Treatment Effects in Bunching Designs: The Impact of the Federal Overtime Rule on Hours

Leonard Goff*

This version: March 13, 2022
For current version click here

Abstract

The Fair Labor Standards Act mandates overtime premium pay for most U.S. workers, but limited variation in the rule has made assessing its impacts on the labor market difficult. With data from individual paychecks, I use the extent to which firms bunch workers at the overtime hours threshold to estimate the rule's effect on work hours. Generalizing previous methods, I show that bunching at a choice-set kink partially identifies an average causal response to the policy switch at the kink, under weak assumptions about preferences and allowing nonparametric heterogeneity. The bounds indicate a relatively small elasticity of demand for hours.

*Department of Economics, University of Georgia. 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, Bhargav Gopal, Jonas Hjort, Wojciech Kopczuk, Bentley MacLeod, Matthew Masten, Serena Ng, José Luis Montiel Olea, Dilip Ravindran, Ian Schmutte, Miguel Urquiola, and seminar participants at 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. Online Appendices available `here`.

1 Introduction

Many countries require premium pay for long work hours, in an effort to limit excessive work schedules and encourage hours to be spread over more workers. In the U.S., such regulation comes through the "time-and-a-half" rule of the Fair Labor Standards Act (FLSA): 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, 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).

Nevertheless, only a small literature has studied the effects of the FLSA overtime rule on the labor market. This stands in marked contrast to the large body of work on the minimum wage, which was also introduced at the federal level by the FLSA in 1938. A key reason for this gap is that the overtime rule has varied little since then: 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.[1]

This paper assesses the effect of the FLSA overtime rule on hours of work, taking a new approach that makes use of variation *within* the rule itself. The policy introduces a sharp discontinuity in the marginal cost of a worker-hour—a convex "kink" in firms' costs—which provides firms with an incentive to set workers' hours exactly at 40. Optimizing behavior by firms predicts that the resulting 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. Combining this observation with assumptions about the shape of the distribution of hours that would be chosen absent the FLSA, I use the bunching mass to identify the effect of the overtime rule on hours.

To do so, I develop a generalization of the "bunching design" identification strategy, which has used bunching at kinks in income tax liability to identify the elasticity of labor supply (Saez 2010; Chetty et al. 2011).[2] In particular, I give new identification results that hold under weaker assumptions and may be suitable to a variety of empirical contexts, showing that the bunching design can be useful for program-evaluation questions such as assessing the effect of the FLSA.

[1]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 30 and 31 for a comparison of my results to these papers. Quach (2021) looks at very recent reforms to eligibility criteria for exemption from the FLSA, estimating effects of the expansion on employment and the incomes of salaried workers, but not on hours of work.

[2]The same basic model has since been applied in a range of settings beyond income taxation. This paper considers only the bunching design for kinks, and not a related method for bunching at *notches* (e.g. Kleven and Waseem 2013).

In income tax settings, the promise of the bunching design is to overcome endogeneity in the marginal tax rates that apply to different individuals while requiring only the cross-sectional distribution of income near a threshold between tax brackets. Analogously, the starting point in the overtime setting is to construct the distribution of hours for which workers are paid in a single week. Administrative hours data at the weekly level has until recently been unavailable, and studies of overtime in the U.S. have typically relied on self-reported integer hours of work from surveys such as the Current Population Survey. I instead obtain detailed data via individual paycheck records from a large payroll processing company. Among workers paid weekly, these paychecks report the exact number of hours that the worker was paid for in a given week, allowing me to construct the distribution of hours-of-pay without rounding or other sources of measurement error.

With these novel data in hand, the goal is to translate features of the observed hours distribution into estimates of the overtime rule's causal effect, under credible assumptions about how weekly working hours are determined. This requires moving beyond the standard bunching-design model popularized in public-finance applications, in which agents have parametric "isoelastic" preferences and strong restrictions are placed on heterogeneity. In the overtime setting, I show that bunching is informative about firms as the decision-maker, choosing the hours of each of their workers in a given week. With this in mind, the identifying assumptions of the bunching design can be separated into two parts: i) assumptions about how individual agents (firms) would make choices given counterfactual choice sets—a *choice model*, and ii) assumptions about the distribution of heterogeneity in choices across observational units (paychecks).

As a first methodological contribution, I show that the class of choice models under which the bunching-design can be useful is considerably more general than the benchmark isoelastic model and its variants. In particular, I find that the method does not rest upon the researcher positing any explicit functional form for decision-makers' (firms') utility; rather, the main prediction about choice driving identification comes from *convexity* of preferences (e.g. weekly profits). In my formulation, agents in the bunching design can furthermore have multiple underlying margins of choice, which might be unobserved to the researcher and vary by observational unit.[3] These findings establish an important robustness property for the bunching design: it rests on a prediction about choice behavior that remains broadly valid even when the parametric utility model typically used to motivate the design is misspecified.

This generality is accomplished by recasting the bunching design in the language of potential outcomes, defining the parameter of interest in terms of a pair of counterfactual *choices* rather than as a preference parameter from an explicit choice model. In the overtime setting these potential outcomes correspond to: a) the number of hours the firm would choose for the worker this week

[3]This property of my choice model also generalizes Blomquist et al. (2021), who discuss a bunching-design setup with nonparametric utility but with a scalar choice variable.

if the worker's normal wage rate applied to all hours this week; and b) the number that the firm would choose if the worker's overtime rate applied to all hours this week. I show that choice from a kinked choice set can be fully characterized by the pair of counterfactuals: agents either choose one of them or they choose the location of the kink. Bunching at the kink then directly identifies a feature of the joint distribution of the potential outcomes, allowing one to make statements about treatment effects purged of selection bias.[4]

In addition to generalizing the choice model underlying the bunching design, I propose a new approach to weakening assumptions about heterogeneity required by the method. Blomquist et al. (2021) emphasize that identification from bunching rests on assumptions regarding the distribution of heterogeneity that cannot be directly verified in the data. In my formulation, such assumptions take the form of extrapolating the marginal distributions of each of the two potential outcomes, which are both observed in a censored manner. To perform this extrapolation I impose a natural nonparametric 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 in-part testable. The restriction affords partial identification of a conditional average treatment effect among units located at the kink, a parameter I call the "buncher ATE". In the overtime context, the buncher ATE represents an average reduced-form wage elasticity of hours demand, which I then use to assess the average effect of the FLSA.

I further 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) the result yields bounds on the derivative of bunching and mean hours with respect to policy parameters. I use this to evaluate proposed reforms to the FLSA: for example lowering the overtime threshold below 40 hours, or increasing the premium pay factor from 1.5 to 2.[5]

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 by making extrapolation assumptions separately for each of *two* counterfactuals. By contrast, existing approaches operate by constraining the distribution of a single scalar heterogeneity parameter, a simplification afforded by the isoelastic choice model. In the context of that model, Bertanha et al. (2020) and Blomquist et al. (2021) obtain bounds on the elasticity when the researcher is willing to put an explicit limit on how quickly the density of heterogeneous choices can change. My approach based on bi-log-concavity avoids the need to choose any such tuning parameters, and is applicable in the general choice model.

