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

Leonard Goff*

This version: April 7, 2024

Abstract

The 1938 Fair Labor Standards Act mandates overtime premium pay for most U.S. workers, but it has proven difficult to assess the policy's impact on the labor market because the rule applies nationally and has varied little over time. I use the extent to which firms bunch workers at the overtime threshold of 40 hours in a week to estimate the rule's effect on hours, drawing on data from individual workers' weekly paychecks. To do so I generalize a popular identification strategy that exploits bunching at kink points in a decision-maker's choice set. Making only nonparametric assumptions about preferences and heterogeneity, I show that the average causal response among bunchers to the policy switch at the kink is partially identified. The bounds indicate a relatively small elasticity of demand for weekly hours, suggesting that the overtime mandate has a discernible but limited impact on hours and employment.

*Department of Economics, University of Calgary. 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, Désiré Kédagni, 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. Some additional results are posted as further appendices on my website `here`. This paper was previously titled "Treatment Effects in Bunching Designs: The Impact of the Federal Overtime Rule on Hours".

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] and remains as the Department of Labor mulls a major expansion to eligibility expected to be announced in late 2023 (U.S. Department of Labor, 2023).

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 previously used bunching at kinks in income tax liability to identify the elasticity of labor supply to the net-of-tax rate (Saez 2010; Chetty et al. 2011).[2] This paper provides new identification results under weakened bunching design assumptions likely suitable to a variety of empirical contexts,

[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 34 and 35 for a comparison of my results to these papers. Quach (2022) 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).

showing that the method can be useful for program-evaluation questions such as assessing the effect of the FLSA.

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 for identification only the cross-sectional distribution of income near a threshold between tax brackets. Analogously, my starting point in the overtime setting is to construct the distribution of weekly work hours. Administrative hours data at the weekly level has previously been unavailable, and studies of overtime in the U.S. have typically relied on self-reported integer hours 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 new data in hand, my 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 decision-makers have parametric "isoelastic" preferences and strong restrictions are placed on heterogeneity. In the overtime setting, I show that bunching is informative about firms (rather that workers) as the decision-maker, choosing the hours of each of their workers in a given week. Having established this, 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 used is considerably more general than the benchmark isoelastic model and its variants. In particular, I find that the method need 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). Agents can furthermore have multiple underlying margins of choice which might be unobserved to the researcher, and preferences can vary flexibly by observational unit. While a non-parametric choice model has previously been considered by Blomquist et al. (2015) to study identification in the bunching design, I extend to a general class of models and distill from this class a common implication about what is observable. My positive identification results then rest on a prediction about choices that remains broadly valid when the isoelastic utility model often employed is misspecified.

The generality of this approach is accomplished by defining the parameter of interest in terms of a pair of counterfactual *choices* rather than as a preference parameter from a parametric choice model, recasting the bunching design in the language of potential outcomes. In the overtime setting these potential outcomes correspond to: a) the number of hours the firm would choose for the worker this week if the worker's normal wage rate applied to all of this week's hours; and b) the number that the firm would choose if the worker's overtime rate applied to all of this week's hours. I show that choice from a kinked choice set can be fully characterized by this pair of counterfactuals: agents choose one or the other of them or they choose the location of the kink. Bunching at the kink directly identifies a feature of the joint distribution of the potential outcomes, allowing one to make statements about treatment effects purged of selection bias.[3]

While generalizing the choice model underlying the bunching design, I also 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 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, is in-part testable, and can be economically motivated in the case of hours. 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 yields an average wage elasticity of hours demand. While the buncher ATE represents a reduced form quantity, I leverage additional assumptions to use it for assessing the overall average effect of the FLSA.

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. (2023) and Blomquist et al. (2021) obtain bounds on the elasticity when the researcher is willing to put an explicit limit on how sharply the density of heterogeneous choices can rise or fall. My approach based on bi-log-concavity avoids the need to choose any such tuning parameters, and is applicable in the general choice model.

I also show that the data in the bunching design are informative about counterfactual policies that change the location or "sharpness" of a kink. To do so, I extend a characterization of bunching from Blomquist et al. (2015), and show that when combined with a general continuity equation (Kasy, 2022) 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: e.g. lowering the overtime

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

threshold below 40 hours (e.g. the *Thirty-Two Hour Workweek Act* proposed in the U.S. House of Representatives in 2021), or increasing the premium pay factor from 1.5 to 2.

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

Empirically, I find that the FLSA overtime rule does in fact reduce hours of work among hourly workers, despite the theoretical possibility that offsetting wage adjustments might eliminate any such effect (Trejo, 1991). My preferred estimate suggests that about one quarter of the bunching observed at 40 among hourly workers is due to the FLSA, and those working at least 40 hours work, on average, about 30 minutes less in a week than they would absent the time-and-a-half rule. Across specifications, I obtain estimates of the local wage elasticity of weekly hours demand near 40 hours in the range -0.04 to -0.19, indicating that firms are fairly resistant to changing hours to avoid overtime payments. A back-of-the-envelope calculation using these effects 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 B expanding on some of the supporting formal results. Section 5 applies these results to estimate effect of the FLSA overtime rule on work hours, as well as the effects of proposed reforms to the FLSA. Section 6 discusses the empirical findings from the standpoint of policy objectives, and 7 concludes.

2 Conceptual framework

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

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

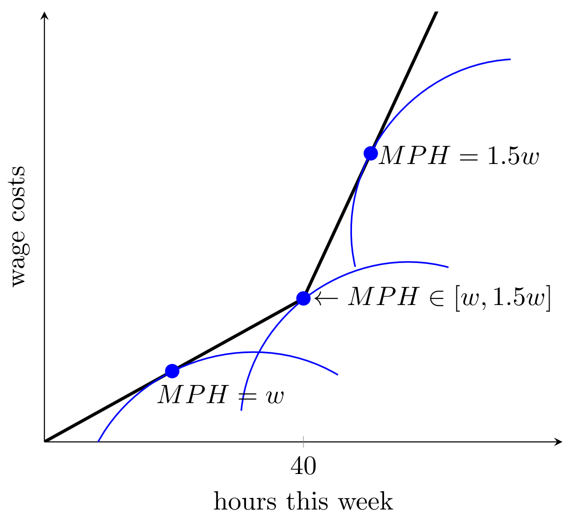

FIGURE 1: With a given worker's straight-time wage fixed at w, labor costs as a function of hours have a convex kink at 40 hours, given the overtime rule. Simple models of week-by-week hours choice (see Section 4.2) yield bunching when for some workers, the marginal product of an hour at 40 is between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

Wages and anticipated hours set at hiring

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

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

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

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

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

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

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

Existing work has investigated whether the fixed-job or fixed-wage model better accords with the observed joint distribution of hourly wages and hours (Trejo, 1991; Barkume, 2010). These

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

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

[6]In Appendix I.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.

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

papers find that wages do tend to be lower among jobs that have overtime pay provisions and more overtime hours, but by a magnitude smaller than would be predicted by the pure fixed job model. These estimates could be driven by selection however, e.g. of lower-skilled workers into covered jobs with longer hours. In Appendix E.3, I construct a new empirical test of Eq. (1) (at the level of individual paychecks), that is instead based on assuming that the conditional distribution of pay is smooth across 40 hours. I find that roughly one quarter of paychecks around 40 hours reflect the wage/hours relationship predicted by the fixed-job model.

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

Dynamic adjustment to hours by week

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

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

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

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

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

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

3 Data and descriptive patterns

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

3.1 Sample description

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

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

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

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

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 firm administrative data when possible).[11]

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

3.2 Hours and wages in the sample

I turn now to the main variables to be used in the analysis. Figure 2 reports the distribution of hours of work in the final sample of paychecks. The graphs indicate a large mass of individuals who were paid for exactly 40 hours that week, amounting to about 11.6% of the sample.[12] Appendix Figure E.9 shows that overtime pay is present in virtually 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, and when hours change the magnitude is about 7 hours on average and roughly symmetric around zero.[13]

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

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

[13]Appendix E reports some further details from the data. Figure 1 shows the distribution of between-paycheck hours changes. Table E.1 documents the prevalence of overtime pay by industry. Table 1 regresses hours, overtime, and bunching on worker and firm characteristics, showing that bunching and overtime hours are predicted by recent hiring at the firm. Table 2 shows that about 63% of variation in total hours can be explained by worker and employer-by-date fixed effects. Figure E.13 considers the joint distribution of wages and hours and reproduces Bick et al.'s (2022) finding that mean wages increase with hours until just beyond 40, before declining.

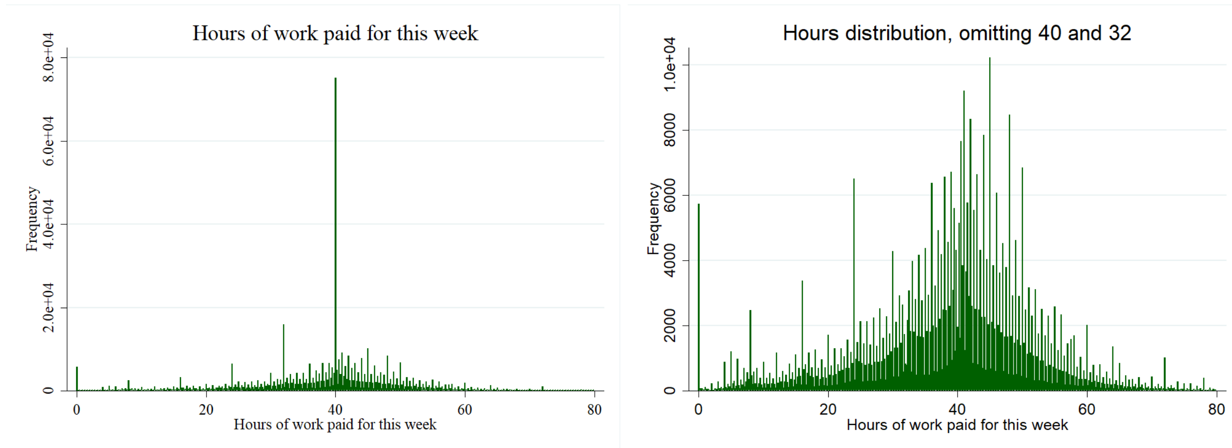

FIGURE 2: Empirical densities of hours worked pooling all paychecks in final estimation sample. Sample is restricted to hourly workers receiving overtime pay at some point (to ensure nearly all are non-exempt from FLSA, see text), and workers having hours variation. The right panel omits the points 40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

	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.

4 Empirical strategy: a generalized kink bunching design

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

The identification results in this section hold in 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 B. Throughout this section I refer to a worker i in week t as a *unit*: an observation of h_{it} for unit it is thus the hours recorded on a single paycheck.

4.1 A general choice model

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

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

where $B_{0it}(h) = w_{it}h$ and $B_{1it}(h) = 1.5w_{it}h - 20w_{it}$. The kinked pay schedule $B_{it}(h)$ is equal to $B_{0it}(h)$ for values $h \leq 40$ and $B_{it}(h)$ is equal to $B_{1it}(h)$ for values $h \geq 40$. The functions B_0 and B_1 recover the two segments in Figure 1 when restricted to these domains respectively (see Appendix Figure B.2). The following definition is generalized in Appendix B:

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

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

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

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

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

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

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

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

For the sake of brevity, I have above stated a version of CONVEX that is a bit stronger than necessary for the identification results below. Appendix B relaxes CONVEX to allow for "double-peaked" preferences with one peak located exactly at the kink (this is relevant if firms have a special preference for a 40 hour work week). The appendix also shows that bunching still has some identifying power without any convexity of preferences. Note that the assumption that firms rather than workers choose hours enters in the claim that π is decreasing (rather than increasing) in z, but Appendix C 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 found at the corner solution of 40 if and only if the two counterfactual outcomes fall on either side, "straddling" the kink.[15] Figure 3 depicts the implications of Eq. (2) for what is therefore observable by the researcher in the bunching design: censored distributions of h_0 and of h_1, and a point-mass of $\mathcal{B} = P(h_{1it} \leq 40 \leq h_{0it})$ at the kink.

