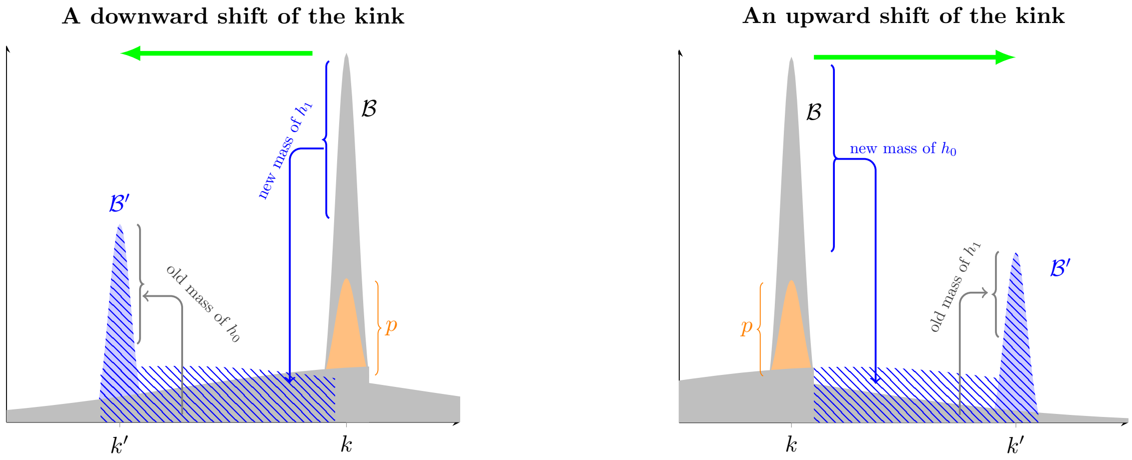
The "wage effects" term in Equation (9) arises because the straight-time wages observed in the data may reflect some adjustment to the FLSA, as we would expect on the basis of the conceptual framework in Section 2.While the "effect of the kink" term is expected to be negative, this second term will be positive if the FLSA causes a reduction in the straight-time wages set at hiring. However, both terms ultimately depend on the same thing: responsiveness of hours to the cost of an hour of work. We can thus use the buncher ATE to compute an approximate upper bound on wage effects by assuming that all straight-time wages are adjusted according to Equation (1) and that the hours response is isoelastic in wages, with anticipated hours approximated by h_{it}. Appendix J.9 provides a visual depiction of the logic. A lower bound on the "wage effects" term, on the other hand, is zero. In practice, the estimated size of the wage effect $\mathbb{E}[h_{0it} - h^*_{0it}]$ is appreciable but still small relative to $\mathbb{E}[h_{it} - h_{0it}]$ (cf. Appendix Table 10).

4.4.2 Forecasting the effects of policy changes

Apart from ex-post evaluation of the overtime rule, policymakers may also be interested in predicting what would happen if the parameters of overtime regulation were modified. Reforms that have been discussed in the U.S. include decreasing "standard hours" k at which overtime pay begins from 40 hours to 35 hours,[33] or increasing the overtime premium from time-and-a-half to "double-time" (Brown and Hamermesh, 2019). This section builds upon Sections 4.1 and 4.3 to show that the bunching-design model is also informative about the impact of such reforms on hours.

[33] Some countries have indeed changed standard hours in recent decades; see Brown and Hamermesh (2019).

Consider changes to standard hours k, for now holding the distributions of h_0 and h_1 fixed across the policy change. Inspection of Equation (2) reveals that as the kink is moved upwards, say from $k = 40$ hours to $k' = 44$ hours, some workers who were previously bunching at k now work h_{0it} hours: namely those for whom $h_{0it} \in [k, k']$. By the same token, some individuals with values of $h_{1it} \in [k, k']$ now bunch at k'. Some individuals who were bunching at k now bunch at k'—namely those for whom $h_{1it} \leq k$ and $h_{0it} \geq k'$. In the case of a reduction in overtime hours, say to $k' = 35$, this logic is reversed. Figure 8 depicts both cases, assuming that the mass of counterfactual bunchers p remains at $k = 40$ after the shift.[34]


A downward shift of the kink

An upward shift of the kink

FIGURE 7: The left panel depicts a shift of the kink point downwards from k to k', while right panel depicts a shift of the kink point upwards. See text for details.

Quantitatively assessing a change to double-time pay requires us to move beyond the two counterfactual choices h_{0it} and h_{1it}: hours that would be worked under straight-wages or under time-and-a-half pay. Recall the notation $h_{it}(\rho)$ introduced in Section 4.3 with $h_{it}(\rho)$ denoting counterfactual hours under a linear pay schedule having a slope of $\rho \cdot w_{it}$ (with w_{it} and hours of other units fixed at their realized levels). Consider a new overtime policy in which a premium pay factor of ρ_1 is due from employers for hours in excess of k, e.g. $\rho_1 = 2$ for a "double-time" policy. Let $h_{it}^{[k,\rho_1]}$ denote realized hours for unit it under this overtime policy as a function of k and ρ_1, and let $\mathcal{B}^{[k,\rho_1]} := P(h_{it}^{[k,\rho_1]} = k)$ be the observable bunching that would occur. I will use ∂_k and ∂_{ρ_1} to denote partial derivatives with respect to k and ρ_1, respectively.

Theorem 2 obtains expressions for the effects of small changes to k or ρ_1 on hours. I continue to assume that counterfactual bunchers $K_{it}^* = 1$ stay at $k^* := 40$, regardless of ρ and k. Let $p(k) = p \cdot \mathbb{1}(k = k^*)$ denote the possible mass of counterfactual bunchers as a function of k.

[34] It is conceivable that some or all counterfactual bunchers locate at 40 because it is the FLSA threshold, while still being non-responsive to the incentives introduced there by the kink. In this case, we might imagine that they would all coordinate on k' after the change. The effects here could then be seen as short-run effects before that occurs.

Theorem 2 (marginal comparative statics in the bunching design). *Under Assumptions CHOICE, CONVEX, SEPARABLE and SMOOTH:*

1. $\partial_k \left\{ \mathcal{B}^{[k,\rho_1]} - p(k) \right\} = f_1(k) - f_0(k)$
2. $\partial_k \mathbb{E}[h_{it}^{[k,\rho_1]}] = \mathcal{B}^{[k,\rho_1]} - p(k)$
3. $\partial_{\rho_1} \mathcal{B}^{[k,\rho_1]} = -k f_{\rho_1}(k) \mathbb{E}\left[\frac{dh_{it}(\rho_1)}{d\rho} \middle| h_{it}(\rho_1) = k \right]$
4. $\partial_{\rho_1} \mathbb{E}[h_{it}^{[k,\rho_1]}] = -\int_k^\infty f_{\rho_1}(h) \mathbb{E}\left[\frac{dh_{it}(\rho_1)}{d\rho} \middle| h_{it}(\rho_1) = h \right] dh$

Proof. See Appendix B. □

The final two assumptions above are given in Appendix B: SEPARABLE requires firm preferences to be quasi-linear in costs, while SMOOTH is a set of regularity conditions which imply that $h_{it}(\rho)$ admits a density $f_\rho(h)$ for all ρ. Theorem 2 also uses a slightly stronger version of Assumption CHOICE that applies to all ρ rather than just ρ_0 and ρ_1. The proof of Theorem 2 builds on results from Blomquist et al. (2015) and Kasy (2022)–see Appendix B for details.