The empirical setting of overtime pay involves confronting two specific challenges not typical of

[4]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 the joint distribution, assuming optimizing behavior.

[5]For example, the bill HR4728 introduced in 2021 would establish a 32 hour rule for overtime pay.

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 always be set unilaterally by one party: in principle either the firm or the worker could have control over a given worker's schedule. I provide evidence that week-to-week variation in hours tends to be driven by firms, and show that even when bargaining weight between workers and firms is arbitrary and heterogeneous bunching at 40 hours is informative about labor demand rather than supply.

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 estimate that the local wage elasticity of weekly hours demand close to 40 falls in the range -0.04 to -0.19, indicating that firms are fairly resistant to changing hours to avoid overtime payments. While the bunching design is only directly informative about the *hours* effects of the overtime rule, a back-of-the-envelope calculation suggests that FLSA regulation creates about 700,000 jobs (relative to an estimated 100 million non-exempt workers), despite a reduction in total hours.

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

2 Conceptual framework

This section outlines a framework for reasoning about the determination of weekly hours, which motivates the identification strategy of Section 4. Readers primarily interested in the bunching design may wish to skip directly to that section.

The conceptual framework is centered around two observations from the data (described in Section 3): weekly hours vary considerably between pay periods for an individual hourly worker, and a given worker's hourly wage tends to change infrequently. I propose to view this as a two stage-process. First, workers are hired with an hourly wage set along with an "anticipated" number of hours they will work per week. 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. Given the FLSA overtime rule and a worker's fixed wage, their employer thus faces a kinked cost schedule when choosing hours in a given week, as pictured in Figure 1.

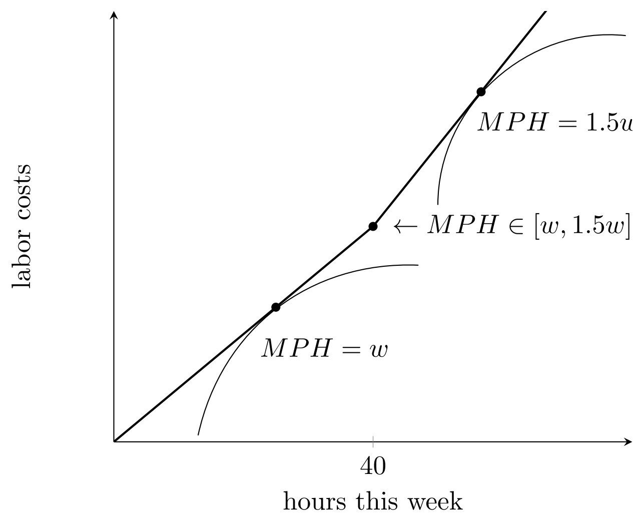

FIGURE 1: With a given worker's straight-time wage fixed at w, labor costs as a function of hours have a convex kink at 40 hours, given the overtime rule. Simple models of week-by-week hours choice (see Section 4.2) yield bunching when for a mass of workers the marginal product of an hour at 40 is between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

Wages and anticipated hours set at hiring

We begin with the hiring stage, which pins down a worker's wage. Throughout the analysis, I focus on workers paid on an hourly basis. Following the literature I refer to the hourly rate of pay that applies to the first 40 of a worker's hours as their *straight-time wage* or simply *straight wage*. This section discusses a benchmark model to endogenize such straight wages, which proves useful in thinking about how these wages may themselves be affected by the overtime rule. The basic bunching design strategy of Section 4 only requires 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. The model described in this section is spelled out formally in Appendix F.1.

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 they are each expected to work per week. The firm faces a labor supply function $N(z, h)$ determined by workers' preferences over the labor-leisure tradeoff,[6] 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 (I will use the term "covered" to indicate workers whose employer is covered by the FLSA and who are non-exempt from the overtime rule).

While labor supply is 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 between the two main views on the likely effects of overtime policy by first supposing that a workers' 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 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.[7] Suppose for the moment that all workers then in fact work exactly h^* hours in all weeks (abstracting away from any reasons for the firm to 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 hours. 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 $N(z, h)$: one that exhibits perfect competition on the quantity $w_s(z, h)$. I show in Appendix F.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.[8]

Trejo (1991) and Barkume (2010) investigate whether the fixed-job or fixed-wage model better

[6] This labor supply function can be viewed as an equilibrium object that reflects both worker preferences and the competitive environment for labor. In Appendix F.2, I endogenize this function in a simple extension of the imperfectly competitive Burdett and Mortensen (1998) search model.

[7] In Appendix F.1 I give a closed-form expression for (z^*, h^*) when both labor supply and production are iso-elastic: 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.

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

accords with the observable joint distribution of hourly wages and hours. They find that wages do tend to be lower among jobs with overtime pay provisions and more overtime hours, but by a magnitude smaller than would be predicted by the fixed-jobs model. However, these estimates could be driven by selection, e.g. of lower-skilled workers into covered jobs with longer hours.

In Appendix C.2, I construct an empirical test of Equation (1) that is instead based on assuming that the conditional distribution (across individual paychecks) of z^* is smooth across $h* = 40$. I find that roughly one quarter of paychecks around 40 hours reflect the wage/hours relationship predicted by the fixed-job model. This finding, along with the observation that hours change much more frequently than wages, is consistent with a model in which once straight-wages are set according to Equation (1), they remain fairly static over time.[9] 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.

Dynamic adjustment to hours by week

Confronted with the observation (presented in Table 1) that workers' hours vary considerably week-to-week in my sample of hourly workers, I assume that this week-level variation reflects choices made by their employers. There are many reasons to expect variation in firms' demand for hours over time. Shocks to product demand or productivity change the number of weekly hours that would be optimal that week from the firm's perspective. 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, variation in productivity across workers may only become apparent to supervisors after workers' straight wages have been set, and it may be profitable to increase the hours of the most productive ones.

Throughout Section 4, I maintain a strong version of the assumption that a firm—rather than a worker—chooses the hours that I observe on each paycheck. In this model workers' preferences *can* matter in the determination of each worker's straight wage at hiring, but I assume the firm has final scheduling rights week to week.[10] This simplification eases notation and emphasizes the intuition behind my identification strategy. Appendix D 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. If just some workers have full control of their weekly hours, then the bunching-design strategy will only be informative about effects of the FLSA among workers

[9] 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).

[10] This can be rationalized on the basis of workers generally having less bargaining power: 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 outside option is having one less worker (Stole and Zwiebel, 1996).

whose final hours are chosen by the firm.

Available survey evidence suggests that this group is the dominant one: a relatively small share of workers report that they choose their own schedules (despite the ongoing increase in flexible work arrangements). 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 control over which shifts they work (Matos et al., 2017).

3 Data and descriptive patterns

The main dataset I use comes from a large payroll processing company. They provided anonymized paychecks for the employees of 10,000 randomly sampled employers, for all pay periods in the years 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 (straight-time) pay, overtime pay, sick leave, 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 their salary if one is stored in the system.