Equation (2) represents a central departure from most previous approaches to the bunching

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

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

Observed distribution of hours

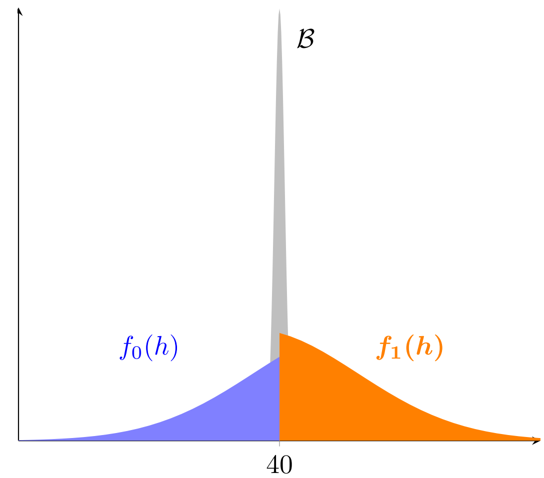

FIGURE 3: Observables in the bunching design, given Equation (2). To the left of the kink at 40, the researcher observes the density $f_0(h)$ of the counterfactual h_{0it}, up to values $h = 40$. To the right of the kink, the researcher observes the density $f_1(h)$ of h_{1it} for values $h > 40$. At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

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

Intuition for Equation (2) in the overtime setting

As an illustration of Equation (2), suppose that firms balance the cost $B_{it}(h)$ against the value of h hours of the worker's labor, in order to maximize that week's profits. Then 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 that unit's 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 $h_{1it} = MPH_{it}^{-1}(1.5w_{it})$.

[16]Blomquist et al. (2015) also derive an expression for $\mathcal{B}$ in terms of agents' choices given all intermediate slopes between those occurring on either side of the kink. I discuss this and offer a generalization in Appendix Lemma 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 B.3 provides some examples that use the full generality of Assumption CONVEX, in which firms simultaneously consider *multiple* margins of choice aside from a given unit's hours. For example, the firm may attempt to mitigate the added cost of overtime by reducing bonuses when a worker works many overtime hours. Eq. (2) remains valid even when such additional margins of choice are unmodeled and unobserved by the econometrician, varying possibly by unit.

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. 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. Section 4.4 considers how in general, interdependencies affect the interpretation of our treatment effects, while Appendix H discusses the impact of nonseparable production functions in more detail.

4.2 The benchmark isoelastic model

This section specializes even further to review a particularly simple special case of the general choice model just presented, which has served as the canonical approach in the bunching-design literature (Saez, 2010; Chetty et al., 2011; Kleven, 2016; Blomquist et al., 2021). This model strengthens Assumption CONVEX to suppose that $\mathbf{x} = h$ and that 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. Letting $\eta_{it} = a_{it}/w_{it}$ denote the ratio of a unit's current productivity factor a_{it} to their straight wage, we have:

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

By Eq. (3), actual hours h_{it} are thus ranked across units in order of η_{it}, and the value of η_{it} determines whether a worker works overtime in a given week. If η_{it} is continuously distributed with support overlapping the interval $[40^{-1/\epsilon}, 1.5 \cdot 40^{-1/\epsilon}]$, then the observed distribution of h_{it} will feature a point mass at 40—"bunching"—and a density elsewhere.

Identification in the isoelastic model

In the context of the isoelastic model, a natural starting place for evaluating the FLSA 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.

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 simply solving Eq. (5) for $|\epsilon|$. However, f_0 is not globally identified from the data: from Figure 3 we can see that f_0 is only identified to the left of the kink, while the density of h_1 is identified to the right of the kink. Since $h_{1it} = 1.5^{\epsilon} \cdot h_{0it}$, it is convenient in the isoelastic model to analyze observables after applying a log transformation to hours: the quantity $\delta = \ln h_{0it} - \ln h_{1it} = |\epsilon| \cdot \ln 1.5$ is homogeneous across all units it, and the density of $\ln h_{1it}$ is thus a simple leftward shift of the density of $\ln h_{0it}$, by δ, as shown in Figure 4.

Standard approaches in the bunching design make parametric assumptions that interpolate f_0 through the missing region of Figure 4 to point-identify ϵ.[17] The approach of Saez (2010) assumes for example that the density of h_0 is linear through the missing region $[40, 40 \cdot e^{\delta}]$ of Figure 4. The

[17] Bertanha et al. (2023) note that given a full parametric distribution for f_0, the entire model could be estimated by maximum likelihood. This approach would enforce (5) automatically while enjoying the efficiency properties of MLE.

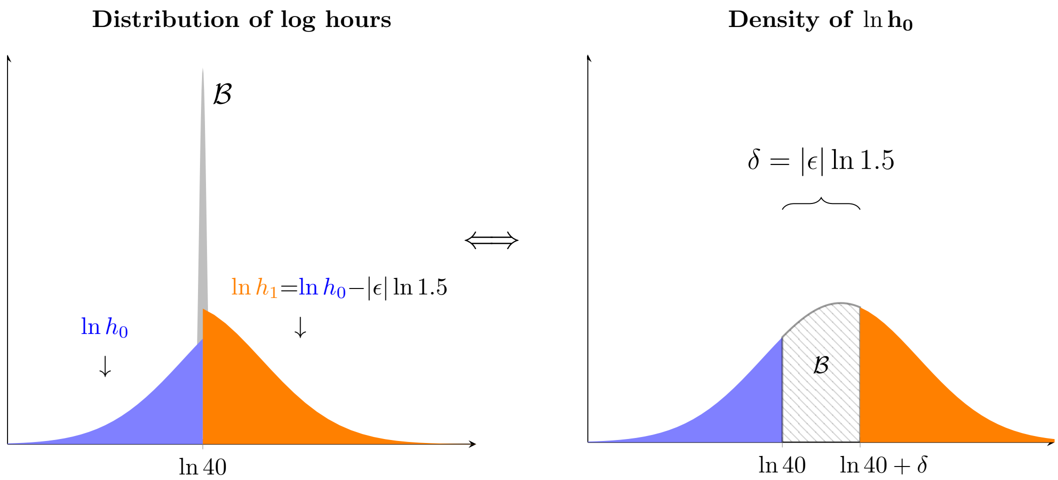

FIGURE 4: The left panel depicts the distribution of observed log hours $\ln h_{it}$ in the isoelastic model, while the right panel depicts the underlying full density of $\ln h_{0it}$. Specializing from the general setting of Figure 3, we have in the isoelastic model that $\tilde{f}_1(h) = \tilde{f}_0(h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

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. Meanwhile, the method of Chetty et al. (2011) ignores the "shift" by δ in the right panel of Figure 4. Both of these approaches ultimately rely on parametric assumptions, and sufficient conditions for each are outlined in Appendix J.2.

If in the other extreme, the researcher is unwilling to assume anything about the density of h_0 in the missing region of Figure 4, then the data are compatible with any finite $\epsilon < 0$, a point emphasized by Blomquist et al. (2021) and Bertanha et al. (2023). In particular, given (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 partial identification. 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 for 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 G.1, assuming the isoelastic model and that h_0 and h_1 are BLC suggests that $\epsilon \in [-.179, -.168]$.[18] Such values are implausible when interpreted through the lens of Equation (4): $\epsilon = -.175$ for example would imply an hours production function of $f(h) \approx -\frac{1}{4.7}h^{-4.7}$ (up to an affine transformation), which features an unrealistic degree of concavity.

In short, the observed bunching at 40 hours is too small to be reconciled with a model in which a single ϵ parameterizes the concavity of weekly production with respect to hours. This motivates a model like the one presented in Section 4.1, in which we can interpret the estimand of the bunching design as a *reduced-form* averaged elasticity of the demand for hours. As described through some examples in Appendix B.3, this elasticity may reflect adjustment by firms along additional margins that can attenuate the hours response, and thus reduce the magnitude of bunching.

4.3 Identifying treatment effects in the general choice model

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

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

[18] 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:[19]

$$\mathcal{B} = F_1(k) - F_0(k) \tag{7}$$

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

Parameter of interest: the buncher ATE

I focus my identification analysis on the average treatment effect among units who locate at exactly 40 hours, a parameter I call the "buncher ATE". In the overtime setting some additional care is needed in defining this parameter, 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.

While the buncher ATE captures a reduced form labor demand response in levels (i.e. measured as a difference in hours), it can be related directly to the elasticity of labor demand by first applying a log transformation to hours. In the isoelastic model, for example, $\delta^*_k := \mathbb{E}[\ln h_{0it} - \ln h_{1it} | h_{it} = k, K^*_{it} = 0] = \epsilon \cdot \ln(1.5)$. This expression holds in general with ϵ replaced by a weighted average of local elasticities among the bunchers—see Appendix K.6 eq. (19) for an explicit expression.[20]

To simplify the discussion, suppose for the moment that $p = 0$, 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

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

[20] Further, the bounds on the buncher ATE presented in Theorem 1 can be easily translated into bounds on the buncher ATE in logs: $\delta^*_k \in [\Delta^L_k / k, \Delta^U_k / k]$. So, Theorem 1 delivers bounds on a weighted-average of the elasticity of demand.

probability $\mathcal{B}$. In Figure 4, we've seen the intuition for this exercise in the context of the isoelastic model, in which there is only a scalar degree 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.:

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

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

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

Assumption RANK. *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 $h_{1it} \in [k - \Delta_1^*, k]$.*

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.[21] For example, suppose that a 50% increase in the wage of worker i would result in their hours being reduced from $h_{0it} = 50$ to $h_{1it} = 45$. If another worker j's hours are instead reduced from $h_{0jt} = 48$ to $h_{1jt} = 46$ under a 50% wage increase, workers i and j will switch ranks, without violating RANK. Note that RANK is also compatible with the existence of counterfactual bunchers $p > 0$.

The right panel of Figure 5 shows an example of a distribution satisfying RANK, which requires the support of (h_0, h_1) to narrow to a point as it crosses $h_0 = k$ or $h_1 = k$. If this is not perfectly satisfied, Appendix B.5 demonstrates how the RHS of Equation (8) will then yield a lower bound on the true buncher ATE (and can still be interpreted as an averaged quantile treatment effect). Ap-

[21] 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 $U_d|(h > k) \sim Unif[F_1(k), 1]$.