Beginning from the actual FLSA policy of $k = 40 = k^*, \rho_1 = 1.5$, the RHS of Items 1 and 2 are in fact point identified from the data, provided that p is known. Item 1 says that if the location k of the kink is changed marginally, the kink-induced bunching probability will change according to the difference between the densities of h_{1i} and h_{0i} at k^*, which are in turn equal to the left and right limits of the observed density $f(h)$ at the kink. This result is intuitive: given continuity of each potential outcome's density, a small increase in k will result in a mass proportional to $f_1(k)$ being "swept in" to the mass point at the kink, while a mass proportional to $f_0(k)$ is left behind. Item 2 aggregates this change in bunching with the changes to non-bunchers' hours as k is increased: the combined effect turns out to be to simply transport the mass of inframarginal bunchers to the new value of k.[35] Making use of Theorem 2 for a discrete policy change like reducing standard hours to 35 requires integrating across the actual range of hypothesized policy variation. We lose point identification, but I use bi-log-concavity of the marginal distributions of h_0 and h_1 to retain bounds.

Now consider the effect of moving from time-and-a-half to double time on average hours worked, in light of Item 4. This scenario, similar to the effect of the kink term in Eq. (9), requires making assumptions about the response of individuals who may locate far above the kink, and for whom the buncher ATE is less directly informative. Integrating Item 4 over ρ we obtain an

[35] Intuitively, "marginal" bunchers who would choose exactly k under one of the two cost functions B_0 or B_1 cease to "bunch" as k increases, but in the limit of a small change they also do not change their realized h. Moore (2021) gives a closely-related result, derived independently of this work. In the context of a tax kink with $\mathbf{x}$ a scalar and $p(k) = 0$, the result of Moore (2021) generalizes Item 2 of Theorem 2, showing that bunching is a sufficient statistic for the effect of a marginal change in k on tax revenue.

expression for the average effect of this reform in terms of local average elasticities of response:

$$\mathbb{E}[h_{it}^{[k,\rho_1]} - h_{it}^{[k,\bar{\rho}_1]}] = \int_{\rho_1}^{\bar{\rho}_1} \left\{ \int_k^{\infty} f_\rho(h) \cdot h \cdot \mathbb{E}\left[\frac{d \ln h_{it}(\rho)}{d \ln \rho} \middle| h_{it}(\rho) = h \right] dh \right\} d \ln \rho$$

Recall that in the isoelastic model the elasticity quantity $\frac{d \ln h_{it}(\rho)}{d \ln \rho} = \frac{dh_{it}(\rho)}{d\rho} \frac{\rho}{h_{it}(\rho)}$ is constant across ρ and across units, and it is partially identified under BLC. Just as a constant proportional response is likely to overstate responsiveness at large values of hours, it is likely to *understate* responsiveness to larger values of ρ. This yields a lower bound on the effect of moving to double-time. For an upper bound on the magnitude of the effect, I assume rather that in levels $\mathbb{E}[h_{it}(\rho_1) - h_{it}(\bar{\rho}_1) | h_{1it} > k]$ is at least as large as $\mathbb{E}[h_{0it} - h_{1it} | h_{1it} > k]$, and that the increase in bunching from a change of ρ_1 to $\bar{\rho}_1$ is as large as the increase from ρ_0 to ρ_1. Additional details are provided in Appendix J.9.

5 Implementation and Results

This section implements the empirical strategy described in Section 4 with the sample of administrative payroll data described in Section 3.

5.1 Identifying counterfactual bunching at 40 hours

To deliver final estimates of the effect of the FLSA overtime rule on hours, it is necessary to first return to an issue raised in the introduction and allowed for in Section 4: that there are other reasons to expect bunching at 40 hours, in addition to being the location of the FLSA kink. For one, 40 may reflect a kind of *status-quo* choice. This effect could be amplified by firms synchronizing the schedules of different workers, requiring *some* common number of hours per week to coordinate around. Finally, if any salaried workers were not correctly so classified and removed from the sample, hours for such workers might be recorded as 40 even as actual hours worked vary.

In terms of the empirical strategy from Section B.2, all of these alternative explanations manifest in the same way: a point mass p at 40 in the distribution of hours that would occur even if pay did not feature a kink at 40. In the notation introduced in Section 4.3, these "counterfactual bunchers" are demarcated by $K_{it}^* = 1$. Let us refer to the $K_{it}^* = 0$ individuals who also locate at the kink as "active bunchers". The mass of active bunchers is $\mathcal{B} - p$. Theorem 1 shows that we can still partially identify the buncher ATE in the presence of counterfactual bunchers, so long as we know what portion of the total bunchers are active versus counterfactual.

I leverage two strategies to provide plausible estimates for the mass of counterfactual bunchers p. My preferred estimate makes use of the fact that when an employee is paid for hours that are not actually worked—including sick time, paid time off (PTO) and holidays—these hours do not contribute to the 40 hour overtime threshold of the FLSA that week. For example, if a worker applies PTO to miss a six hour shift, then they are not required to be paid overtime until they reach 46 total paid hours in that week. Thus while the kink remains at 40 hours *worked*, non-work hours like PTO shift the location of the kink in hours of *pay*.

The identifying assumption that I rely on is that individuals who still work 40 hours a week, even when they have non-work hours (and are hence paid for more than 40), are all active bunchers: they would not be located at forty hours in the counterfactuals h_{0it} and h_{1it}. This reflects the idea that additional explanations for bunching at 40 hours operate at the level of hours paid, rather than hours worked. Letting n_{it} indicate non-work hours of pay for paycheck it, I make two assumptions:

1. $P(h_{it} = 40 | n_{it} > 0) = P(h_{it} = 40 \text{ and } K^*_{it} = 0 | n_{it} > 0)$
2. $P(h_{it} = 40 \text{ and } K^*_{it} = 0 | n_{it} > 0) = P(h_{it} = 40 \text{ and } K^*_{it} = 0 | n_{it} = 0)$

The first item reflects the above logic, and allows me to identify the mass of active bunchers in the $n_{it} > 0$ conditional distribution of hours. The second item says that this conditional mass is representative of the unconditional mass of active bunchers. To increase the plausibility of this assumption, I focus on η as paid time off because it is generally planned in advance, yet has somewhat idiosyncratic timing.[36]