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 sample only of workers who are non-exempt from the FLSA overtime rule. For the purposes of a), I keep paychecks from workers who are paid on a weekly basis (roughly half of the workers in the sample). To achieve b) I focus on hourly workers, since nearly all workers who are paid hourly are subject to FLSA regulation. I also drop any workers who never receive overtime pay during the study period. The final sample includes 630,217 paychecks for 12,488 workers across 566 firms. See Appendix C.1 for further details of the sample construction.

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) 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 administrative data when possible.[11]

	(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	12488	149459	63404	228773

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.

The sample I use is somewhat 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. In particular, conditioning on workers that are paid on a weekly basis oversamples industries that tend to have more men, and tend to pay somewhat lower wages. Appendix C 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 10 shows that overtime pay is present in nearly all weekly paychecks that report more than 40 hours, in line with the presumption that workers in the final sample are not FLSA-exempt.

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. When 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, holiday, and sick pay hours as discussed in Section 5.

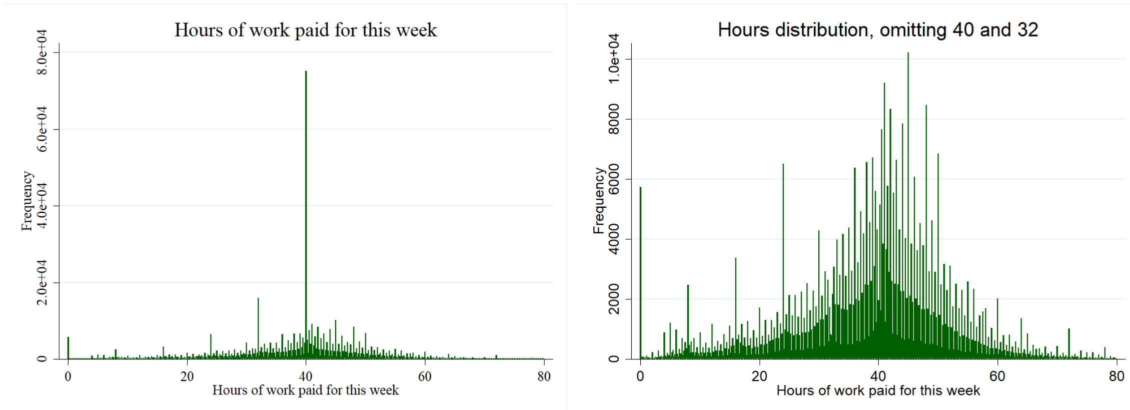

FIGURE 2: Empirical densities of hours worked pooling all paychecks in final estimation sample. Sample is restricted to hourly workers receiving overtime pay at some point (to ensure nearly all are non-exempt from FLSA, see text), and workers having hours variation. The right panel omits the points 40 and 32 to improve visibility elsewhere. Bins have a width of 1/8 of an hour, based on the observed granularity of hours (see Appendix Figure 12 for details).

change the magnitude is about 7 hours on average, and roughly symmetric around zero (see Appendix Figure 12 for the distribution of hours changes).

	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.

Appendix C reports some further details on the variation in hours and wages present in the data. Appendix Table 2 regresses hours, overtime hours, and an indicator for bunching on worker observables, and shows that after controlling for worker and date fixed effects bunching and overtime hours are both predicted by recent hiring at the firm, lending further evidence for the assumption that shocks to labor demand drive variation in hours. Appendix Table 3 shows that overall, about 63% of variation in total hours can be explained by worker and employer-by-date fixed effects. Appendix Table 1 documents heterogeneity in the prevalence of overtime pay across industry classifications. Appendix Figure 5 studies the joint distribution of wages and hours and reproduces Bick et al.'s (2022) finding that wages increase with hours until just beyond 40, then decline some.

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 that worker's wage fixed and costs a kinked function of hours as depicted in Figure 1. This section shows that under weak assumptions, firms facing such kinks 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 a much 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 A. Throughout the present 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. Probability statements are to be understood with respect to such paycheck-level units.

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 this week w_{it}. 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

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

where $B_{0it}(h) = w_{it}h$ and $B_{1it}(h) = 1.5w_{it}h - 20w_{it}$. The kinked pay schedule $B_{kit}(h)$ is equal to $B_{0it}(h)$ for values $h \leq 40$ and $B_{kit}(h)$ 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 A:

Definition (potential outcomes). *Let h_{0it} denote the hours of work that of unit it would be paid for if instead of $B_{kit}(h)$, 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 $B_{1it}(h)$.*

By contrast, 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 $B_{kit}(h)$ 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.

Our second assumption is that each unit's firm optimizes some vector $\mathbf{x}$ 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 $\mathbf{x}$ (Appendix A.3 gives some examples). Firm preferences are taken to be convex in $\mathbf{x}$ 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 $\mathbf{x}$ for each unit.*

For the sake of brevity, I here state a version of CONVEX that is a bit stronger than necessary for the identification results below. In particular, Appendix A 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 under no assumptions about convexity of preferences. The assumption that firms rather than workers choose hours enters in the claim that π is decreasing (rather than increasing) in z, but Appendix D 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 1 shows that Assumptions CHOICE AND CONVEX imply that:

$$h_{it} = \begin{cases} h_{0it} & \text{if} \quad h_{0it} < 40 \\ 40 & \text{if} \quad h_{1it} \leq 40 \leq h_{0it} \\ h_{1it} & \text{if} \quad h_{1it} > 40 \end{cases} \tag{2}$$

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 located at the corner solution of 40 if and only if the two counterfactual outcomes fall on either side, "straddling" the kink.[13] Figure 3 depicts the implications of Eq. (2) for what is observable by the researcher in the bunching design: censored distributions of both h_0 and of h_1, and a point-mass of size $\mathcal{B} = P(h_{1it} \leq 40 \leq h_{0it})$ at the kink.

[13] "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 A).

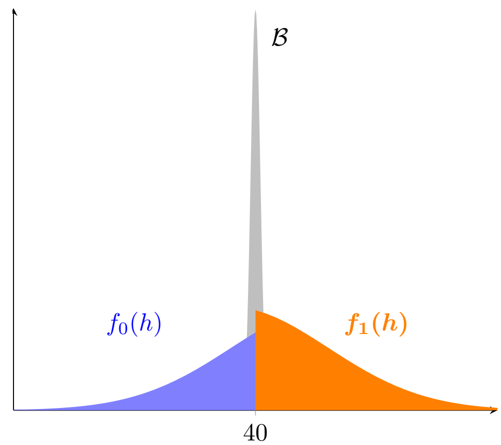

FIGURE 3: Observables in the bunching design, given Equation (2). To the left of the kink at 40, the researcher observes the density $f_0(h)$ of the counterfactual h_{0it}, up to values $h = 40$. To the right of the kink, the researcher observes the density $f_1(h)$ of h_{1it} for values $h > 40$. At 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 departure from previous approaches to the bunching design, which characterize bunching in terms of the counterfactual h_0 only.[14] I show below that this 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 A 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" decision-makers dislike.