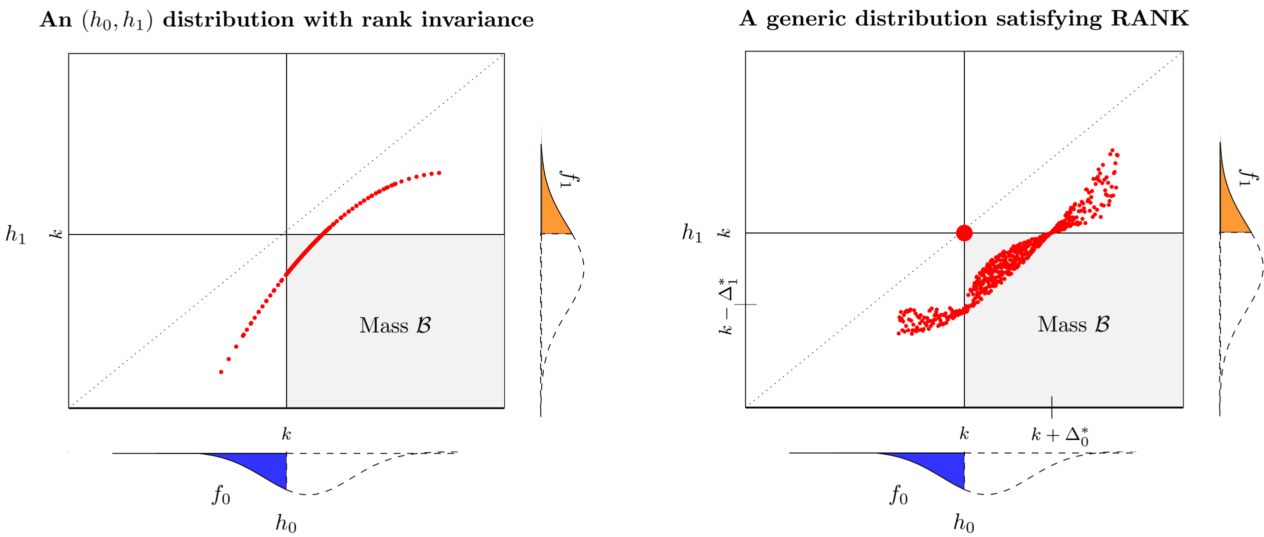

FIGURE 5: The joint distribution of (h_{0it}, h_{1it}) (in red), comparing an example satisfying rank invariance (left) to a case satisfying Assumption RANK (right). RANK allows the support of the joint distribution to "fan-out" from perfect co-dependence of h_0 and h_1, except when either outcome is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

pendix Figure C.6 generalizes RANK to case in which some workers choose their hours, resulting in mass also appearing in the north-west quadrant of Figure 5.

Remark: Neither of Assumptions RANK nor CONVEX require that $h_{0it} \geq h_{1it}$ with probability one. While this is true in the examples of Figure 5 above, Appendix Figure B.5 depicts an example of a joint distribution satisfying RANK in which some units it have negative treatment effects.

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

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

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

If F is BLC then it admits a strictly positive density f that is itself differentiable with locally bounded derivative: $\frac{-f(h)^2}{1-F(h)} \leq f'(h) \leq \frac{f(h)^2}{F(h)}$ (Dümbgen et al., 2017). Intuitively, this rules out cases in which the density of h_0 or h_1 ever spikes or falls *too* quickly on the interior of its support, leading to non-identification of the type discussed in Section 4.2.[22] Note that for a given value $f(h)$, BLC constrains $f'(h)$ more the closer h is to the median of distribution F.

The assumption that h_0 and h_1 admit BLC distributions can be justified in three primary ways. First, it weakens parametric distributions distributional assumed by previous bunching design studies. BLC nests as a special case distributions with log-concave densities, such as the linear counterfactual density assumption used by Saez (2010), and more generally polynomial densities (when they have real roots) used by Chetty et al. 2011.[23] Secondly, the BLC property is partially testable in the bunching design, since $F_0(y)$ is observable for all $h < k$ and $F_1(h)$ is observable for all $h > k$. Appendix Figure E.10 shows that the observable portions of F_0 and F_1 indeed satisfy BLC. Identification then simply requires us to believe that BLC *also* holds in the unobserved portions of F_0 and F_1.

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

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

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

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

2. *The buncher ATE Δ^*_k lies in the interval $\left[\Delta^L_k, \Delta^U_k\right]$, where:*

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

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

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

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

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

Proof. See Appendix A. □

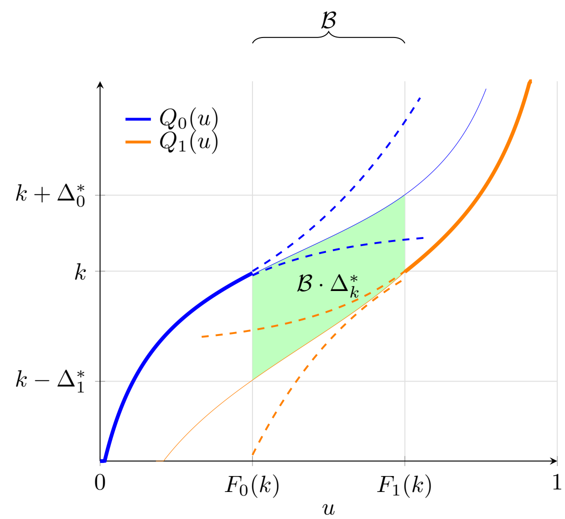

FIGURE 6: Extrapolating the quantile functions for h_0 and h_1 (blue and orange, respectively) to place bounds on the buncher ATE (case depicted has no counterfactual bunchers). The observed portions of each quantile function are depicted by thick curves, while the unobserved portions are indicated by thinner curves. The dashed curves represent upper and lower bounds for this unobserved portion coming from an assumption of bi-log-concavity. The buncher ATE is equal to the area shaded in green, divided by the bunching probability $\mathcal{B}$.[24]The quantities Δ_0^* and Δ_1^* are defined in Assumption RANK below.

Combining Items 1 and 2 of Theorem 1, it follows that the bounds Δ_k^L and Δ_k^U on the buncher ATE are identified, given the CDF $F(h)$ of hours and p.[25] Inspection of the expressions appearing in Theorem 1 reveals that Δ_k^U is always weakly larger than Δ_k^L, and the difference between the two grows the larger the net bunching probability $\mathcal{B} - p$. Some algebra also shows that when net bunching $\mathcal{B} - p$ is strictly positive $\Delta_k^L > 0$, so that the buncher ATE can be bounded away from zero.

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

[24]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 B.4). 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.

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

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

A second existing result 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.[26] Theorem 1 avoids the need for such an assumption.

4.4 Estimating policy relevant parameters

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

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

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

[26] 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.

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. Letting h^*_{it} indicate the hours unit it would work absent the FLSA, consider the parameter $\theta := \mathbb{E}[h_{it}] - \mathbb{E}^*[h^*_{it}]$, where the second expectation $\mathbb{E}^*$ is over units of workers that would exist in the no-FLSA counterfactual and be covered were it introduced.[27] Defining θ in this way allows us to remain agnostic as to whether the FLSA changes employment, and hence the population of workers it applies to. However, I assume that the hours among any workers who enter or exit employment due to the FLSA are not systematically different from those who would exist without it, so that we may rewrite θ as $\theta = \mathbb{E}[h_{it} - h^*_{it}]$, averaging over individual-level causal effects in the population that does exist given the FLSA.

Next, decompose θ as:

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

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

The first term in Equation (9) reflects the "effect of the kink" quantity $h_{it} - h_{0it}$ examined in Section 4.2, and I view it as the first-order object of interest. The second term reflects that straight-time wages w_{it} may differ from those that workers would face without the FLSA, denoted by w^*_{it}. The third term is zero when firms' choice of hours for their workers decomposes into separate optimization problems for each unit, as in the benchmark model from Section 4.2. More generally, it will capture any interdependencies in hours across units, for instance due to different workers' hours being not linearly separable in production. In Appendix H 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.[28]

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

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

[28]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 H. If the third term of Eq. (9) is not zero, my strategy still estimates the average of a unit-level labor demand elasticity in which the hours of a worker's colleagues are fixed.

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

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

and thus $\mathbb{E}[h_{it} - h_{0it}] = \mathcal{B} \cdot \mathbb{E}[k - h_{0it}|h_{it} = k] - P(h_{it} > k)\mathbb{E}[\Delta_{it}|h_{it} > k]$. To identify this quantity we must extrapolate from the buncher ATE to obtain an estimate of $\mathbb{E}[\Delta_{it}|h_{it} > k]$, the average effect for units who work overtime. To do this, I assume that the Δ_{it} of units working more than 40 hours are at least as large on average as those who work exactly 40, but that the reduced-form *elasticity* of their response is no greater than that of the bunchers. The logic is as follows: assuming a constant percentage change between h_0 and h_1 over units would imply responses that grow in proportion to h_1, eventually becoming implausibly large. On the other hand, it would be an underestimate to assume high-hours workers, say at 60 hours, have the same effect in levels $h_0 - h_1$ as those closer to 40. Finally, I use bi-log-concavity of h_0 to put bounds on the average effect of the kink among bunchers $\mathcal{B} \cdot \mathbb{E}[k - h_{0it}|h_{it} = k]$. Details are provided in Appendix K.6.

The "wage effects" term in Equation (9) arises because the straight-time wages observed in the data may reflect some adjustment to the FLSA, as we would expect on the basis of the conceptual framework in Section 2.While the "effect of the kink" term is expected to be negative, this second term will be positive if the FLSA causes a reduction in the straight-time wages set at hiring. However, both terms ultimately depend on the same thing: responsiveness of hours to the cost of an hour of work. We can thus use the buncher ATE to compute an approximate upper bound on wage effects by assuming that all straight-time wages are adjusted according to Equation (1) and that the hours response is iso-elastic in wages, with anticipated hours approximated by h_{it}. Appendix K.6 provides a visual depiction of the logic. A lower bound on the "wage effects" term, on the other hand, is zero. In practice, the estimated size of the wage effect $\mathbb{E}[h_{0it} - h^*_{0it}]$ is appreciable but still small relative to $\mathbb{E}[h_{it} - h_{0it}]$ (cf. Appendix Table 7).

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,[29] 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.

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

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, assuming that the mass of counterfactual bunchers p remains at $k = 40$ after the shift.[30]

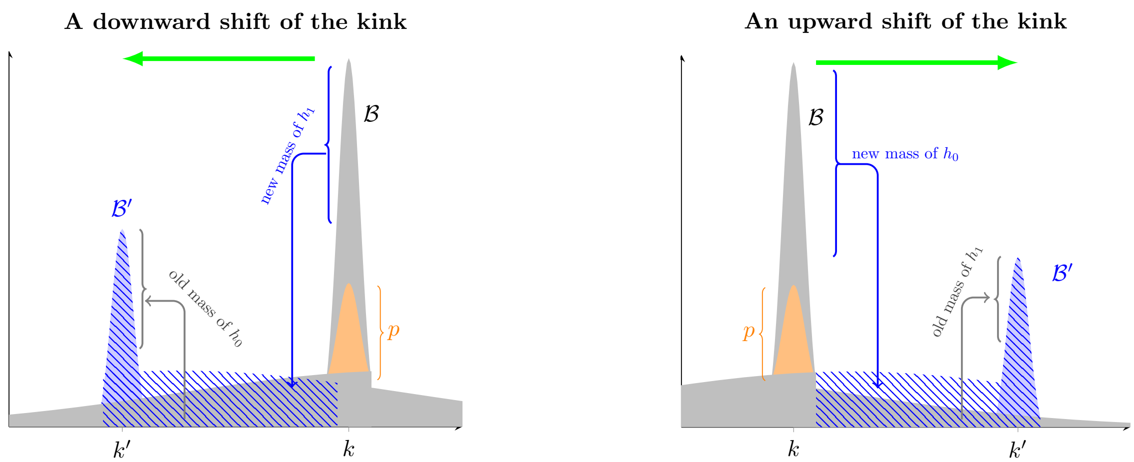

FIGURE 7: The left panel depicts a shift of the kink point downwards from k to k', while right panel depicts a shift of the kink point upwards. See text for details.