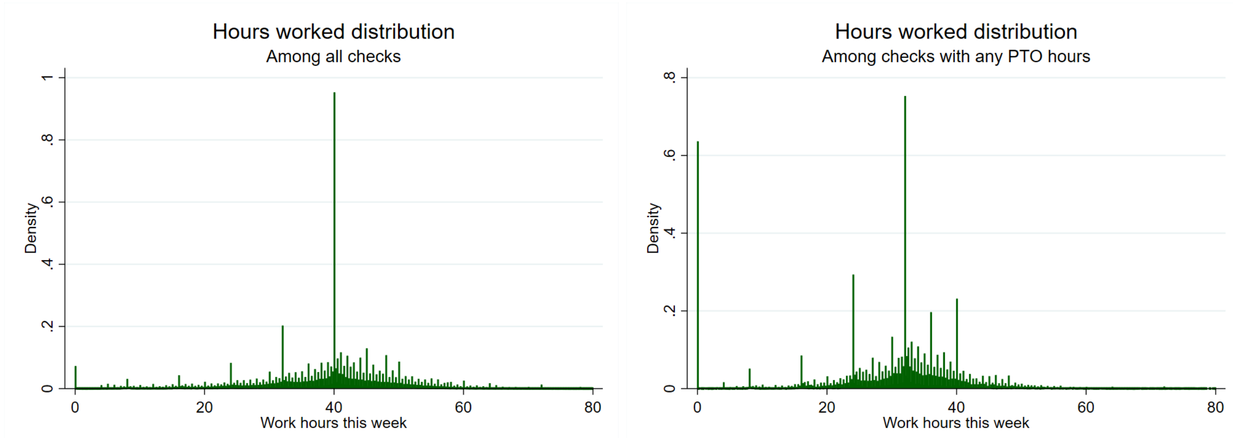


FIGURE 8: The right panel shows a histogram of hours worked when paid time off hours are positive ($\eta_{it} > 0$). The left panel shows the unconditional distribution. While $\mathcal{B} \approx 11.6\%$, $P(h_{it} = 40 | n_{it} > 0) \approx 2.7\%$.

Together, the two assumptions above imply that $p = P(K^*_{it} = 1 \text{ and } h_{it} = 40)$ is identified as $\mathcal{B} - P(h_{it} = 40 | \eta_{it} > 0)$. Figure 8 shows the conditional distribution of hours paid for work when the paycheck contains a positive number of PTO hours ($n_{it} > 0$). The figure reveals that

[36] By contrast, sick pay is often unanticipated so the firm may not be able to re-optimize total hours within the week in which a worker calls in sick. Holiday pay is known in advance, but holidays are unlikely to be representative in terms of other factors important for hours determination (e.g. product demand).

when moving from the unconditional (left panel) to positive-PTO conditional (right panel) distribution, most of the point mass at 40 hours moves away, largely concentrating now at 32 hours (corresponding to the PTO covering eight hours). Of the total bunching of $\mathcal{B} \approx 11.6\%$ in the unconditional distribution, I estimate that only $P(h_{it} = 40|n_{it} > 0) \approx 2.7\%$ are active bunchers, leaving $p \approx 8.9\%$. Thus roughly three quarters of the individuals at 40 hours are counterfactual rather than active bunchers.

As a secondary strategy, I estimate an upper bound for p by using the assumption that the potential outcomes of counterfactual bunchers are relatively "sticky" over time. If the hours of counterfactual bunchers are at 40 for behavioral or administrative reasons, it is reasonable to assume that these external considerations are fairly static, preventing latent hours h_{0it} from changing much between adjacent weeks. In particular, assume that in a given week t nearly all of the counterfactual bunchers are also "non-changers" of hours from week $t-1$. Then:

$$p = P(h_{0it} = 40) \approx P(h_{0it} = h_{0it-1} = 40) \leq P(h_{it} = h_{i,t-1} = 40),$$

where the inequality follows from $(h_{0it} = 40) \implies (h_{it} = 40)$ by Lemma B.1. The probability $P(h_{it} = h_{i,t-1} = 40)$ can be directly estimated from the data, yielding $p \leq 6\%$.

5.2 Estimation and inference

Given Theorem 1 and a value of p, computing bounds on the buncher ATE requires estimates of the right and left limits of the CDF and density of hours at the kink. I use the local polynomial density estimator of Cattaneo, Jansson and Ma (2020) (CJM), which is well-suited to estimating a CDF and its derivatives at boundary points. A local-linear CJM estimator of the left limit of the CDF and density at k, for instance, is:

$$(\hat{F}_-(k), \hat{f}_-(k)) = \underset{(b_1,b_2)}{\text{argmin}} \sum_{it:h_{it}<k} (F_n(h_{it}) - b_1 - b_2 h_{it})^2 \cdot K\left(\frac{h_{it}-k}{\alpha}\right) \tag{11}$$

where $F_n(y) = \frac{1}{n}\sum_{it} \mathbb{1}(h_{it} \leq y)$ is the empirical CDF of a sample of size n, $K(\cdot)$ is a kernel function, and α is a bandwidth. The right limits $F_+(k)$ and $f_+(k)$ are estimated analogously using observations for which $h_{it} > k$. I use a triangular kernel, and choose h as follows: first, I use CJM's mean-squared error minimizing bandwidth selector to produce bandwidth choices for the left and right limits at $k = 40$. I then average the two bandwidths, and use this common bandwidth in the final calculation of both limits. In the full sample, the bandwidth chosen by this procedure is about 1.7 hours, and is somewhat larger for estimates that condition on a single industry.

To construct confidence intervals for parameters that are partially identified (e.g. the buncher ATE), I use adaptive critical values proposed by Imbens and Manski (2004) and Stoye (2009) that are valid for the underlying parameter. To easily incorporate sampling uncertainty in all of $\hat{F}_-(k), \hat{f}_-(k), \hat{F}_+(k), \hat{f}_+(k)$ and $\hat{p}$, I estimate variances by a cluster nonparametric bootstrap that resamples at the firm level. This allows arbitrary autocorrelation in hours across pay periods for a single worker, and between workers within a firm. All standard errors use 500 bootstrap samples.