Intuition for Equation (2) in the overtime setting

As an intuitive illustration of Equation (2), suppose that firms balance the cost $B_{kit}(h)$ against the value of h hours of the worker's labor, in order to maximize that week's profits. Then Eq. (2) can be written:

$$h_{it} = \begin{cases} MPH_{it}^{-1}(w_{it}) & \text{if} \quad MPH_{it}(40) < w_{it} \\ 40 & \text{if} \quad MPH_{it}(40) \in [w_{it}, 1.5w_{it}] \\ MPH_{it}^{-1}(1.5w_{it}) & \text{if} \quad MPH_{it}(40) > 1.5w_{it} \end{cases} \tag{3}$$

where denotes $MPH_{it}(h)$ is the marginal product of an hour of labor for unit it, as a function of hours h. Assuming that production is strictly concave, the function $MPH_{it}(h)$ will be strictly decreasing in h, and we have that $h_{0it} = MPH_{it}^{-1}(w_{it})$ and $h_{1it} = MPH_{it}^{-1}(1.5w_{it})$.

[14] 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 2.

Figure 1 depicts Eq. (3) 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.

While Eq. (3) provides a natural nonparametric characterization of when the firm will ask a worker to work overtime (when the ratio of producitivity to wages is high), it is still more restrictive than necessary for the purposes of the bunching design. Appendix A.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.

Note that if production depends jointly on the hours of all workers within a firm, we may expect the function $MPH_{it}(h)$ in Eq. (3) to depend on the hours of worker i's colleagues in week t. In this case the quantities h_{0it} and h_{1it} hold the hours of i's colleagues fixed at their *realized* values: they contemplate ceteris paribus counterfactuals in which the pay schedule for a single unit it is varied, and nothing else. This affects the interpretation of our treatment effects, as discussed in Section 4.4. In the baseline isoelastic model that we consider in the next section, such interdependencies between workers' hours are ruled out by assuming that production is linearly separable across units. Appendix E discusses in more detail the case of a general nonseparable production function.

4.2 The benchmark isoelastic model

The canonical approach from the bunching-design literature (Saez, 2010; Chetty et al., 2011; Kleven, 2016), strengthens Assumption CONVEX to suppose that $\mathbf{x} = h$ and decision-makers' utility follows an isoelastic functional form, with preferences identical between units up to a scalar heterogeneity parameter. This corresponds to a model in which firm profits from unit it are:

$$\pi_{it}(z, h) = a_{it} \cdot \frac{h^{1+\frac{1}{\epsilon}}}{1 + \frac{1}{\epsilon}} - z \tag{4}$$

where $\epsilon < 0$ is common across units, and z represents wage costs for worker i in week t. Eq. (4) 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. Let $\eta_{it} = a_{it}/w_{it}$ denote the ratio of a unit's current productivity factor a_{it} to their straight wage. In the isoelastic model

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

and by Eq. (3) actual hours h_{it} are ranked across units in order of η_{it}. 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. In the isoelastic model whether a worker works overtime in a given week is determined by the scalar η_{it}: a worker with a wage w_{it} fixed throughout the year may for example work overtime only in periods when a_{it} is relatively high due to seasonally elevated productivity.

Identification in the isoelastic model

In the context of the isoelastic model, a natural starting place for evaluating the FLSA is to estimate the parameter ϵ. Ignoring for the moment any effects of the policy on straight-wages, the effect of the time-and-a-half rule on unit it's hours will simply be the difference $h_{it} - h_{0it}$, what we might call the *effect of the kink*. It follows from the above that the effect of the kink is $h_{it} \cdot (1 - 1.5^{-\epsilon})$ for any unit such that $h_{it} > 40$. Provided the value of ϵ, we could thus evaluate the effect of the time-and-a-half rule for any paycheck recording overtime using that unit's observed hours.

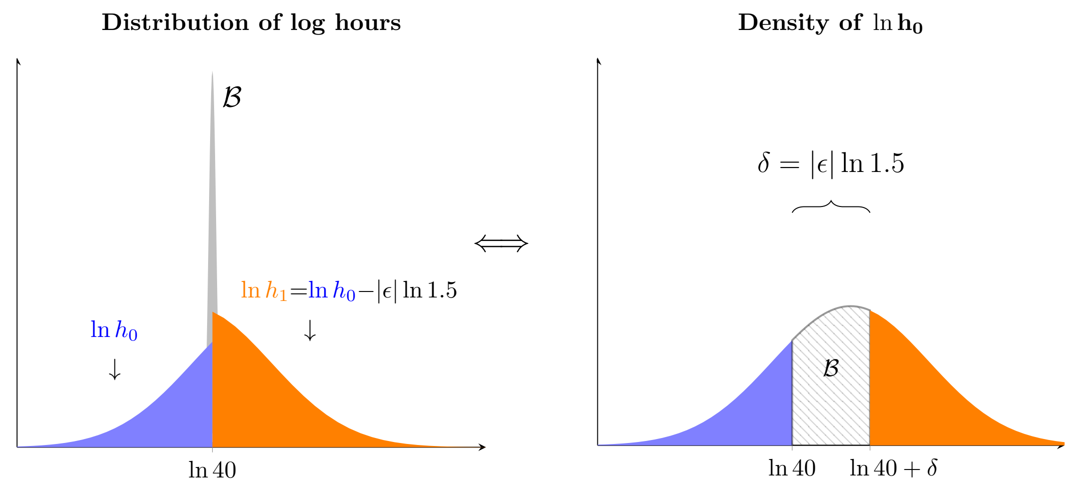

FIGURE 4: The left panel depicts the distribution of observed log hours $\ln h_{it}$ in the isoelastic model, while the right panel depicts the underlying full density of $\ln h_{0it}$. Specializing from the general setting of Figure 3, we have in the isoelastic model that $f_1(h) = f_0(h + |\epsilon| \cdot \ln 1.5)$. Thus, the full density of f_0 is related to the observed distribution by "sliding" the observed distribution for $h > 40$ out 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 $\mathcal{B}$.

The classic bunching-design method pioneered by Saez (2010) identifies ϵ by relating it to the observable bunching probability:

$$\mathcal{B} := P(h_{it} = 40) = \int_{40}^{1.5^{|\epsilon|}\cdot 40} f_0(h) \cdot dh \tag{5}$$

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 solving Eq. (5) for $|\epsilon|$. However, f_0 is not globally identified from the data: by 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.[15] 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 particularly suitable in the overtime context. 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. The method of Chetty et al. (2011) ignores the "shift" by δ in the right panel of Figure 4. Both of these approaches rely on parametric assumptions, and sufficient conditions for each are outlined in Appendix G.2.

If on the other hand, 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 $\epsilon < 0$, a point emphasized by Blomquist et al. (2021) and Bertanha et al. (2020). In particular, given the integration constraint (5), 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. I find a middle ground by imposing a nonparametric shape constraint on h_0: *bi-log-concavity* (BLC), leading to a partial identification result. A detailed discussion of BLC is given in Section 4.3.