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 pay factor of ρ_1 is due from employers for hours in excess of k, e.g. $\rho_1 = 2$ for a "double-time" policy. Let $h_{it}^{[k,\rho_1]}$ denote realized hours for unit it under this overtime policy as a function of k and ρ_1, and let $\mathcal{B}^{[k,\rho_1]} := P(h_{it}^{[k,\rho_1]} = k)$ be the observable bunching that would occur. I will use ∂_k and ∂_{ρ_1} to denote partial derivatives with respect to k and ρ_1, respectively.

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

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

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

1. $\partial_k \left\{ \mathcal{B}^{[k,\rho_1]} - p(k) \right\} = f_1(k) - f_0(k)$

2. $\partial_k \mathbb{E}[h_{it}^{[k,\rho_1]}] = \mathcal{B}^{[k,\rho_1]} - p(k)$

3. $\partial_{\rho_1} \mathcal{B}^{[k,\rho_1]} = -k f_{\rho_1}(k) \mathbb{E}\left[\frac{dh_{it}(\rho_1)}{d\rho} \middle| h_{it}(\rho_1) = k \right]$

4. $\partial_{\rho_1} \mathbb{E}[h_{it}^{[k,\rho_1]}] = -\int_k^\infty f_{\rho_1}(h) \mathbb{E}\left[\frac{dh_{it}(\rho_1)}{d\rho} \middle| h_{it}(\rho_1) = h \right] dh$

Proof. See Appendix B. □

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

Beginning from the actual FLSA policy of $k = 40 = k^*, \rho_1 = 1.5$, the RHS of Items 1 and 2 are in fact point identified from the data, provided that p is known. Item 1 says that if the location k of the kink is changed marginally, the kink-induced bunching probability will change according to the difference between the densities of h_{1i} and h_{0i} at k^*, which are in turn equal to the left and right limits of the observed density $f(h)$ at the kink. This result is intuitive: given continuity of each potential outcome's density, a small increase in k will result in a mass proportional to $f_1(k)$ being "swept in" to the mass point at the kink, while a mass proportional to $f_0(k)$ is left behind. Item 2 aggregates this change in bunching with the changes to non-bunchers' hours as k is increased: the combined effect turns out to be to simply transport the mass of inframarginal bunchers to the new value of k.[31] Making use of Theorem 2 for a discrete policy change like reducing standard hours to 35 requires integrating across the actual range of hypothesized policy variation. We lose point identification, but I use bi-log-concavity of the marginal distributions of h_0 and h_1 to retain bounds.

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

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

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

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

Recall that in the isoelastic model the elasticity quantity $\frac{d \ln h_{it}(\rho)}{d \ln \rho} = \frac{dh_{it}(\rho)}{d\rho}\frac{\rho}{h_{it}(\rho)}$ is constant across ρ and across units, and it is partially identified under BLC. Just as a constant proportional response is likely to overstate responsiveness at large values of hours, it is likely to *understate* responsiveness to larger values of ρ. This yields a lower bound on the effect of moving to double-time. For an upper bound on the magnitude of the effect, I assume rather that in levels $\mathbb{E}[h_{it}(\rho_1) - h_{it}(\bar{\rho}_1)|h_{1it} > k]$ is at least as large as $\mathbb{E}[h_{0it} - h_{1it}|h_{1it} > k]$, and that the increase in bunching from a change of ρ_1 to $\bar{\rho}_1$ is as large as the increase from ρ_0 to ρ_1. Additional details are provided in Appendix K.6.

5 Implementation and Results

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

5.1 Identifying counterfactual bunching at 40 hours

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

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

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

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

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

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

The first item reflects the above logic, and allows me to identify the mass of active bunchers in the $n_{it} > 0$ conditional distribution of hours. The second item says that this conditional mass is representative of the unconditional mass of active bunchers. To increase the plausibility of this assumption, I focus on η as paid time off because it is generally planned in advance, yet has somewhat idiosyncratic timing.[32]

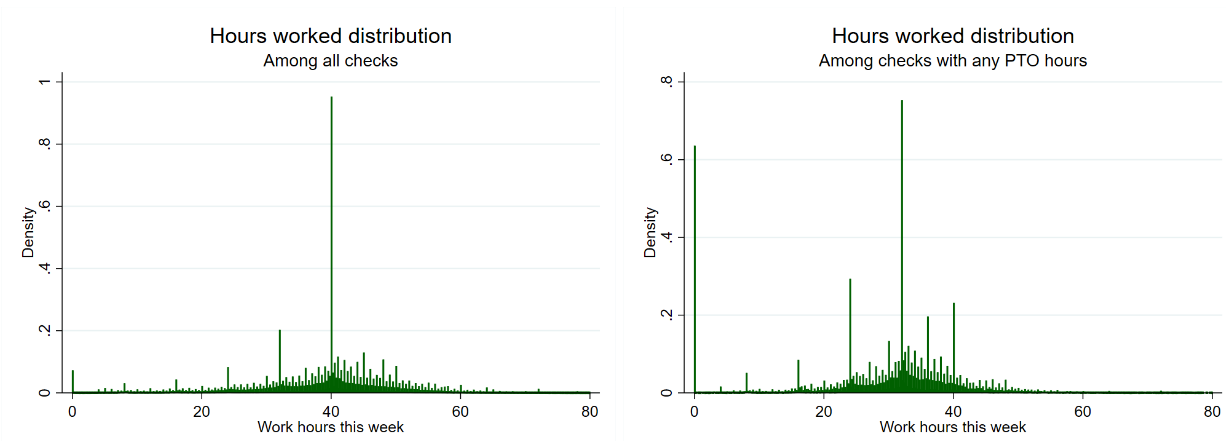

FIGURE 8: The right panel shows a histogram of hours worked when paid time off hours are positive ($\eta_{it} > 0$). The left panel shows the unconditional distribution. While $\mathcal{B} \approx 11.6\%$, $P(h_{it} = 40 | n_{it} > 0) \approx 2.7\%$.

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

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

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

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

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

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

5.2 Estimation and inference

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

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

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

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

5.3 Results of the bunching estimator: the buncher ATE

Table 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 close to each other and precisely estimated. One can show from the expressions for the bounds in Theorem 1 that if $f_0(k) \approx f_1(k)$ and $p \approx 0$, the bounds will tend to be narrower when $F_0(k)$ is closer to $(1 - \mathcal{B})/2$, i.e. the kink is close to the median of the latent hours distribution. This provides some intuition for why the bounds are reasonably narrow, since hours are roughly evenly divided to either side of 40 hours (cf. Figure 2).

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

TABLE 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. Firms would ask this group to work on average about 40 minutes more that week if they were paid their straight-time wage for all hours, compared with a counterfactual in which they are paid their overtime rate for all hours. If we instead attribute all of the observed bunching mass to active bunchers ($p = 0$), then the buncher ATE is estimated to be at least 2.6 hours. In Appendix E I also report estimates based on alternative shape constraints and assumptions about effect heterogeneity (with similar results).

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,[33] which I use as an input in these calculations. The next two rows report bounds on $\mathbb{E}[h_{it} - h^*_{it}]$ and $\mathbb{E}[h_{it} - h^*_{it}|h_{1it} \geq 40, K^*_{it} = 0]$, respectively. The second row is the overall ex-post effect of the FLSA on hours, averaged over workers and pay periods, and the third row conditions on paychecks reporting at least 40 hours (omitting counterfactual bunchers). The final row reports an estimate of the effect of moving to double-time pay.

	p=0	*p* from non-changers	*p* from PTO
Buncher ATE as elasticity	[-0.188,-0.161]	[-0.088,-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 $\mathbb{E}[h_{it} - h^*_{it}]$,[33] 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.

Taking the PTO-based estimate of p as yielding a lower bound on treatment effects, the estimates suggest that workers work at least about 1/4 of an hour less on average in a given week than

[33] This is $\hat{\Delta}^*_k/(40\ln(1.5))$ where $\hat{\Delta}_k$ is the estimate of the buncher ATE presented in Table 3. This is numerically equivalent to the elasticity implied by the buncher ATE in logs $\mathbb{E}[\ln h_{0it} - \ln h_{1it}|h_{it} = k, K^*_{it} = 0]/(\ln 1.5)$ estimated under assumption that $\ln h_0$ and $\ln h_1$ are BLC.

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

Table 4 also suggests that a move to double-time pay would introduce a further reduction in hours comparable to the existing ex-post effect of the FLSA, but the bounds are wider. These estimates include the effects of possible adjustments to straight-time wages, which tend to attenuate the impact of the policy change. Appendix Table 7 replicates Table 4 neglecting wage adjustments, which might be viewed as a short-run response to the FLSA before wages adjust.

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

Appendix Figure E.12 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 E.11 shows these bounds as k' ranges all the way from 0 to 80, for the $p = 0$ case.

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 E.12 shows estimates for the average effect on hours of changing standard hours, inclusive of wage effects (see Appendix K.6 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.[34] The size of these effects is not precisely estimated for changes larger than

[34] 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.

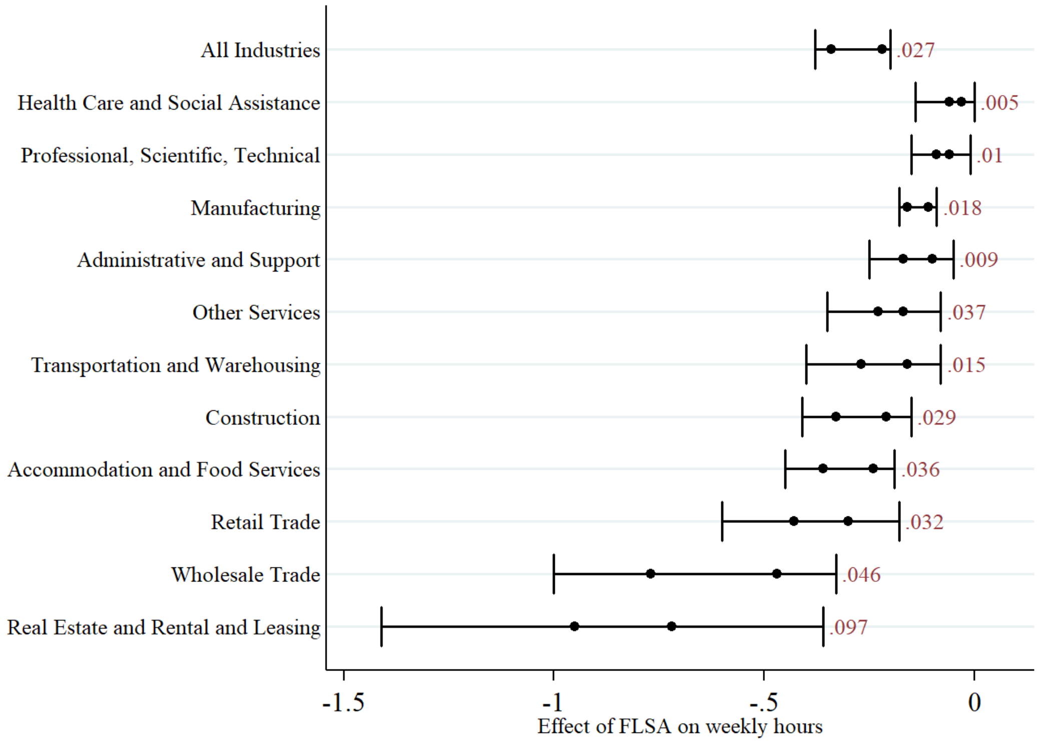

FIGURE 9: 95% confidence intervals for the effect of the FLSA on hours by industry, using PTO-based estimates of p for each. Dots are point estimates of the upper and lower bounds. The number to the right of each range is the point estimate of the net bunching $\mathcal{B} - p$ for that industry.