5.3 Results of the bunching estimator: the buncher ATE

Table 4 reports treatment effect estimates based on Theorem 1, when p is either assumed to be zero or is estimated by one of the two methods described in Section 5.1. The first row reports the corresponding estimate of the net bunching probability $\mathcal{B} - p$, while the second row reports the bounds on the buncher ATE $\mathbb{E}[h_{0it} - h_{1it}|h_{it} = k, K^*_{it} = 0]$. Within a fixed estimate of p, the bounds on the buncher ATE based on bi-log-concavity are quite informative: the upper and lower bounds are close to each other and precisely estimated. One can show from the expressions for the bounds in Theorem 1 that if $f_0(k) \approx f_1(k)$ and $p \approx 0$, the bounds will tend to be narrower when $F_0(k)$ is closer to $(1 - \mathcal{B})/2$, i.e. the kink is close to the median of the latent hours distribution. This provides some intuition for why the bounds are reasonably narrow, since hours are roughly evenly divided to either side of 40 hours (cf. Figure 2).

	p=0	*p* from non-changers	*p* from PTO
Net bunching:	0.116	0.057	0.027
	[0.112, 0.120]	[0.055, 0.058]	[0.024, 0.030]
Buncher ATE	[2.614, 3.054]	[1.324, 1.435]	[0.640, 0.666]
	[2.493, 3.205]	[1.264, 1.501]	[0.574, 0.736]
Num observations	630217	630217	630217
Num clusters	566	566	566

TABLE 4: Estimates of net bunching $\mathcal{B} - p$ and the buncher ATE: $\Delta^*_k = \mathbb{E}[h_{0it} - h_{1it}|h_{it} = k, K^*_{it} = 0]$, across various strategies to estimate counterfactual bunching $p = P(K^*_{it} = 1)$. Unit of analysis is a paycheck, and 95% bootstrap confidence intervals (in gray) are clustered by firm.

The PTO-based estimate of p provides the most conservative treatment effect estimate, attributing roughly one quarter of the observed bunching to active rather than counterfactual bunchers. Nevertheless, this estimate still yields a highly statistically significant buncher ATE of about 2/3 of an hour, or 40 minutes. This estimate has the following interpretation: consider the group of workers that are in fact working 40 hours in a given pay period and are not counterfactual bunchers. Firms would ask this group to work on average about 40 minutes more that week if they were paid their straight-time wage for all hours, compared with a counterfactual in which they are paid their overtime rate for all hours. If we instead attribute all of the observed bunching mass to active bunchers ($p = 0$), then the buncher ATE is estimated to be at least 2.6 hours. In Appendix E I also report estimates based on alternative shape constraints and assumptions about effect heterogeneity (with similar results), including the method described in Section 4.3.4 of polynomial extrapolation of quantiles (see Table 3).

5.4 Estimates of policy effects

I now use estimates of the buncher ATE and the results of Section 4.4 to estimate the overall causal effect of the FLSA overtime rule, and simulate changes based on modifying standard hours or the premium pay factor. Table 5 first reports an estimate of the buncher ATE expressed as a reduced-form hours demand elasticity,[37] which I use as an input in these calculations. The next two rows report bounds on $\mathbb{E}[h_{it} - h^*_{it}]$ and $\mathbb{E}[h_{it} - h^*_{it}|h_{1it} \geq 40, K^*_{it} = 0]$, respectively. The second row is the overall ex-post effect of the FLSA on hours, averaged over workers and pay periods, and the third row conditions on paychecks reporting at least 40 hours (omitting counterfactual bunchers). The final row reports an estimate of the effect of moving to double-time pay.

	p=0	*p* from non-changers	*p* from PTO
Buncher ATE as elasticity	[-0.188,-0.161] [-0.198,-0.154]	[-0.088,-0.082] [-0.093,-0.078]	[-0.041,-0.039] [-0.045,-0.035]
Average effect of FLSA on hours	[-1.466, -1.026] [-1.535, -0.977]	[-0.727, -0.486] [-0.762, -0.463]	[-0.347, -0.227] [-0.384, -0.203]
Avg. effect among directly affected	[-2.620, -1.833] [-2.733, -1.750]	[-1.453, -0.972] [-1.518, -0.929]	[-0.738, -0.483] [-0.812, -0.434]
Double-time, average effect on hours	[-2.604, -0.569] [-2.707, -0.547]	[-1.239, -0.314] [-1.285, -0.300]	[-0.580, -0.159] [-0.638, -0.143]

TABLE 5: Estimates of the buncher ATE expressed as an elasticity, the average ex-post effect of the FLSA $\mathbb{E}[h_{it} - h^*_{it}]$,[37] the effect among directly affected units $\mathbb{E}[h_{it} - h^*_{it}|h_{it} \geq k, K^*_{it} = 0]$ and predicted effects of a change to double-time. 95% bootstrap confidence intervals in gray, clustered by firm.

[37] This is $\hat{\Delta}^*_k/(40\ln(1.5))$ where $\hat{\Delta}_k$ is the estimate of the buncher ATE presented in Table 4. This is numerically equivalent to the elasticity implied by the buncher ATE in logs $\mathbb{E}[\ln h_{0it} - \ln h_{1it}|h_{it} = k, K^*_{it} = 0]/(\ln 1.5)$ estimated under assumption that $\ln h_0$ and $\ln h_1$ are BLC.

Taking the PTO-based estimate of p as yielding a lower bound on treatment effects, the estimates suggest that workers work at least about 1/4 of an hour less on average in a given week than they would absent overtime regulation: about one third the magnitude of the buncher ATE. When I focus on those workers that are directly affected in a given week, the figure is about twice as high: roughly 30 minutes. Since my data has been restricted to hourly workers paid on a weekly basis, these estimates should be interpreted as holding for that population only. While one might assume that similar effects hold for hourly workers paid at other intervals (e.g. bi-weekly), speaking to the hours effects of the FLSA on salary workers is beyond the scope of this study.

Table 5 also suggests that a move to double-time pay would introduce a further reduction in hours comparable to the existing ex-post effect of the FLSA, but the bounds are wider. These estimates include the effects of possible adjustments to straight-time wages, which tend to attenuate the impact of the policy change. Appendix Table 10 replicates Table 5 neglecting wage adjustments, which might be viewed as a short-run response to the FLSA before wages adjust.

Figure 9 breaks down estimates of the ex-post effect of the overtime rule by major industries, revealing considerable heterogeneity between them. The estimates suggest that Real Estate & Rental and Leasing as well as Wholesale Trade see the highest average reduction in hours. The least-affected industries are Health Care and Social Assistance and Professional Scientific and Technical, with the average worker working just about 6 minutes less per week due to the overtime rule. Appendix E reports estimates broken down by gender, finding that the FLSA has considerably higher effects on the hours of men compared with women.

Appendix Figure 6 looks at the effect of changing the threshold for overtime hours k from 40 to alternative values k'. Moving standard hours to 35 is thus predicted to completely eliminate bunching due to the overtime kink in the short run, before any adjustment to latent hours (e.g. through changes to straight-time wages). Even for the preferred estimate of p from PTO, increasing the overtime threshold as high as 43 hours is estimated to increase average working hours by an amount distinguishable from zero.