Limitations of the isoelastic model

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

[15] Bertanha et al. (2020) note that given a full parametric distribution for f_0, the entire model could be estimated by maximum likelihood. This approach would enforce (5) automatically while enjoying the efficiency properties of MLE.

C.5, assuming the isoelastic model and that h_0 and h_1 are BLC (which nests a linear density as a special case) suggests that $\epsilon \in [-.179, -.168]$.[16] Such values are implausible when interpreted through the lens of Equation (4): $\epsilon = -.2$ for exmaple would imply an hours production function of $f(h) = -\frac{1}{4}h^{-4}$ (up to an affine transformation), which features an unrealistic degree of concavity. Allowing a more general production function $f(h)$ (separable between units) is also not much help, as the standard bunching design approach then estimates an averaged local inverse elasticity of $f(h)$ (see Appendix C.5).

In short, the observed bunching at 40 hours is too small to be reconciled with a model in which ϵ parameterizes the concavity of weekly production with respect to hours. The estimand of the bunching design should instead be interpreted as a *reduced-form* elasticity of the demand for hours, which may reflect adjustment by firms along additional margins that can attenuate the hours response, and thus reduce the magnitude of bunching. Appendix A.3 provides some examples.

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 parameter of interest 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. Δ_{it} thus represents the causal effect of a one-period 50% increase in worker i's wage on their hours in week t: the difference between the hours that unit's firm would choose if the worker were paid at their straight-time rate versus at their overtime rate for all hours in that week (asuming quasi-linearity of firm preferences). As such we would expect that $\Delta_{it} \geq 0$. A unit's treatment effect can be contrasted with the "effect of the kink" quantity $h_{it} - h_{0it}$ introduced before, but importantly the two are related: by Eq. (2) the effect of the kink is $-\Delta_{it}$ for all units working overtime.

In the isoelastic model $\Delta_{it} = h_{0it} \cdot (1 - 1.5^{\epsilon})$, representing a special case in which treatment effects are homogenous 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

[16]The width of these bounds is about 4 times smaller than if BLC is assumed for h_0 only. These estimates 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.

see this, let $k = 40$ denote the location of the kink, and write the bunching probability as:

$$\mathcal{B} = P(h_{1it} \leq k \leq h_{0it}) = P(h_{0it} \in [k, k + \Delta_{it}]) = P(h_{1it} \in [k - \Delta_{it}, k]), \tag{6}$$

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: $\mathcal{B} = F_1(k) - F_0(k)$, provided that each potential outcome is continuously distributed and with F_0 and F_1 their cumulative distribution functions.

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 care is needed in defining this parameter, allowing 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, and would not locate at 40 hours otherwise. 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 a strategy estimate it empirically in Section 5.1.

To simplify the discussion, suppose for the moment that there are no counterfactual bunchers, so that $\Delta_k^* = \mathbb{E}[\Delta_{it} | h_{it} = k]$. Our goal is to invert (6) 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 isoelastic model, in which there is only a scalar dimension 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 (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 $F_0(k)$ and $F_1(k) = F_0(k) + \mathcal{B}$, where Q_d is the quantile function of h_{dit}, i.e.:

$$\Delta_k^* = \frac{1}{\mathcal{B}} \int_{F_0(k)}^{F_1(k)} [Q_0(u) - Q_1(u)]du, \tag{7}$$

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. (7) to hold:

Assumption RANK. *There exist fixed values* Δ_0^* *and* Δ_1^* *such that* $h_{0it} \in [k, k + \Delta_{it}]$ *iff* $h_{0it} \in [k, k + \Delta_0^*]$, *and* $h_{1it} \in [k - \Delta_{it}, k]$ *iff* $h_{1it} \in [k - \Delta_1^*, k]$.

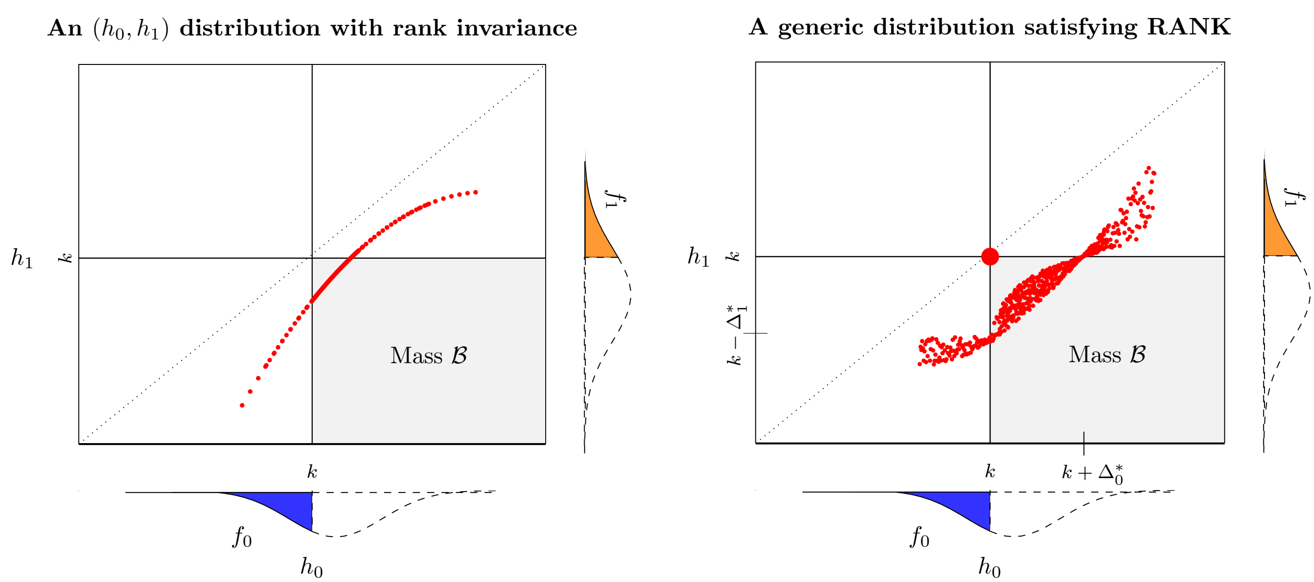

FIGURE 5: The joint distribution of (h_{0it}, h_{1it}) (in red), comparing an example satisfying rank invariance (left) to a case satisfying Assumption RANK (right). RANK allows the support of the joint distribution to "fan-out" from perfect co-dependence of h_0 and h_1, except when either outcome is 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.

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.[17] For example,

[17] 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 Equation (2) hold for the unit. Specifically, for both $d = 0$ and $d = 1$: $U_d|(h < k) \sim Unif[0, F_0(k)]$, $U_d|(h = k) \sim Unif[F_0(k), F_1(k)]$, and $U_d|(h > k) \sim Unif[F_1(k), 1]$.

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 $h_1 = k$. When this is not perfectly satisfied, Appendix A.5 demonstrates how the RHS of Equation (7) will then yield a lower bound on the true buncher ATE (and can still be interpreted as an averaged quantile treatment effect). Appendix Figure 15 generalizes RANK to case in which some workers choose their hours, resulting in mass also appearing in the north-west quadrant of Figure 5.