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.

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.[35] By contrast, my estimates use an identification strategy that does not require focusing on the subpopulation affected by a natural experiment, and are based on recent and administrative data.

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

Nevertheless, the overall impact of the FLSA overtime rule on workers is still notable. The 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 E.4 for details). To get an overall upper bound on the size of employment effects, one can instead attribute all of the bunching at 40 to the FLSA and assume that the total number of worker-hours is not reduced by the FLSA. By this estimate the FLSA increases employment by at most 3 million jobs, or roughly 3% among covered workers. A reasonable range of parameter values in this simple calculation rules out that the FLSA overtime rule has negative overall employment effects on hourly workers.

7 Conclusion

This paper has provided a new interpretation of the popular bunching-design method in the language of treatment effects, showing that the basic identifying power of the method is robust to a wide variety of underlying choice models. Across such models, the parameter of interest remains a reduced-form average treatment effect (local to the kink) between two appropriately-defined counterfactual choices, which is partially identified under a natural nonparametric assumption about

[35] 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.

those counterfactuals' distributions. This provides conditions under which the bunching design can be useful to answer program evaluation questions in a variety of contexts, particularly beyond those in which the researcher is prepared to posit a parametric model of agents' preferences.

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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).

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

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

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

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

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

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

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

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

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

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

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

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

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

KIMBALL, W. and MISHEL, L. (2015). “Estimating the Number of Workers Directly Benefiting from the Proposed Increase in the Overtime Salary Threshold”. *Economic Policy Institute Technical Paper*, p. 12.

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.

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

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

QUACH, S. (2022). “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.

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

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

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

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

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

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

U.S. DEPARTMENT OF LABOR (2023). *Unified Agenda: DOL Wage and Hour Division Rule 1235-AA39*. https://www.reginfo.gov/public/do/eAgendaViewRule?pubId=202304&RIN=1235-AA39.

A Main proofs

A.1 Proof of Theorem 1

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

Call a distribution function $F(h)$ *locally BLC* on an interval N, if $\ln F(h)$ and $\ln(1 - F(h))$ are concave for $h \in N$. Given Assumption RANK, we know that 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 $h_{1it} \in [k - \Delta_1^*, k]$. In the proof that follows we assume that conditional on $K_{it}^* = 0$, h_0 is BLC on $[k, k + \Delta_0^*]$ and h_1 is BLC on $[k - \Delta_1^*, k]$. This is of course implied if these distributions are globally BLC as assumed in the main text.

Let $\mathcal{B}^* := P(h_{it} = k | K^* = 0)$. Given Assumption RANK:

$$
\begin{aligned}
E[h_{0it} - h_{1it} | h_{it} = k, K_{it}^* = 0] &= \frac{1}{\mathcal{B}^*} \int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k)+\mathcal{B}^*} \{Q_{0|K^*=0}(u) - Q_{1|K^*=0}(u)\} du \\
&= \frac{1}{\mathcal{B}^*} \int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k)+\mathcal{B}^*} \{Q_{0|K^*=0}(u) - k\} du + \frac{1}{\mathcal{B}^*} \int_{F_{1|K^*=0}(k)-\mathcal{B}^*}^{F_{1|K^*=0}(k)} \{k - Q_{1|K^*=0}(v)\} dv \\
&= \frac{1}{\mathcal{B}^*} \int_{F_{0|K^*=0}(k)}^{F_{0|K^*=0}(k+\Delta_0^*)} \{Q_{0|K^*=0}(u) - k\} du + \frac{1}{\mathcal{B}^*} \int_{F_{1|K^*=0}(k-\Delta_1^*)}^{F_{1|K^*=0}(k)} \{k - Q_{1|K^*=0}(v)\} dv
\end{aligned}
$$

To replace $F_{0|K^*=0}(k) + \mathcal{B}^*$ with $F_{0|K^*=0}(k + \Delta_0^*)$ and $F_{1|K^*=0}(k) - \mathcal{B}^*$ with $F_{1|K^*=0}(k - \Delta_1^*)$, we use Equation (2), which is established as Lemma 1 in Appendix B. A proof of Lemma 1 is provided below.

Consider first the implication of local BLC that $F_{d|K^*=0}(h)$ is log-concave on an interval between k and $k + t$ for some t. In what follows, we will consider positive $t \in [0, \Delta_0^*]$ for $d = 0$ and negative such $t \in [-\Delta_1^*, 0]$ for $d = 1$). Concavity implies that a first-order Taylor expansion for $\log F_{d|K^*=0}(k + t)$ around k overshoots: i.e. $\log F_{d|K^*=0}(k + t) \leq \log F_{d|K^*=0}(k) + t \cdot \frac{d}{dh} \log F_{d|K^*=0}(k)$. Similarly, that $\log(1 - F_{d|K^*=0}(h))$ is concave on an interval $[k, k + t]$ implies that $\log(1 - F_{d|K^*=0}(k + t)) \leq \log(1 - F_{d|K^*=0}(k)) + t \cdot \frac{d}{dh} \log(1 - F_{d|K^*=0}(k))$. These two inequalities can be rearranged to put upper and lower bounds on $F_{d|K^*=0}(k + t)$:

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

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

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

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

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

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

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

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

where

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

and

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

Similarly, an upper bound is:

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

where

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

and

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

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

$$
F_{0|K^*=0}(k) = \frac{F_0(k)-p}{1-p} \quad \text{and} \quad F_{1|K^*=0}(k) = \frac{F_1(k)-p}{1-p} \quad \text{and} \quad \mathcal{B}^* = \frac{\mathcal{B}-p}{1-p}
$$

$$
f_{0|K^*=0}(k) = \frac{f_0(k)}{1-p} \quad \text{and} \quad f_{1|K^*=0}(k) = \frac{f_1(k)}{1-p}
$$

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

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

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

To obtain the final result, note that the function $g(a, b, x)$ is homogeneous of degree zero. Thus $\Delta_k^* \in [\Delta_k^L, \Delta_k^U]$, with

$$\Delta_k^L := g\left(F_-(k), f_-(k), \mathcal{B} - p\right) + g\left(1 - F(k), f_+(k), \mathcal{B} - p\right)$$

$$\Delta_k^U := -g\left(1 - p - F_-(k), f_-(k), p - \mathcal{B}\right) - g\left(F(k) - p, f_+(k), p - \mathcal{B}\right)$$

where $-$ and $+$ subscripts denote left and right limits.

Sharpness of the bounds: To see that the above bounds $[\Delta_k^L, \Delta_k^L]$ are sharp, we need to show that for each of the bounds Δ_k^L and Δ_k^L, there exists a distribution of potential outcomes consistent with the data and the assumptions of the theorem, for which Δ_k^* is equal to that bound. The approach below bears some similarity to the sharpness proof in Kédagni and Mourifié (2020). Since specifying a joint distribution of (h_0, h_1) is equivalent to specifying the joint distribution of $(h_0, h_1)|K^* = 0$ as well as p (which is known), I focus on the distribution of $(h_0, h_1)|K^* = 0$.

Let us consider the lower bound first. Let $Q(u)$ denote the quantile function of the data (corresponding to the CDF F). Now specify the marginal distribution of h_0 (conditional on $K^* = 0$) to follow the following quantile function:

$$Q^L_{0|K^*=0}(u) := \begin{cases} Q(u) & \text{if } 0 \leq u \leq F_{0|K^*=0}(k) \\ k + \frac{F_{0|K^*=0}(k)}{f_{0|K^*=0}(k)} \cdot \ln\left(\frac{u}{F_{0|K^*=0}(k)}\right) & \text{if } F_{0|K^*=0}(k) \leq u \leq 1 \end{cases}$$

where I've constructed the extrapolated portion for $u \geq F_{0|K^*=0}(k)$ from the lower bound on $Q^L_{0|K^*=0}$ arising from (A.2). Similarly, construct the marginal distribution of h_1 (conditional on $K^* = 0$) as:

$$Q^L_{1|K^*=0}(u) := \begin{cases} k - \frac{1 - F_{1|K^*=0}(k)}{f_{1|K^*=0}(k)} \cdot \ln\left(\frac{1-u}{1 - F_{1|K^*=0}(k)}\right) & \text{if } 0 \leq u \leq F_{1|K^*=0}(k) \\ Q(u) & \text{if } F_{1|K^*=0}(k) \leq u \leq 1 \end{cases}$$

using the upper bound on $Q^L_{1|K^*=0}$ arising from (A.2).

It can be readily verified that both $Q^L_{d|K^*=0}$ above are valid quantile functions defined on the unit interval $u \in [0,1]$: they are increasing and left continuous on $[0,1]$. Furthermore, $Q^L_{0|K^*=0}(u)$ and $Q^L_{1|K^*=0}(u)$ are locally BLC inside the bunching region by construction, and are also globally BLC provided that $F(h)$ is BLC on the regions $(0,k)$ and (k,∞).[36]

To build a *joint* distribution of $(h_0,h_1)|K^*=0$ from the $Q^L_{0|K^*=0}$ and $Q^L_{1|K^*=0}$ functions above, let us impose rank invariance on our constructed distribution. That is, let

$$(h_0,h_1)|K^*=0 \sim (Q^L_{0|K^*=0}(U), Q^L_{1|K^*=0}(U)) \tag{A.4}$$

where U is a uniform $[0,1]$ random variable. Then RANK holds immediately for this distribution.

Note that $Q^L_{0|K^*=0}(u)$ and $Q^L_{1|K^*=0}(u)$ recover the observed distribution $Q(u)$ of h, via Eq. (2). Lastly, we must show that $\Delta^*_k = \Delta^L_k$ when $(h_0,h_1)|K^*=0$ follows (A.4). This follows from the same steps used above to prove that $\Delta^*_k \geq \Delta^L_k$ generally, with the weak inequalities replaced as equalities.

To build a distribution of (h_0,h_1) that meets the upper bound Δ^U_k, we proceed analogously. That is, let

$$Q^U_{0|K^*=0}(u) := \begin{cases} Q(u) & \text{if } 0 \leq u \leq F_{0|K^*=0}(k) \\ k - \frac{1-F_{0|K^*=0}(k)}{f_{0|K^*=0}(k)} \cdot \ln\left(\frac{1-u}{1-F_{0|K^*=0}(k)}\right) & \text{if } F_{0|K^*=0}(k) \leq u \leq 1 \end{cases}$$

and

$$Q^U_{1|K^*=0}(u) := \begin{cases} k + \frac{F_{1|K^*=0}(k)}{f_{1|K^*=0}(k)} \cdot \ln\left(\frac{u}{F_{1|K^*=0}(k)}\right) & \text{if } 0 \leq u \leq F_{1|K^*=0}(k) \\ Q(u) & \text{if } F_{1|K^*=0}(k) \leq u \leq 1 \end{cases}$$

and again impose rank invariance as before.