6 Implications of the estimates for overtime policy

The estimates from the preceding section suggest that FLSA regulation indeed has real effects on hours worked, in line with labor demand theory when wages do not fully adjust to absorb the added cost of overtime hours. When averaged over affected workers and across pay periods, I find that hourly workers in my sample work at least 30 minutes less per week than they would without the overtime rule. This lower bound is broadly comparable to the few causal estimates that exist in the literature, including Hamermesh and Trejo (2000) who assess the effects of expanding California's daily overtime rule to cover men in 1980, and Brown and Hamermesh (2019) who use the erosion of

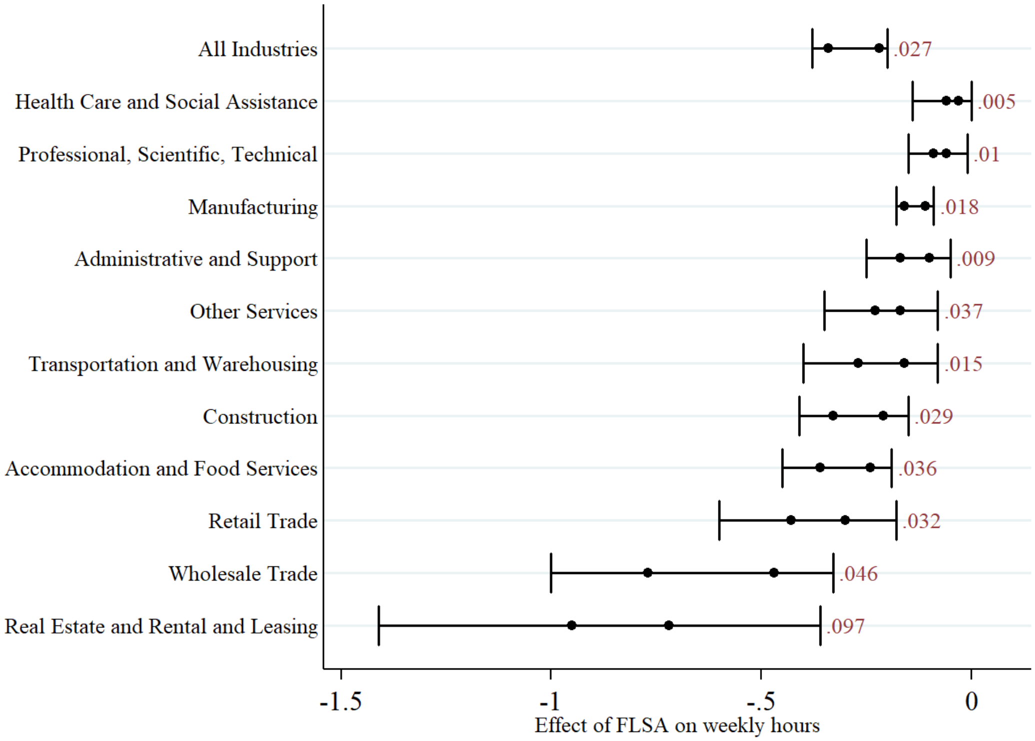


FIGURE 9: 95% confidence intervals for the effect of the FLSA on hours by industry, using PTO-based estimates of p for each. Dots are point estimates of the upper and lower bounds. The number to the right of each range is the point estimate of the net bunching $\mathcal{B} - p$ for that industry.

the salary threshold for exemption of white-collar jobs in real terms over the last several decades.[38] By contrast, my estimates use an identification strategy that does not require focusing on the subpopulation affected by a natural experiment, and are based on recent and administrative data.

My estimates speak to the substitutability of hours of labor between workers. The primary justifications for overtime regulation have been to reduce excessive workweeks, while encouraging hours to be distributed over more workers (Ehrenberg and Schumann, 1982). How well this plays out in practice hinges on how easily an hour of work can be moved from one worker to another or across time, from the perspective of the firm. The results of this paper find hours demand to be relatively inelastic: hours cannot be easily so reallocated between workers or weeks. This suggests that ongoing efforts to expand coverage of the FLSA overtime rule may have limited scope to dramatically affect the hours of U.S. workers.

Nevertheless, the overall impact of the FLSA overtime rule on workers is still notable. The

[38] Hamermesh and Trejo (2000) and Brown and Hamermesh (2019) report estimates of -0.5 and -0.18 for the elasticity of overtime hours with respect to the overtime rate. My preferred estimate of -0.04 for the buncher ATE as an elasticity is the elasticity of *total* hours, including the first 40. An elasticity of overtime hours can be computed from this using the ratio of mean hours to mean overtime hours in the sample, resulting in an estimate of roughly -0.45.

data suggest that at least about 3% and as many as about 12% of workers' hours are adjusted to the threshold introduced by the policy, indicating that it may have distortionary impacts for a significant portion of the labor force. The policy may also have important effects on unemployment. While an assessment of the employment effects of the FLSA overtime rule is beyond the scope of this paper, my estimates of the hours effect can be used to build a back-of-the-envelope calculation, following Hamermesh (1993). As detailed in Appendix E.5 I assume a value for the rate at which firms substitute labor for capital to obtain a "best-guess" estimate that the FLSA overtime rule creates about 700,000 jobs. To get an overall upper bound on the size of employment effects, one can instead attribute *all* of the bunching at 40 to the FLSA and assume that the total number of worker-hours is not reduced by the FLSA. By this estimate the FLSA increases employment by at most 3 million jobs, or roughly 3% among covered workers. A reasonable range of parameter values in this simple calculation rules out that the FLSA overtime rule has negative overall employment effects on hourly workers.

7 Conclusion

This paper has provided a new interpretation of the popular bunching-design method in the language of treatment effects, showing that the basic identifying power of the method is robust to a wide variety of underlying choice models. Across such models, the parameter of interest remains a reduced-form average treatment effect (local to the kink) between two appropriately-defined counterfactual choices, which is partially identified under a natural nonparametric assumption about those counterfactuals' distributions. This provides conditions under which the bunching design can be useful to answer program evaluation questions in a variety of contexts, particularly beyond those in which the researcher is prepared to posit a parametric model of agents' preferences.

By leveraging these insights with a new payroll dataset recording exact weekly hours paid at the individual level, I estimate that U.S. hourly workers subject to the Fair Labor Standard Act work shorter hours due to its overtime provision, which may lead to positive employment effects. Given the large amount of within-worker variation in hours observed, the modest size of the FLSA effects estimated in this paper suggest that firms do face significant incentives to maintain longer working hours, countervailing against the ones introduced by policies intended to reduce them.

References

BARKUME, A. (2010). "The Structure of Labor Costs with Overtime Work in U.S. Jobs". *Industrial and Labor Relations Review* 64 (1).

BARLOW, R., PROSCHAN, F. and HUNTER, L. (1996). *Mathematical Theory of Reliability*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.

BERTANHA, M., MCCALLUM, A. H. and SEEGERT, N. (2023). “Better bunching, nicer notching”. *Journal of Econometrics* 237 (2, Part A), p. 105512.

BERTHANA, M., CAETANO, C., JALES, H. and SEEGERT, N. (2023). “Bunching Estimation Methods”. *Handbook of Labor, Human Resources and Population Economics* (forthcoming).

BEST, M. C., BROCKMEYER, A., KLEVEN, H. J., SPINNEWIJN, J. and WASEEM, M. (2015). “Production vs Revenue Efficiency With Limited Tax Capacity: Theory and Evidence From Pakistan”. *Journal of Political Economy* 123 (6), p. 48.