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

Given Eq. (7), 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:

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 emphasized by Blomquist et al., 2021 and discussed in Section 4.2.[18]

The family of BLC distributions includes parametric distributions assumed by previous bunching design studies, such as those with uniform or linear densities (Saez, 2010), or those with polynomial densities as in Chetty et al. 2011 (provided they have real roots). All globally log-concave distributions are BLC.[19] Importantly, the BLC property is partially testable in the bunching design, since $F_0(y)$ is identified for all $h < k$ and $F_1(h)$ is identified for all $h > k$. Appendix Figure 9 shows that the observable portions of F_0 and F_1 indeed satisfy BLC. Identification requires us to believe that BLC *also* holds in the unobserved portions of F_0 and F_1.

We are now ready to state the main identification result. Its logic is summarized by Figure 6: given the general choice model, RANK converts identification of the buncher ATE into a pair of

[18] Bertanha et al. (2020) 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.

[19] However unlike log-concave densities, BLC distributions need not be unimodal (Dümbgen et al., 2017).

extrapolation problems, each of which are approached by assuming bi-log-concavity of the corresponding marginal potential outcome distribution. Let $F(h) := P(h_{it} \le h)$ be the CDF of observed hours.

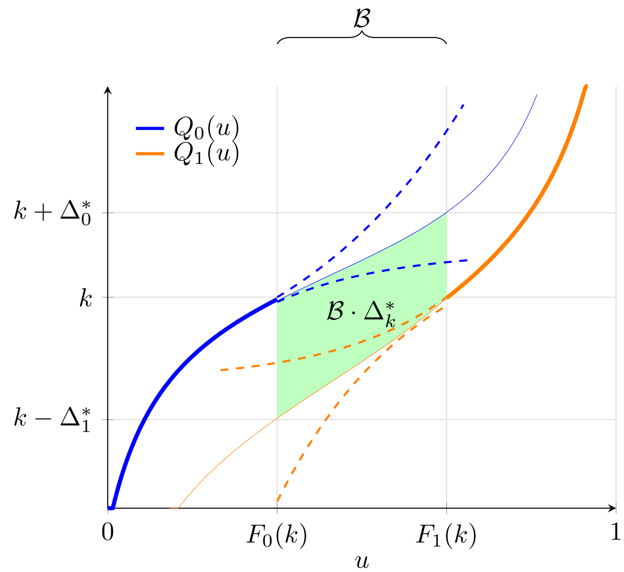

FIGURE 6: Extrapolating the quantile functions for h_0 and h_1 (blue and orange, respectively) to place bounds on the buncher ATE (case depicted has no counterfactual bunchers). The observed portions of each quantile function are depicted by thick curves, while the unobserved portions are indicated by thinner curves. The dashed curves represent upper and lower bounds for this unobserved portion coming from an assumption such as bi-log-concavity. The buncher ATE is equal to the area shaded in green, divided by the bunching probability $\mathcal{B}$.[20] The quantities Δ_0^* and Δ_1^* are defined in Assumption RANK below.

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:*

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

and

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

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

Proof. See Appendix B. □

Combining Items 1 and 2 of Theorem 1, it follows that the sharp bounds Δ_k^L and Δ_k^U on the buncher ATE are identified, given the CDF of the data $F(h)$ and p.[21] Inspection of the expressions appearing in Theorem 1 reveals that the bounds become wider the larger the net bunching probability $\mathcal{B} - p$. When $f_0(k) \approx f_1(k)$ and $p = 0$, 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 helps explain why the estimated bounds in Section 5 turn out to be quite informative.

Comparison to existing results. The existing bunching design literature does contain a few identification results that circle the common intuition that bunching is informative about a local average responsiveness, when responsiveness to incentives varies by observational 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 8 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 the small-kink 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 $\Delta_k^* \approx \frac{\mathcal{B}-p}{2f_0(k)} + \frac{\mathcal{B}-p}{2f_1(k)}$.

A second existing result comes from Blomquist et al. (2015), who show that bunching identifies a certain weighted average of compensated elasticities in a nonparametric labor supply model, if the density of choices at an income tax kink is assumed to be linear across counterfactual tax rates. But as these authors point out, such a parametric assumption would be difficult to motivate.[22]

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

[20] 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 tradeoff.

[21] 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.

[22] In particular, the data identifies the density at the kink for two particular tax rates only, so cannot provide evidence of such linearity. Theorem 1 instead requires assumptions only about the two counterfactuals that are in fact observed.

at their overtime rate? This section discusses how we may now 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 some additional assumptions, which I continue to approach from a partial identification perspective.

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 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 average weekly hours among hourly workers with and without the FLSA: $\theta := \mathbb{E}[h_{it}] - \mathbb{E}^*[h_{it}^*]$, where h_{it}^* indicates the hours unit it would work absent the FLSA, and the second expectation $\mathbb{E}^*$ is over units corresponding to workers that would exist in the no-FLSA counterfactual and be covered were it introduced.[23] 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 anyways, 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:

$$\theta = \mathbb{E}[h_{it}(w_{it}, \mathbf{h}_{-i,t}) - h_{0it}(w_{it}^*, \mathbf{h}_{-i,t}^*)] = \mathbb{E}[\underbrace{h_{it}(w_{it}, \mathbf{h}_{-i,t}) - h_{0it}(w_{it}, \mathbf{h}_{-i,t})}_{\text{“effect of the kink”}}]$$
$$+ \mathbb{E}[\underbrace{h_{0it}(w_{it}, \mathbf{h}_{-i,t}) - h_{0it}(w_{it}^*, \mathbf{h}_{-i,t})}_{\text{“wage effects”}}] + \mathbb{E}[\underbrace{h_{0it}(w_{it}^*, \mathbf{h}_{-i,t}) - h_{0it}(w_{it}^*, \mathbf{h}_{-i,t}^*)}_{\text{“interdependencies”}}], \quad (8)$$

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 (8) 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

[23] Note that h_{it}^* in this section differs from the "anticipated" hours quantity h^* in Section 2.

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 Appendix E 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.[24]

Turning first to the "effect of the kink" term, note that with straight-wages and the hours of other units fixed, the kink only has such direct effects on those units working at least $k = 40$ hours:

$$h_{it} - h_{0it} = \begin{cases} 0 & \text{if } \; h_{it} < k \\ k - h_{0it} & \text{if } \; h_{it} = k \\ -\Delta_{it} & \text{if } \; h_{it} > k \end{cases} \tag{9}$$

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 Δ_{it} of units working more than 40 hours are at least as large on average as those who work 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, to put bounds on the average effect of the kink among bunchers $\mathcal{B} \cdot \mathbb{E}[k - h_{0it}|h_{it} = k]$, I use bi-log-concavity of h_0. Details are provided in Appendix H.9.

The "wage effects" term in Equation (8) 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 on the basis of expected hours. However, both terms ultimately depend on the same thing: responsiveness of hours to the cost of an hour of work. I 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 iso-elastic in wages, with anticipated hours approximated by h_{it}. Appendix H.9 provides a visual depiction of these definitions. 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 11).