Note that the quantile functions $Q^B_{d|K^*=0}$ for $d \in \{0,1\}$ and $B \in \{L,U\}$ are valid quantile functions globally across the unit interval, despite the fact that the functions of t defining the upper and lower BLC bounds in (A.1) are not valid CDF functions globally in t. While those functions are continuous and increasing for all t, they will exit the unit interval if extrapolated too far in either direction. This does not affect the constructions $Q^B_{d|K^*=0}(u)$ because they are only defined within the unit interval. Intuitively, the BLC extrapolations of $Q(u)$ in quantile space only need to extend across the bunching interval $u \in [F_{0|K^*=0}(k), F_{1|K^*=0}(k)]$, which is an identified subset of the unit interval (note that the $Q^B_{d|K^*=0}$ are defined above to continue the BLC extrapolation beyond

[36]To see this, note that if $Q(u)$ is BLC on $(0,k)$ and (k,∞), the functions $Q^L_{d|K^*=0}(u)$ are differentiable everywhere on $(0,1)$, even at the points $F_{0|K^*=0}(k)$ and $F_{1|K^*=0}(k)$. This is because the density associated with the BLC extrapolation is itself continuous at the point of extrapolation (one can see this by differentiating the bounds in (A.1) at $t=0$). Thus the log of the CDF $F^L_{d|K^*=0}(h)$ corresponding to each $Q^L_{d|K^*=0}$ is piecewise concave and continuous and with no kink at $h=k$, which is thus a concave function globally. The same applies to the log of $(1-F^L_{d|K^*=0}(h))$.

that for concreteness, and remain in $[0,1]$).

A.2 Proof of Lemma 1

Theorem 1 makes use of Equation (2), which relates observable choices h_{it} to the counterfactual choices h_{0it} and h_{1it}. The result is stated formally in Appendix B as Lemma 1, but I include the proof here as it is central to the identification logic. This proof follows the notation of Appendix B, dropping the t index from observational units for simplicity. I also use the generalized bunching design setting described in Appendix B, where the continuous function $h_i(\mathbf{x})$ relating hours to choice variables can vary by i.

The proof proceeds in the following two steps:

i) First, I show that $h_{0i} \leq k$ implies that $h_i = h_{0i}$, and similarly $h_{1i} \geq k$ implies that $h_i = h_{1i}$. This holds under CONVEX but also under the weaker assumption of WARP.

ii) Second, I show that under CONVEX $h_i < k \implies h_i = h_{0i}$ and $h_i > k \implies h_i = h_{1i}$.

Item i) above establishes the first and third cases of Lemma 1. The only remaining possible case is that $h_{1i} \leq k \leq h_{0i}$. However, to finish establishing Lemma 1, we also need the reverse implication: that $h_{1i} \leq k \leq h_{0i}$ implies $h_i = k$. This comes from taking the contrapositive of each of the two claims in item ii).

Proof of i): Let $\mathcal{X}_{0i} = \{\mathbf{x} : h_i(\mathbf{x}) \leq k\}$ and $\mathcal{X}_{1i} = \{\mathbf{x} : h_i(\mathbf{x}) \geq k\}$. If $h_{0i} \leq k$, then by CHOICE $\mathbf{x}_{B_{0i}}$ is in $\mathcal{X}_0$, where for any budget constraint B, $(z_{Bi}, \mathbf{x}_{Bi})$ are the choices the decision-maker would make under B. Since $B_i(\mathbf{x}) = B_{0i}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}_0$, it follows that $z_{B_{0i}i} \geq B_i(\mathbf{x}_{B_{0i}i})$, i.e. the decision-maker's choice under B_0 is feasible under the kinked budget constraint B. Note that $B_i(\mathbf{x}) \geq B_{0i}(\mathbf{x})$ for all $\mathbf{x}$. By WARP then $(z_{B_ii}, \mathbf{x}_{B_ii}) = (z_{B_{0i}i}, \mathbf{x}_{B_{0i}i})$. Thus $h_i = h_i(\mathbf{x}_{B_ii}) = h_i(\mathbf{x}_{B_{0i}i}) = h_{0i}$. So $h_{0i} \leq k \implies h_i = h_{0i}$. By the same logic we can show that $h_{1i} \geq k \implies h_i = h_{1i}$.

Proof of ii): For any convex budget function $B(\mathbf{x})$, $(z_{Bi}, \mathbf{x}_{Bi}) = \text{argmax}_{z,\mathbf{x}} \{u_i(z, \mathbf{x}) \text{ s.t. } z \geq B(\mathbf{x})\}$. If $u_i(z, \mathbf{x})$ is strictly quasi-concave, then the RHS exists and is unique since it maximizes u_i over the convex domain $\{(z, \mathbf{x}) : z \geq B(\mathbf{x})\}$. Furthermore, by monotonicity of $u(z, \mathbf{x})$ in z we may substitute in the constraint $z = B(\mathbf{x})$ and write

$$\mathbf{x}_{Bi} = \text{argmax}_{\mathbf{x}} u_i(B(\mathbf{x}), \mathbf{x})$$

Suppose that $h_i(\mathbf{x}_{Bi}) \neq k$, and consider any $\mathbf{x} \neq \mathbf{x}_{Bi}$ such that $h_i(\mathbf{x}) \neq k$. Let $\tilde{\mathbf{x}} = \theta\mathbf{x} + (1-\theta)\mathbf{x}^*$ where $\mathbf{x}^* = \mathbf{x}_{Bi}$ and $\theta \in (0,1)$. Since $B(\mathbf{x})$ is convex in $\mathbf{x}$ and $u_i(z,\mathbf{x})$ is weakly decreasing in z:

$$u_i(B(\tilde{\mathbf{x}}),\tilde{\mathbf{x}}) \geq u_i(\theta B(\mathbf{x}) + (1-\theta)B(\mathbf{x}^*),\tilde{\mathbf{x}}) > \min\{u_i(B(\mathbf{x}),\mathbf{x}), u_i(B(\mathbf{x}^*),\mathbf{x}^*)\} = u_i(B(\mathbf{x}),\mathbf{x}) \tag{A.5}$$

where I have used CONVEX in the second step, and that $\mathbf{x}^*$ is a maximizer in the third. This result implies that for any such $\mathbf{x} \neq \mathbf{x}^*$, if one draws a line between $\mathbf{x}$ and $\mathbf{x}^*$, the function $u_i(B(\mathbf{x}),\mathbf{x})$ is strictly increasing as one moves towards $\mathbf{x}^*$. When $\mathbf{x}$ is a scalar, this argument is used by Blomquist et al. (2015) (see Lemma A1 therein) to show that $u_i(B(\mathbf{x}),\mathbf{x})$ is strictly increasing to the left of $\mathbf{x}^*$, and strictly decreasing to the right of $\mathbf{x}^*$. Note that for any (binding) linear budget constraint $B(\mathbf{x})$, the result still holds without monotonicity of $u_i(z,\mathbf{x})$ in z. This is useful for Theorem 1* of Appendix C in which some workers choose their hours.

For any function B, let $u_{Bi}(\mathbf{x}) = u_i(B(\mathbf{x}),\mathbf{x})$, and note that

$$u_{B_i i}(\mathbf{x}) = \begin{cases} u_{B_{0i}i}(\mathbf{x}) & \text{if } \ \mathbf{x} \in \mathcal{X}_{0i} \\ u_{B_{1i}i}(\mathbf{x}) & \text{if } \ \mathbf{x} \in \mathcal{X}_{1i} \end{cases}$$

where B_i is the actual, kinked budget constraint faced by i. Let $\mathbf{x}_{ki}$ be the unique maximizer of $u_{B_i i}(\mathbf{x})$, where $h_i = h_i(\mathbf{x}_{ki})$. Suppose that $h_i < k$. Suppose furthermore that $h_{0i} \neq h_i$, with $h_{0i} = h_i(\mathbf{x}_{0i})$ and $\mathbf{x}_{0i}$ the maximizer of $u_{B_{0i}i}(\mathbf{x})$. Note that we must have that $\mathbf{x}_{0i} \notin \mathcal{X}_{0i}$, because $B_{0i} = B_i$ in $\mathcal{X}_{0i}$ so we can't have $u_{B_{0i}i}(\mathbf{x}_{0i}) > u_{B_{0i}i}(\mathbf{x}_{ki})$ (since $\mathbf{x}_{ki}$ maximizes $u_{B_i}(\mathbf{x})$). Thus $h_{0i} > k$.

By continuity of $h_i(\mathbf{x})$, $\mathcal{X}_{0i}$ is a closed set and $\mathbf{x}_{ki}$ belongs to the interior of $\mathcal{X}_{0i}$. Thus, while $\mathbf{x}_{0i}$ is not in $\mathcal{X}_{0i}$, there exists a point $\tilde{\mathbf{x}} \in \mathcal{X}_{0i}$ along the line between $\mathbf{x}_{0i}$ to $\mathbf{x}_{ki}$. Since $h_i \neq k$ and $h_{0i} \neq k$, Eq. (A.5) then implies that $u_{B_i i}(\tilde{\mathbf{x}}) > u_{B_i i}(\mathbf{x}_{0i})$. Since $u_{B_{0i}i}(\mathbf{x}) = u_{B_i i}(\mathbf{x})$ for all $\mathbf{x}$ in $\mathcal{X}_{0i}$, it follows that $u_{B_{0i}i}(\tilde{\mathbf{x}}) > u_{B_{0i}i}(\mathbf{x}_{0i})$. However, this contradicts the premise that $\mathbf{x}_{0i}$ maximizes $u_{B_{0i}i}(\mathbf{x})$. Thus, $h_i < k$ implies $h_i = h_{0i}$. Figure A.1 depicts the logic visually. The proof that $h_i > k$ implies $h_i = h_{1i}$ is analogous.

A.3 Proof of Theorem 2

This proof continues with the notation of Appendix B, using i rather than it indices. Throughout this proof we let $h_i(\rho,k) = h_i(\rho)$, given Assumption SEPARABLE. By Lemmas 2 and 3

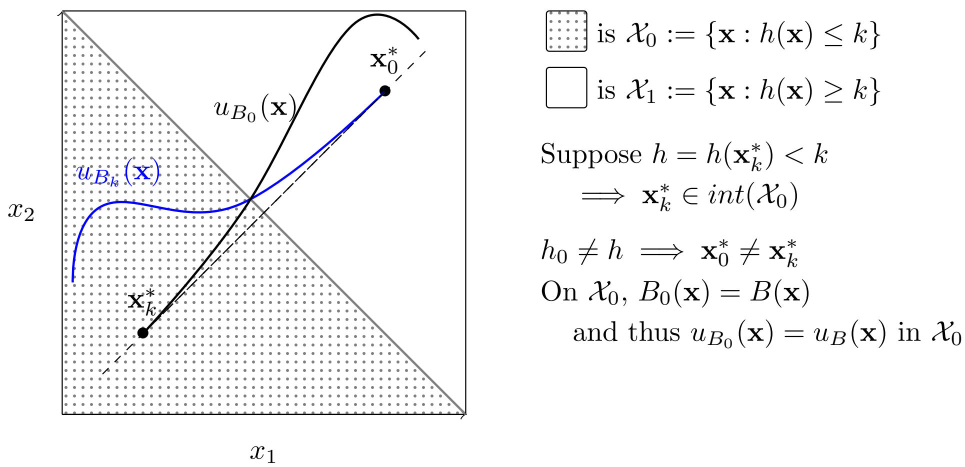

FIGURE A.1: Depiction of the step establishing $(h < k) \implies (h = h_0)$ in the proof of Lemma 1. Since the result considers a single decision-maker i, I supress this index in the Figure. In this example $z = (x_1, x_2)$ and $y(\mathbf{x}) = x_1 + x_2$. Proof is by contradiction. If $h_0 \neq Y$, then $\mathbf{x}_k^* \neq \mathbf{x}_0^*$, where $\mathbf{x}_k^*$ and $\mathbf{x}_0^*$ are the unique maximizers of $u_B(\mathbf{x})$ and $u_{B_0}(\mathbf{x})$, respectively. By Equation A.5, we have that the function $u_{B_0}(\mathbf{x})$, depicted heuristically as a solid black curve, is strictly increasing as one moves along the dotted line from $\mathbf{x}_k^*$ towards $\mathbf{x}_0^*$. Similarly, the function $u_{B_0}(\mathbf{x})$, depicted as a solid blue curve, is strictly increasing as one moves in the opposite direction along the same line, from $\mathbf{x}_0^*$ towards $\mathbf{x}_k^*$. By the assumption that $h < k$, then using continuity of $h(\mathbf{x})$ it must be the case that $\mathbf{x}_k^*$ lies in the interior of $\mathcal{X}_0$, the set of $\mathbf{x}$'s that make $h(\mathbf{x}) \leq k$. This means that there is some interval of the dotted line that is within $\mathcal{X}_0$. On this interval, the functions B_0 and B are equal, and thus so must be the functions u_B and u_{B_0}. Since the same function cannot be both strictly increasing and strictly decreasing, we have obtained a contradiction.