BICK, A., BLANDIN, A. and ROGERSON, R. (2022). “Hours and Wages”. *The Quarterly Journal of Economics* (forthcoming).

BISHOW, J. L. (2009). “A Look at Supplemental Pay: Overtime Pay, Bonuses, and Shift Differentials”. *Monthly Labor Review*. Publisher: Bureau of Labor Statistics, U.S. Department of Labor.

BLOCK, H. W., SAVITS, T. H. and SINGH, H. (1998). “The Reversed Hazard Rate Function”. *Probability in the Engineering and Informational Sciences* 12 (1), 69–90.

BLOMQUIST, S., KUMAR, A., LIANG, C.-Y. and NEWEY, W. (2015). “Individual heterogeneity, nonlinear budget sets and taxable income”. *The Institute for Fiscal Studies Working Paper* CWP21/15.

— (2021). “On Bunching and Identification of the Taxable Income Elasticity”. *Journal of Political Economy* 129 (8).

BLOMQUIST, S., NEWEY, W., KUMAR, A. and LIANG, C.-Y. (2017). *On Bunching and Identification of the Taxable Income Elasticity*. Working Paper 24136. National Bureau of Economic Research.

BRECHLING, F. P. R. (1965). “The Relationship Between Output and Employment in British Manufacturing Industries”. *The Review of Economic Studies* 32 (3), p. 187.

BROWN, C. and HAMERMESH, D. S. (2019). “Wages and Hours Laws: what do we know? what can be done?” *The Russell Sage Foundation Journal of the Social Sciences* 5 (5), pp. 68–87.

BURDETT, K. and MORTENSEN, D. T. (1998). “Wage Differentials, Employer Size, and Unemployment”. *International Economic Review* 39 (2), p. 257.

BUREAU OF LABOR STATISTICS (2020). *National Compensation Survey (Restricted-Use Microdata)*. www.bls.gov/ncs/ncs-data-requests.htm.

CAHUC, P. and ZYLBERBERG, A. (2014). *Labor economics*. 2nd. Cambridge, Mass.: MIT Press.

CATTANEO, M. D., JANSSON, M. and MA, X. (2020). "Simple Local Polynomial Density Estimators". *Journal of the American Statistical Association* 115 (531), pp. 1449–1455.

CHERNOZHUKOV, V. and HANSEN, C. (2005). "An IV Model of Quantile Treatment Effects". *Econometrica* 73 (1), pp. 245–261.

CHETTY, R., FRIEDMAN, J. N., OLSEN, T. and PISTAFERRI, L. (2011). "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." *Quarterly Journal of Economics* 126 (2), pp. 749–804.

COSTA, D. L. (2000). "Hours of Work and the Fair Labor Standards Act: A Study of Retail and Wholesale Trade, 1938–1950". *Industrial and Labor Relations Review*, p. 17.

DUBE, A., MANNING, A. and NAIDU, S. (2020). "Monopsony, Misoptimization, and Round Number Bunching in the Wage Distribution". *NBER Working Paper* w24991.

DÜMBGEN, L., KOLESNYK, P. and WILKE, R. A. (2017). "Bi-log-concave distribution functions". *Journal of Statistical Planning and Inference* 184, pp. 1–17.

EHRENBERG, R. and SCHUMANN, P. (1982). *Longer hours or more jobs? : an investigation of amending hours legislation to create employment*. New York State School of Industrial and Labor Relations, Cornell University.

EHRENBERG, R. G. (1971). "The Impact of the Overtime Premium on Employment and Hours in U . S . Industry". *Economic Inquiry* 9 (2).

EINAV, L., FINKELSTEIN, A. and SCHRIMPF, P. (2017). "Bunching at the kink: Implications for spending responses to health insurance contracts". *Journal of Public Economics* 146, pp. 27–40.

GRIGSBY, J., HURST, E. and YILDIRMAZ, A. (2021). "Aggregate Nominal Wage Adjustments: New Evidence from Administrative Payroll Data". 11 (2), pp. 428–71.

GUPTA, R. D. and NANDA, A. K. (2001). "Some results on reversed hazard rate". *Communications in Statistics - Theory and Methods* 30 (11), pp. 2447–2457. eprint: `https://www.tandfonline.com/doi/pdf/10.1081/STA-100107697`.

HAMERMESH, D. S. (1993). *Labor demand*. Princeton, NJ: Princeton Univ. Press.

HAMERMESH, D. S. and TREJO, S. J. (2000). “The Demand for Hours of Labor : Direct Evidence from California”. *The Review of Economics and Statistics* 82 (1), pp. 38–47.

HART, R. A. (2004). *The economics of overtime working*. Cambridge, UK: Cambridge University Press.

HJORT, J., LI, X. and SARSONS, H. (2020). “Across-Country Wage Compression in Multinationals”. *NBER Working Paper* w26788.

IMBENS, G. W. and MANSKI, C. F. (2004). “Confidence Intervals for Partially Identified Parameters”. *Econometrica* 72, p. 14.

JOHNSON, J. (2003). “The Impact of Federal Overtime Legislation on Public Sector Labor Markets”. *Journal of Labor Economics* 21 (1), pp. 43–69.

KASY, M. (2022). “Who wins, who loses? Identification of the welfare impact of changing wages”. *Journal of Econometrics* 226 (1), pp. 1–26.

KEILSON, J. (1971). “Log-Concavity and Log-Convexity in Passage Time Densities of Diffusion and Birth-Death Processes”. *Journal of Applied Probability* 8 (2), pp. 391–398.

KIJIMA, M. (1998). “Hazard Rate and Reversed Hazard Rate Monotonicities in Continuous-Time Markov Chains”. *Journal of Applied Probability* 35 (3), pp. 545–556.

KLEVEN, H. J. (2016). “Bunching”. *Annual Review of Economics* 8, pp. 435–464.

KLEVEN, H. J. and WASEEM, M (2013). “Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan”. *The Quarterly Journal of Economics* 128 (2), pp. 669–723.

KLINE, P. and TARTARI, M. (2016). “Bounding the Labor Supply Responses to a Randomized Welfare Experiment: A Revealed Preference Approach”. *American Economic Review* 106 (4), pp. 972–1014.

KÉDAGNI, D. and MOURIFIÉ, I. (2020). “Generalized instrumental inequalities: testing the instrumental variable independence assumption”. *Biometrika* 107 (3), pp. 661–675. eprint: `https://academic.oup.com/biomet/article-pdf/107/3/661/33658405/asaa003.pdf`.

MATOS, K., GALINSKY, E. and BOND, J. (2017). “National Study of Employers”. *Society for Human Resource Management*, p. 79.

MILGROM, P. and ROBERTS, J. (1996). “The LeChatelier Principle”. *American Economic Review* 1 (86), pp. 173–179.