[24] 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 E. If the third term of Eq. (8) 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.

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,[25] 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.

Let us begin by considering 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. I assume that the mass of counterfactual bunchers p remains at $k = 40$ after the shift.[26]

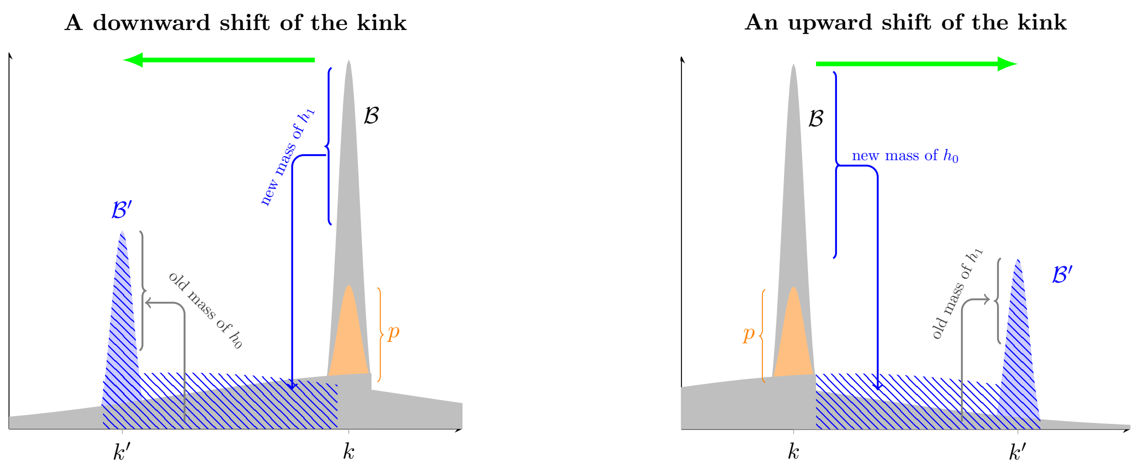

FIGURE 7: The left panel depicts a shift of the kink point downwards from k to k', while right panel depicts a shift of the kink point upwards. See text for details.

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. Let $h_{it}(\rho)$ be the hours that it would work if their employer faced a linear pay schedule at rate $\rho \cdot w_{it}$ (with w_{it} and hours of other units fixed at their realized levels). In this notation, $h_{0it} = h_{it}(1)$ and $h_{0it} = h_{it}(1.5)$. Now consider a new overtime policy in which a premium

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

[26]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.

pay factor of ρ_1 is required 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 $\mathcal{B}^{[k,\rho_1]} := P(h_{it}^{[k,\rho_1]} = k)$ the observable bunching that would occur.

Theorem 2 allows one to discuss 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.

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 A. □

The final two assumptions above are given in Appendix A: 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 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 A for details.

Beginning from the actual FLSA policy of $k = 40 = k^*, \rho_1 = 1.5$, the RHS of Items 1 and 2 are 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.[27] Making use of Theorem 2 for a discrete policy change like reducing standard hours

[27]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.

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. (8), 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 expression for the average effect of this reform in terms of local average elasticities of response:

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

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. I argue that 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, thus yielding 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 H.9.

5 Implementation and Results

This section implements the empirical strategy described in the last section 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 a kind of *status-quo* choice, being chosen even when it is not exactly profit maximizing for the firm. 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 successfully 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 A.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 workers' 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 uses 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. 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 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.[28]

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 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.

[28]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).

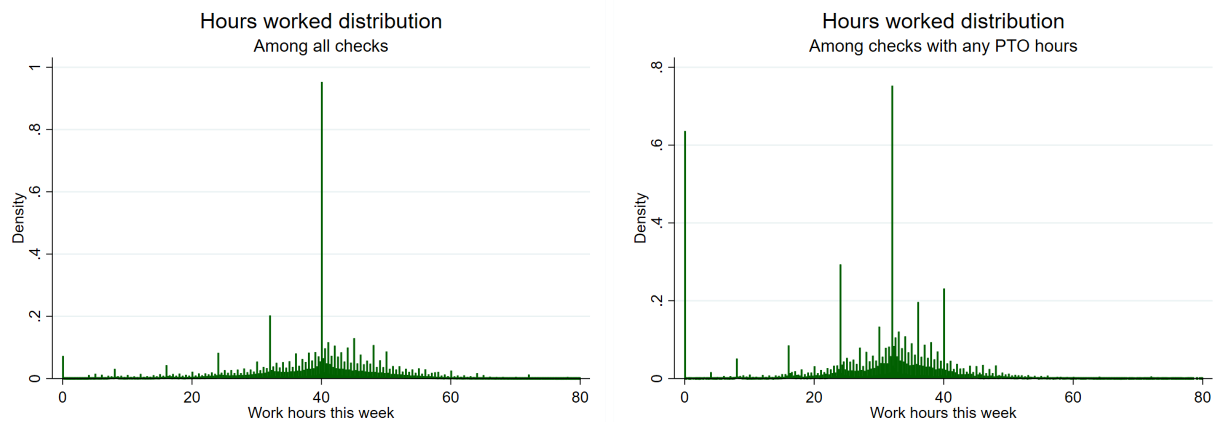

FIGURE 8: The right panel shows a histogram of hours worked when paid time off hours are positive ($\eta_{it} > 0$). The left panel shows the unconditional distribution. While $\mathcal{B} \approx 11.6\%$, $P(h_{it} = 40 | n_{it} > 0) \approx 2.7\%$.

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-movers 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 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. For instance, a local-linear CJM estimator provides a smoothed estimate of the left limit of the CDF and density at k as:

$$(\hat{F}_-(k), \hat{f}_-(k)) = \underset{(b_1, b_2)}{\operatorname{argmin}} \sum_{it: h_{it} < k} (F_n(h_{it}) - b_1 - b_2 h_{it})^2 \cdot K\left(\frac{h_{it} - k}{\alpha}\right) \tag{10}$$

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 a bandwidth choice using the data on either side of $k = 40$ (for the left and right limits, respectively). I then average the two bandwidths, and use this as the bandwidth in the final calculation of both the right and left limits. In the full sample, the bandwidth chosen by this procedure is about 1.7 hours, and is somewhat larger for subsamples 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 3 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 always close to each other and precisely estimated. Appendix C reports estimates based on alternative shape constraints and assumptions about effect heterogeneity, which deliver similar results.

	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 3: 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. This group would 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 this buncher ATE parameter is estimated to be at least 2.6 hours. In Appendix C I report estimates of the buncher ATE for each of the largest industries in the sample, and also plot estimates directly as a function of the assumed mass p of counterfactual bunchers at 40 hours.

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 4 first reports an estimate of the buncher ATE expressed as a reduced-form hours demand elasticity,[29] 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. I provide details of the calculations in Appendix H.9.

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 in any given week than they would absent overtime regulation: about one third the magnitude of the buncher ATE in levels. 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. However, we might assume that similar effects hold for hourly workers paid at other intervals (e.g. bi-weekly) and salary workers covered by the time-and-a-half rule have similar.