established in Appendix B, the effect of changing k on bunching is:

$$\begin{aligned}
\partial_k \{\mathcal{B} - p(k)\} &= -\frac{\partial}{\partial k} \int_{\rho_0}^{\rho_1} f_\rho(k) \mathbb{E}\left[\frac{h_i(\rho)}{d\rho} \middle| h_i(\rho) = k \right] d\rho \\
&= - \int_{\rho_0}^{\rho_1} \frac{\partial}{\partial k} \left\{ f_\rho(k) \mathbb{E}\left[\frac{h_i(\rho)}{d\rho} \middle| h_i(\rho) = k \right] \right\} d\rho = \int_{\rho_0}^{\rho_1} \partial_\rho f_\rho(k) d\rho = f_1(k) - f_0(k)
\end{aligned}$$

Turning now to the total effect on average hours.

$$\begin{aligned}
\partial_k E[h_i^{[k,\rho_1]}] &= \partial_k \{P(h_i(\rho_0) < k) \mathbb{E}[h_i(\rho_0) | h_i(\rho_0) < k]\} + k \partial_k \left(\mathcal{B}^{[k,\rho_1]} - p(k) \right) + \mathcal{B}^{[k,\rho_1]} - p(k) \\
&\quad + \partial_k \{P(h_i(\rho_1) > k) \mathbb{E}[h_i(\rho_1) | h_i(\rho_1) > k]\} \\
&= \partial_k \int_{-\infty}^{k} y \cdot f_{\rho_0}(y) \cdot dy + k \left(f_0(k) - f_1(k) \right) + \mathcal{B}^{[k,\rho_1]} - p(k) + \partial_k \int_{k}^{\infty} y \cdot f_{\rho_1}(y) \cdot dy \\
&= \cancel{k f_0(k)} + \cancel{k \left(f_1(k) - f_0(k) \right)} + \mathcal{B}^{[k,\rho_1]} - p(k) - \cancel{k f_1(k)}
\end{aligned}$$

Meanwhile: $\partial_{\rho_1}\mathbb{E}[h_i^{[k,\rho_1]}] = -\int_k^\infty f_{\rho_1}(y)\mathbb{E}\left[\left.\frac{dh_i(\rho_1)}{d\rho}\right| h_i(\rho_1)=y\right]dy$ follows directly from Lemma 2 and differentiating both sides with respect to ρ_1, and thus

$$\partial_{\rho_1}E[h_i^{[k,\rho_1]}] = k\partial_{\rho_1}\mathcal{B}^{[k,\rho_1]} + \partial_{\rho_1}\left\{P(h_i(\rho_1)>k)\mathbb{E}[h_i(\rho_1)|h_i(\rho_1)>k]\right\} = k\partial_{\rho_1}\mathcal{B}^{[k,\rho_1]} + \int_k^\infty y\cdot\partial_{\rho_1}f_{\rho_1}(y)\cdot dy$$

$$= -kf_{\rho_1}(k)\mathbb{E}\left[\left.\frac{h_i(\rho_1)}{d\rho}\right| h_i(\rho_1)=k\right] - \int_k^\infty y\cdot\partial_y\left\{f_{\rho_1}(y)\mathbb{E}\left[\left.\frac{dh_i(\rho_1)}{d\rho}\right| h_i(\rho_1)=y\right]\right\}dy$$

$$= \cancel{-kf_{\rho_1}(k)\mathbb{E}\left[\left.\frac{h_i(\rho_1)}{d\rho}\right| h_i(\rho_1)=k\right]} + \cancel{yf_{\rho_1}(y)\mathbb{E}\left[\left.\frac{dh_i(\rho_1)}{d\rho}\right| h_i(\rho_1)=y\right]\Bigg|_\infty^k} - \int_k^\infty f_{\rho_1}(y)\mathbb{E}\left[\left.\frac{dh_i(\rho_1)}{d\rho}\right| h_i(\rho_1)=y\right]dy$$

where I have used Lemma 2 with the Leibniz rule (establishing Item 3 in Theorem 2) as well as Lemma 3 in the third step, and then integration by parts along with the boundary condition that $\lim_{y\to\infty} y\cdot f_{\rho_1}(y)=0$, implied by Assumption SMOOTH.

B Identification in a generalized bunching design

This section presents some generalizations of the bunching-design model used in the main text. While the FLSA will provide a running example throughout, I largely abstract from the overtime context to emphasize the general applicability of the results.

To facilitate comparison with the existing literature on bunching at kinks – which has mostly considered cross-sectional data – I throughout this section suppress time indices and use the single index i to refer to each unit of observation (a paycheck in the overtime setting). Further, the "running variable" of the bunching design is typically denoted by Y rather than h, and so the random variable Y_i will play the role of h_{it} from the main text. This is done to emphasize the link to the treatment effects literature, while also allowing a distinction that is in some cases useful (e.g. in the overtime setting, models in which hours of pay for work differ from actual hours of work).

B.1 The policy environment

Here we abstract from the conventional piece-wise linear kink setting that appears in tax examples as well as the main body of this paper. Consider a population of observational units indexed by i. For each i, a decision-maker $d(i)$ chooses a point $(z,\mathbf{x})$ in some space $\mathcal{X}\subseteq\mathbb{R}^{m+1}$ where z is a scalar and $\mathbf{x}$ a vector of m components, subject to a constraint of the form:

$$z \geq \max\{B_{0i}(\mathbf{x}), B_{1i}(\mathbf{x})\} \tag{B.6}$$

The functions $B_{0i}(\mathbf{x})$ and $B_{1i}(\mathbf{x})$ are taken to be continuous and weakly convex functions of the vector **x**, and assume that there exist continuous scalar functions $y_i(\mathbf{x})$ and a scalar k such that:

$$B_{0i}(\mathbf{x}) > B_{1i}(\mathbf{x}) \text{ whenever } y_i(\mathbf{x}) < k \quad \text{and} \quad B_{0i}(\mathbf{x}) < B_{1i}(\mathbf{x}) \text{ whenever } y_i(\mathbf{x}) > k$$

The value k is taken to be common to all units i, and is assumed to be known by the researcher.[37] In the overtime setting, $y_i(\mathbf{x})$ represents the hours of work for which a worker is paid in a given week, $k = 40$, and $B_{0i}(\mathbf{x}) = w_i y_i(\mathbf{x})$ and $B_{1i}(\mathbf{x}) = 1.5 w_i y_i(\mathbf{x}) - 20 w_i$. In most applications of the bunching design, the decision-maker $d(i)$ is simply i themself, for example a worker choosing their labor supply subject to a tax kink. In the overtime application however i is a worker-week pair, and $d(i)$ is the worker's firm.

Let X_i be i's realized outcome of **x**, and $Y_i = y_i(X_i)$. I assume that Y_i is observed by the econometrician, but not that X_i is. In the overtime setting this means that the econometrician observes hours for which workers are paid, but not necessarily all choices made by firms that pin down those hours (for example, how many hours to allow the worker to stay "on the clock" during paid breaks–see Section B.3).

In general, the functions B_{0i}, B_{1i} will represent a schedule of some kind of "cost" as a function of the choice vector **x**, with two regimes of costs that are separated by the condition $y_i(\mathbf{x}) = k$, characterizing the locus of points at which the two cost functions cross. Let $B_i(\mathbf{x}) := \max\{B_{0i}(\mathbf{x}), B_{1i}(\mathbf{x})\}$ denote the actual constraint function that applies to z. A budget constraint like Eq. $z \geq B_i(\mathbf{x})$ is typically "kinked" because while the function $B_i(\mathbf{x})$ is continuous, it will generally be non-differentiable at the **x** for which $y_i(\mathbf{x}) = k$.[38] While the functions B_0, B_1 and y can all depend on i, I will often suppress this dependency for clarity of notation.

Discussion of the general model:

In the most common cases from the literature, no distinction is made between the "running variable" y of the kink and any underlying choice variables **x**. This corresponds to a setting in which **x** is a scalar and $y_i(x) = x$. For example, the seminal bunching design papers Saez (2010) and Chetty et al. (2011) considered progressive taxation with z being tax liability (or credits), $y = x$ corresponding to taxable income, and B_0 and B_1 linear tax functions on either side of a threshold y between two adjacent tax/benefit brackets. Similarly, in the overtime context, the functions B_0 and B_1 are linear and only depend on hours $y_i(\mathbf{x})$, as depicted in Figure B.2. Appendix J discusses a tax setting in the literature in which the functions B_0

[37]This comes at little cost of generality since with heterogeneous k_i this could be subsumed as a constant into the function $y_i(\mathbf{x})$, so long as the k_i are observed by the researcher.

[38]In particular, the subgradient of $\max\{B_{0i}(\mathbf{x}), B_{1i}(\mathbf{x})\}$ will depend on whether one approaches from the $y_i(\mathbf{x}) > k$ or the $y_i(\mathbf{x}) < k$ side. With a scalar x and linear B_0 and B_1, the derivative of $B_i(x)$ discontinuously rises at **x** for which $y_i(\mathbf{x}) = k$.

and B_1 are linear but depend directly on a vector **x** of two components.[39] This represents a non-standard bunching-design setting, but fits naturally within the framework of this section.

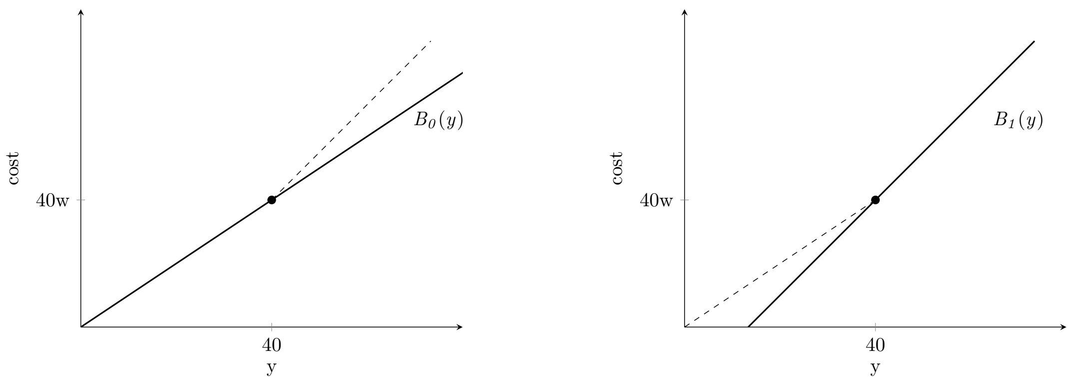

FIGURE B.2: Definition of counterfactual cost functions B_0 and B_1 that firms could have faced, absent the overtime kink. Regardless of what choice variables are in **x**, these functions only depend on $y_i(\mathbf{x})$ and are thus depicted as a function of y. Dashed lines show the rest of actual kinked-cost function in comparison to the counterfactual as a solid line. Note that we use the notation y here to indicate hours, rather than the h used in the main text.