MOORE, D. T. (2021). “Evaluating Tax Reforms without Elasticities: What Bunching Can Identify”. *Mimeo*, p. 61.

PENCAVEL, J. (2015). “THE PRODUCTIVITY OF WORKING HOURS”. *The Economic Journal* 125 (589), pp. 2052–2076.

POLLINGER, S. (2023). “Kinks Know More: Policy Evaluation Beyond Bunching with an Application to Solar Subsidies”. (hal-04182085).

QUACH, S. (2024). “The Labor Market Effects of Expanding Overtime Coverage”. *SSRN Working Paper* 100613.

ROSEN, S. (1968). “Short-Run Employment Variation on Class-I Railroads in the U.S., 1947-1963”. *Econometrica* 36 (3), p. 511.

SAEZ, E. (2010). “Do Taxpayers Bunch at Kink Points?” *American Economic Journal: Economic Policy* 2 (3), pp. 180–212.

SAUMARD, A. (2019). “Bi-log-concavity: some properties and some remarks towards a multi-dimensional extension”. *arXiv:1903.07347*.

SAUMARD, A. and WELLNER, J. A. (2014). “Log-concavity and strong log-concavity: A review”. *Statistics Surveys* 8 (none), pp. 45 –114.

STOLE, L. A. and ZWIEBEL, J. (1996). “Intra-Firm Bargaining under Non-Binding Contracts”. *The Review of Economic Studies* 63 (3), pp. 375–410.

STOYE, J. (2009). “More on Confidence Intervals for Partially Identified Parameters”. *Econometrica* 77 (4), pp. 1299–1315.

TAYLOR, H. and KARLIN, S. (1994). *An Introduction to Stochastic Modeling*. Academic Press.

TREJO, B. S. J. (1991). “The Effects of Overtime Pay Regulation on Worker Compensation”. *American Economic Review* 81 (4), pp. 719–740.

U.S. DEPARTMENT OF LABOR (2024). *Defining and Delimiting the Exemptions for Executive, Administrative, Professional, Outside Sales, and Computer Employees.* https://www.federalregister.gov/documents/2024/04/26/2024-08038/defining-and-delimiting-the-exemptions-for-executive-administrative-professional-outside-sales-and.

A Proof of Theorem 1

In proving Theorem 1, we may relax somewhat the assumption that h_1 and h_0 are everywhere BLC (conditional on $K^* = 0$). What we in fact need is for these distributions to each be "locally" BLC over a particular region containing the kink. This relaxation may be of interest when motivating the BLC assumption, as discussed in Appendix C.

Call a distribution function $F(h)$ *locally BLC* on an interval N, if $\ln F(h)$ and $\ln(1 - F(h))$ are concave for $h \in N$. In what follows we assume that conditional on $K^*_{it} = 0$, h_0 is BLC on $[k, k + \Delta^*_0]$ and h_1 is BLC on $[k - \Delta^*_1, k]$. This is of course implied if these distributions are globally BLC as assumed in the main text. The constants Δ^*_0 and Δ^*_1 come from Assumption RANK.

Let $\mathcal{B}^* := P(h_{it} = k|K^* = 0)$. Recall that RANK says that there exist positive Δ^*_0 and Δ^*_1 such that $h_{0it} \in [k, k + \Delta^*_0]$ holds if and only if $h_{1it} \in [k - \Delta^*_1, k]$. Note that either of these conditions then implies $h_{it} = k$, by Eq. (2). Meanwhile if $h_{it} \neq k$ then either $h_{0it} < k$ or $h_{1it} > k$ by Eq. (2), and so neither $h_{0it} \in [k, k + \Delta^*_0]$ nor $h_{1it} \in [k - \Delta^*_1, k]$ can hold given RANK. Thus under Assumption RANK both $h_{0it} \in [k, k + \Delta_{it}]$ and $h_{1it} \in [k - \Delta^*_1, k]$ are equivalent to $h_{it} = k$.

Under RANK, we can therefore write the buncher ATE as:

$$\begin{aligned}
E[h_{0it} - h_{1it}|h_{it} = k, K^*_{it} = 0] &= E[h_{0it}|h_{0it} \in [k, k + \Delta^*_0], K^*_{it} = 0] - E[h_{1it}|h_{1it} \in [k - \Delta^*_1, k], K^*_{it} = 0] \\
&= \frac{1}{\mathcal{B}^*}\int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k+\Delta^*_0)} Q_{0|K^*=0}(u)du - \frac{1}{\mathcal{B}^*}\int_{F_{1|K^*=0}(k-\Delta^*_1)}^{F_{1|K^*=0}(k)} Q_{1|K^*=0}(v)dv \\
&= \frac{1}{\mathcal{B}^*}\int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k+\Delta^*_0)} \{Q_{0|K^*=0}(u) - k\}du + \frac{1}{\mathcal{B}^*}\int_{F_{1|K^*=0}(k-\Delta^*_1)}^{F_{1|K^*=0}(k)} \{k - Q_{1|K^*=0}(v)\}dv
\end{aligned}$$

That $\mathcal{B}^* = F_{0|K^*=0}(k + \Delta^*_0) - F_{0|K^*=0}(k) = F_{1|K^*=0}(k) - F_{1|K^*=0}(k - \Delta^*_1)$ under Assumption RANK follows from Equation (2), which is established as Lemma B.1 in Appendix B. A proof of Lemma B.1 is provided in the supplemental material.

Consider first the implication of local BLC that $F_{d|K^*=0}(h)$ is log-concave on an interval between k and $k + t$ for some t. In what follows, we will consider positive $t \in [0, \Delta^*_0]$ for $d = 0$ and negative such $t \in [-\Delta^*_1, 0]$ for $d = 1$). Concavity implies that a first-order Taylor expansion for $\log F_{d|K^*=0}(k + t)$ around k overshoots: i.e. $\log F_{d|K^*=0}(k + t) \leq \log F_{d|K^*=0}(k) + t \cdot \frac{d}{dh}\log F_{d|K^*=0}(k)$. Similarly, that $\log(1 - F_{d|K^*=0}(h))$ is concave on an interval $[k, k + t]$ implies that $\log(1-F_{d|K^*=0}(k+t)) \leq \log(1-F_{d|K^*=0}(k)) + t \cdot \frac{d}{dh}\log(1-F_{d|K^*=0}(k))$. These two inequalities can be rearranged to put upper and lower bounds on $F_{d|K^*=0}(k+t)$:

$$1-(1-F_{d|K^*=0}(k))e^{-\frac{f_{d|K^*=0}(k)}{1-F_{d|K^*=0}(k)}t} \leq F_{d|K^*=0}(k+t) \leq F_{d|K^*=0}(k)e^{\frac{f_{d|K^*=0}(k)}{F_{d|K^*=0}(k)}t} \tag{A.1}$$

An analagous expression is obtained in Theorem 1 of Dümbgen et al. (2017).