I estimate that a move to double-time pay would introduce a further reduction in hours comparable to the existing overall ex-post effect, but with substantially wider bounds. These estimates include the effects of possible adjustments to straight-time wages, which tend to attenuate the ef-

[29] This is $\hat{\Delta}_k^*/(40\ln(1.5))$ where $\hat{\Delta}_k$ is the estimate of the buncher ATE presented in Table 3, which turns out to be 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.

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

TABLE 4: Estimates of the buncher ATE expressed as an elasticity, the average ex-post effect of the FLSA $\mathbb{E}[h_{it} - h^*_{it}]$,[29] 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.

fects of the policy change. Appendix Table 11 replicates Table 4 neglecting these wage adjustments, which might be viewed as a short-run response to the FLSA before wages have time to 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. Appendix Figure 8 compares the hours distribution for Real Estate & Rental and Leasing with the distribution for of Professional Scientific and Technical, showing that the difference in their effects is explained both by a larger value of $\mathcal{B} - p$ and a lower density of hours close to the kink for Real Estate & Rental and Leasing. Appendix C reports estimates broken down by gender, finding that the FLSA has considerably higher effects on the hours of men compared with women.

Appendix Figure 14 looks at the effect of changing the threshold for overtime hours k from 40 to alternative values k'. The left panel reports estimates of the identified bounds on $\mathcal{B}^{[k',\rho_1]}$ as well as point-wise 95% confidence intervals (gray) across values of k' between 35 and 45, for each of the three approaches to estimating p. In all cases, the upper bound on bunching approaches zero as k' is moved farther from 40. This is sensible if the h_0 and h_1 distributions are roughly unimodal with modes around 40: straddling of potential outcomes becomes less and less likely as one moves away from where most of the mass is. Appendix Figure 13 shows these bounds as k' ranges all the way from 0 to 80, for the $p = 0$ case. These estimates do not account for adjustment to straight-time wages, so should be viewed as quantifying short-run responses.

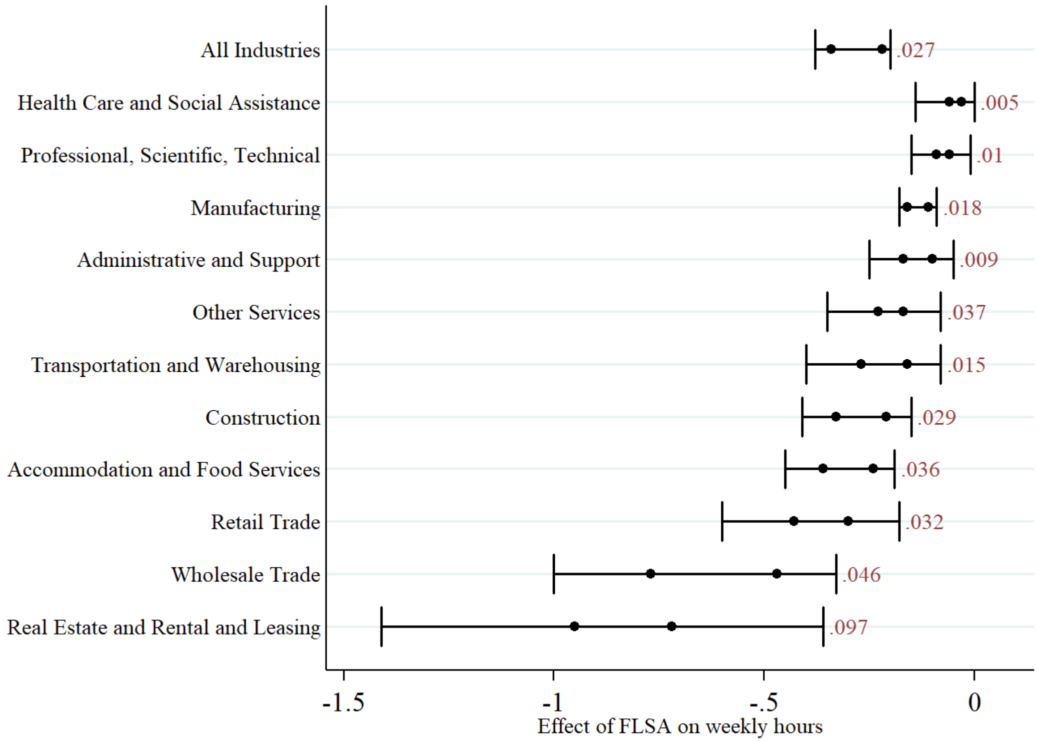

FIGURE 9: 95% confidence intervals for the effect of the FLSA on hours by industry, using PTO-based estimates of p for each. Dots are point estimates of the upper and lower bounds. The number to the right of each range is the point estimate of the net bunching $\mathcal{B} - p$ for that industry.

When p is estimated using PTO or non-changers between periods, we see that the upper bound of the identified set for $\mathcal{B}^{[k',\rho_1]}$ in fact reaches zero quite quickly in 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). The right panel of Appendix Figure 14 shows estimates for the average effect on hours of changing standard hours, inclusive of wage effects (see Appendix H.9 for details). Increases to standard hours cause an increase in hours per worker, as overtime policy becomes less stringent, and reductions to standard hours reduce hours.[30] The size of these effects is not precisely estimated for changes larger than a couple of hours, however the range of statistically significant effects depends on p. 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.

[30] The magnitudes are consistent with estimates by Costa (2000), that hours fell by 0.2-0.4 on average during the phased introduction of the FLSA in which standard hours declined by 2 hours in 1939 and 1940.

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 the salary threshold for exemption of white-collar jobs in real terms over the last several decades.[31] By contrast, my estimates carry the strengths of an identification strategy that does not require focusing on the sub-population affected by a natural experiment, and use much more recent 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—and related policies that have received recent interest like work-sharing programs—play 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 could be substantial. The 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 a full 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), 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 (see Appendix C.6 for details). To get an overall upper bound on the size of employment effects, I 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 3% among covered workers. A reasonable range of parameter values in this simple

[31] 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.

calculation rules out negative overall employment effects from the FLSA.

7 Conclusion

This paper has provided a reinterpretation 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 variety of specific underlying choice models. Across such modeling choices, the parameter of interest remains a reduced-form local average treatment effect between two appropriately-defined counterfactual choices, which can be partially identified by making nonparametric assumptions about the 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 decision-makers' preferences.

By leveraging these insights with a new payroll dataset recording exact weekly hours paid at the individual level, I estimate that U.S. 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 in the data, 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).

BERTANHA, M., MCCALLUM, A. H. and SEEGERT, N. (2020). "Better Bunching , Nicer Notching". *SSRN Working Paper*.

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.

BLOMQUIST, S., KUMAR, A., LIANG, C.-Y. and NEWEY, W. (2021). "On Bunching and Identification of the Taxable Income Elasticity". *Journal of Political Economy* 129 (8).

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

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.

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

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

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).

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

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.

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.

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

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

QUACH, S. (2021). "The Labor Market Effects of Expanding Overtime Coverage". *MPRA 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.

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.

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