Defining $u = F_{0|K^*=0}(k+t)$, we can use the substitution $t = Q_{0|K^*=0}(u) - k$ to translate the above into bounds on the conditional quantile function of h_{0it}, evaluated at u:

$$\frac{F_{0|K^*=0}(k)}{f_{0|K^*=0}(k)} \cdot \ln\left(\frac{u}{F_{0|K^*=0}(k)}\right) \leq Q_{0|K^*=0}(u) - k \leq -\frac{1-F_{0|K^*=0}(k)}{f_{0|K^*=0}(k)} \cdot \ln\left(\frac{1-u}{1-F_{0|K^*=0}(k)}\right) \tag{A.2}$$

And similarly for h_1, letting $v = F_{1|K^*=0}(k-t)$:

$$\frac{1-F_{1|K^*=0}(k)}{f_{1|K^*=0}(k)} \cdot \ln\left(\frac{1-v}{1-F_{1|K^*=0}(k)}\right) \leq k - Q_{1|K^*=0}(v) \leq -\frac{F_{1|K^*=0}(k)}{f_{1|K^*=0}(k)} \cdot \ln\left(\frac{v}{F_{1|K^*=0}(k)}\right) \tag{A.3}$$

A lower bound for $E[h_{0it} - h_{1it} | h_{it} = k, K^*_{it} = 0]$ is thus:

$$\begin{aligned}
&\frac{F_{0|K^*=0}(k)}{f_{0|K^*=0}(k)\cdot\mathcal{B}^*}\int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k)+\mathcal{B}^*} \ln\left(\frac{u}{F_{0|K^*=0}(k)}\right)du + \frac{1-F_{1|K^*=0}(k)}{f_{1|K^*=0}(k)\cdot\mathcal{B}^*}\int_{F_{1|K^*=0}(k)-\mathcal{B}^*}^{F_{1|K^*=0}(k)} \ln\left(\frac{1-v}{1-F_{1|K^*=0}(k)}\right)dv \\
&= g(F_{0|K^*=0}(k), f_{0|K^*=0}(k), \mathcal{B}^*) + h(F_{1|K^*=0}(k), f_{1|K^*=0}(k), \mathcal{B}^*)
\end{aligned}$$

where

$$\begin{aligned}
g(a,b,x) &:= \frac{a}{bx}\int_a^{a+x} \ln\left(\frac{u}{a}\right)du = \frac{a^2}{bx}\int_1^{1+\frac{x}{a}} \ln(u)\,du \\
&= \frac{a^2}{bx}\left\{u\ln(u) - u\right\}\Big|_1^{1+\frac{x}{a}} = \frac{a^2}{bx}\left\{\left(1+\frac{x}{a}\right)\ln\left(1+\frac{x}{a}\right) - \frac{x}{a}\right\} = \frac{a}{bx}(a+x)\ln\left(1+\frac{x}{a}\right) - \frac{a}{b}
\end{aligned}$$

and

$$h(a,b,x) := \frac{1-a}{bx}\int_{a-x}^{a} \ln\left(\frac{1-v}{1-a}\right)dv = \frac{(1-a)^2}{bx}\int_1^{1+\frac{x}{1-a}} \ln(u)\,du = g(1-a,b,x)$$

Similarly, an upper bound is:

$$
\begin{aligned}
&-\frac{1-F_{0|K^*=0}(k)}{f_{0|K^*=0}(k)(\mathcal{B}^*)}\int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k)+\mathcal{B}^*}\ln\left(\frac{1-u}{1-F_{0|K^*=0}(k)}\right)du \\
&\qquad\qquad\qquad -\frac{F_{1|K^*=0}(k)}{f_{1|K^*=0}(k)(\mathcal{B}^*)}\int_{F_{1|K^*=0}(k)-\mathcal{B}^*}^{F_{1|K^*=0}(k)}\ln\left(\frac{v}{F_{1|K^*=0}(k)}\right)dv \\
&= \tilde{g}(F_{0|K^*=0}(k), f_{0|K^*=0}(k), \mathcal{B}^*) + \tilde{h}(F_{1|K^*=0}(k), f_{1|K^*=0}(k), \mathcal{B}^*)
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{g}(a,b,x) &:= -\frac{1-a}{bx}\int_a^{a+x}\ln\left(\frac{1-u}{1-a}\right)du = -\frac{(1-a)^2}{bx}\int_{1-\frac{x}{1-a}}^{1}\ln(u)\,du \\
&= \frac{(1-a)^2}{bx}\left\{u-u\ln(u)\right\}\Big|_{1-\frac{x}{1-a}}^{1} = \frac{1-a}{b}+\frac{1-a}{bx}(1-a-x)\ln\left(1-\frac{x}{1-a}\right) \\
&= -g(1-a,b,-x)
\end{aligned}
$$

and

$$
\tilde{h}(a,b,x) := -\frac{a}{bx}\int_{a-x}^{a}\ln\left(\frac{v}{a}\right)dv = -\frac{a^2}{bx}\int_{1-\frac{x}{a}}^{1}\ln(u)\,du = \tilde{g}(1-a,b,x) = -g(a,b,-x)
$$

Given p, we relate the $K^*=0$ conditional quantities to their unconditional analogues:

$$
F_{0|K^*=0}(k) = \frac{F_0(k)-p}{1-p}, \quad F_{1|K^*=0}(k) = \frac{F_1(k)-p}{1-p}, \quad \mathcal{B}^* = \frac{\mathcal{B}-p}{1-p}, \quad f_{d|K^*=0}(k) = \frac{f_d(k)}{1-p} \quad \forall d\in\{0,1\}
$$

Let $F(h) = P(h_{it}\leq h)$ be the CDF of the data, and define $f(h) = \frac{d}{dh}P(h_{it}\leq h)$ for $h\neq k$. By Proposition B.2 and the BLC assumption, the above quantities are related to observables as:

$$
F_0(k) = \lim_{h\uparrow k}F(h)+p, \qquad F_1(k) = F(k), \qquad f_0(k) = \lim_{h\uparrow k}f(h), \quad \text{and} \quad f_1(k) = \lim_{h\downarrow k}f(h)
$$

As shown by Dümbgen et al. (2017), BLC implies the existence of a continuous density function, which assures that the required density limits exist, and delivers Item 1. of the theorem.

To obtain the final result, note that the function $g(a,b,x)$ is homogeneous of degree zero. Thus $\Delta_k^*\in[\Delta_k^L,\Delta_k^U]$, with $\Delta_k^L := g\left(F_-(k), f_-(k), \mathcal{B}-p\right) + g\left(1-F(k), f_+(k), \mathcal{B}-p\right)$ and $\Delta_k^U := -g\left(1-p-F_-(k), f_-(k), p-\mathcal{B}\right) - g\left(F(k)-p, f_+(k), p-\mathcal{B}\right)$, where $-$ and $+$ subscripts denote left and right limits. A proof that the bounds Δ_k^L,Δ_k^U are sharp is presented in Appendix D.