Even when the functions B_0 and B_1 only depend on **x** through $y_i(\mathbf{x})$, as in standard settings, the bunching design is compatible with models in which multiple margins of choice respond to the incentives provided by the kink. As discussed in the overtime context, the econometrician may be agnostic as to even what the full set of components of **x** are, with $B_{0i}(\cdot)$, $B_{1i}(\cdot)$, and $y_i(\cdot)$ depending only on various subsets of the **x** that are possibly heterogeneous by i (this is allowable because y need only be continuous in **x**, and the cost functions only need to be continuous and *weakly* convex in **x**, both of which are compatible with zero dependence on some of its components). Appendix J.5 gives an example in which the overtime kink gives firms an incentive to reduce bonuses, which appear in firm costs but not in the kink the variable y.

In general, the bunching design allows us to conduct causal inference on $Y_i = y_i(X_i)$, but not directly on the underlying choice variables X_i. For example in the overtime setting with possible evasion (see Sec. B.3), bunching at 40 hours will be informative about the effect of a move from B_0 to B_1 on reported hours worked y. However, it will not disentangle whether the effect on hours actually worked is attenuated by, for example, an increase in hours worked off-the-clock. The empirical setting of Best et al. (2015) provides another environment in which this point is relevant (see Appendix J.5).

[39] Best et al. (2015) study firms in Pakistan that pay either a tax on output or a tax on profit, whichever is higher. The two tax schedules cross when the ratio of profits to output crosses a certain threshold that is pinned down by the two respective tax rates. In this case, the variable y depends both on production and on reported costs, leading to two margins of response to the kink: one from choosing the scale of production and the other from choosing whether and how much to misreport costs. In this setting a distinction between y and **x** cannot be avoided. The authors use features of the function $y_i(\mathbf{x})$ to argue that the bunching reveals changes mostly to reported costs rather than to output (see Appendix J.5 for details).

B.2 Potential outcomes as counterfactual choices

Here I restate slightly more general versions of assumptions CONVEX and CHOICE from Section 4, in the present notation. As in Section 4, let us define a pair of potential outcomes as what would occur if the decision-maker faced either of the functions B_0 or B_1 globally, without the kink.

Definition (potential outcomes). *Let Y_{0i} be the value of $y_i(\mathbf{x})$ that would occur for unit i if $d(i)$ faced the constraint $z \geq B_0(\mathbf{x})$, and let Y_{1i} be the value that would occur under the constraint $z \geq B_1(\mathbf{x})$.*

I again make explicit the assumption that these potential outcomes reflect choices made by the decision-maker. For any function B let Y_{Bi} be the outcome that would occur under the choice constraint $z \geq B(\mathbf{x})$, with Y_{0i} and Y_{1i} shorthands for $Y_{B_{0i}i}$ and $Y_{B_{0i}i}$, respectively. In this notation, the actual outcome Y_i observed by the econometrician is equal to $Y_{B_i i}$.

Assumption CHOICE (perfect manipulation of y). *For any function $B(\mathbf{x})$, $Y_{Bi} = y_i(\mathbf{x}_{B_i})$, where $(z_{Bi}, \mathbf{x}_{Bi})$ is the choice that $d(i)$ would make under the constraint $z \geq B(\mathbf{x})$.*

Assumption CHOICE rules out for example optimization error, which could limit the decision-maker's ability to exactly manipulate values of $\mathbf{x}$ and hence y. It also takes for granted that counterfactual choices are unique, and rules out some kinds of extensive margin effects in which a decision-maker would not choose any value of Y at all under B_1 or B_0. Note that CHOICE here is slightly stronger than the version given in the main text in that it applies to all functions B, not just B_0, B_1 and B_k (this is useful for Theorem 2).

The central behavioral assumption that allows us to reason about the counterfactuals Y_0 and Y_1 is that decision-makers have convex preferences over $(c, \mathbf{x})$ and dislike costs z:

Assumption CONVEX (strictly convex preferences except at kink, decreasing in z). *For each i and any function $B(\mathbf{x})$, choice is $(z_{Bi}, \mathbf{x}_{Bi}) = argmax_{z,\mathbf{x}}\{u_i(z, \mathbf{x}) : z \geq B(\mathbf{x})\}$ where $u_i(z, \mathbf{x})$ is weakly decreasing in z and satisfies*

$$u_i(\theta z + (1-\theta)z^*, \theta\mathbf{x} + (1-\theta)\mathbf{x}^*) > \min\{u_i(z, \mathbf{x}), u_i(z^*, \mathbf{x}^*)\}$$

for any $\theta \in (0, 1)$ and points $(z, \mathbf{x})$,$(z^*, \mathbf{x}^*)$ such that $y_i(\mathbf{x}) \neq k$ and $y_i(\mathbf{x}^*) \neq k$.

Note: The function $u_i(\cdot)$ represents preferences over choice variables for unit i, but the preferences are those of the decision maker $d(i)$. I avoid more explicit notation like $u_{d(i),i}(\cdot)$ for brevity. In the overtime setting with firms choosing hours, $u_i(z, \mathbf{x})$ corresponds to the firm's profit function π as a function of the hours of a particular worker this week, and costs this week z for that worker.

Note: The second part of Assumption CONVEX is implied by strict quasi-concavity of the function $(z, \mathbf{x})$, corresponding to strictly convex preferences. However it also allows for decision-makers preferences to have "two peaks", provided that one of the peaks is located exactly at the kink. This is useful in cases in which the kink is located at a point that has particular value to decision-makers, such as firms setting weekly hours. For example, suppose that firms choose hours only $\mathbf{x} = h$, and have preferences of the form:

$$u_i(z, h) = af(h) + \phi \cdot \mathbb{1}(h = 40) - z \tag{B.7}$$

where $f(h)$ is strictly concave. This allows firms to have a behavioral "bias" towards 40 hours, or to extract extra profits when $h = 40$ exactly. Figure B.3 depicts an example of such preferences, given an arbitrary linear budget function $B(h)$. Note that if a mass of firms were to have preferences of this form, then it would be natural to expect bunching in the distributions of h_{0it} and h_{1it}, which I allow in Section 5.

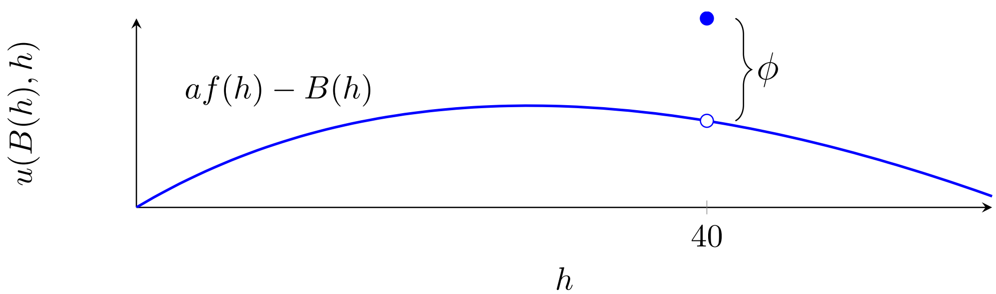

FIGURE B.3: An example of preferences that satisfy CONVEX but are not strictly convex, cf. Eq. (B.7).

Note: Some departures from CONVEX are allowable without compromising it's main implication for the bunching-design, which is given in Lemma 1 below. If B_0 and B_1 are linear in $\mathbf{x}$ and the constraint $z \geq B(\mathbf{x})$ can be assumed to bind (hold as an equality), then the assumption that u_i is decreasing in z from CONVEX can be dropped (see Assumption CONVEX* in Appendix C). If by contrast B_0 and B_1 were strictly (rather than weakly) convex, strict convexity of preferences could be replaced with weakly convex preferences along with an assumption that u_i are strictly decreasing in z (see Eq. (A.5) in the Proof of Lemma 1).

Note: The notation of Assumption CONVEX does not make explicit any dependence of the functions $u_i(\cdot)$ on the choices made for other observational units $i' \neq i$. When the functions $u_i(\cdot)$ are indeed invariant over such counterfactual choices, we have a version of the no-interference condition of the stable unit treatment values assumption (SUTVA). Maintaining SUTVA is not necessary to define treatment effects in the bunching design, provided that the variables y and z can be coherently defined at the individual unit i level (see Appendix H for details). Nevertheless, the interpretation of the treatment effects identified by the bunching design is most straightforward when SUTVA does hold. This assumption is standard in the bunching design.[40]

[40] I note that SUTVA issues like those addressed in Appendix H could also occur in canonical bunching designs: for example if spouses choose their labor supply jointly, the introduction of a tax kink may cause one spouse to increase labor supply while the other decreases theirs.

A weaker assumption than CONVEX that still has identifying power is simply that decision-makers' choices do not violate the weak axiom of revealed preference:

Assumption WARP (rationalizable choices). *Consider two budget functions B and B' and any unit i. If $d(i)$'s choice under B' is feasible under B, i.e. $z_{B'i} \geq B(\mathbf{x}_{B'i})$, then $(z_{Bi}, \mathbf{x}_{Bi}) = (z_{B'i}, \mathbf{x}_{B'i})$.*

I make the stronger assumption CONVEX for most of the identification results, but Assumption WARP still allows a version of many of them in which equalities become weak inequalities, indicating a degree of robustness with respect to departures from convexity (see Propositions 1 and 2 below). Note that the monotonicity assumption in CONVEX implies that choices will always satisfy $z = B(\mathbf{x})$, i.e. agents' choices will lay on their cost functions (despite Eq. B.6 being an inequality, indicating "free-disposal").

B.3 Examples from the general choice model in the overtime setting

To demonstrate the flexibility of the general choice model CONVEX, I below present some examples for the overtime setting. These examples are meant only to be illustrative, and each could apply to a different subset of units in the population. In these examples we continue to take the decision-maker for a given unit to be the firm employing that worker.[41]

Example 1: Substitution from bonus pay

Let the firm's choice vector be $\mathbf{x} = (h, b)'$, where $b \geq 0$ indicates a bonus (or other fringe benefit) paid to the worker. Firms may find it optimal to offer bonuses to improve worker satisfaction and reduce turnover. Suppose firm preferences are: $\pi(z, h, b) = f(h) + g(z + b - \nu(h)) - z - b$, where z continues to denote wage compensation this week, $z + b - \nu(h)$ is the worker's utility with $\nu(h)$ a convex disutility from labor h, and $g(\cdot)$ increasing and concave. In this model firms will choose the surplus maximizing choice of hours $h_m := \text{argmax}_h f(h) - \nu(h)$, provided that the corresponding optimal bonus is non-negative. Bonuses fully adjust to counteract overtime costs, and $h_0 = h_1 = h_m$.

Example 2: Off-the-clock hours and paid breaks

Suppose firms choose a pair $\mathbf{x} = (h, o)'$ with h hours worked and o hours worked "off-the-clock", such that $y(\mathbf{x}) = h - o$ are the hours for which the worker is ultimately paid. Evasion is harder the larger o is, which could be represented by firms facing a convex evasion cost $\phi(o)$, so that firm utility is $\pi(z, h, o) = f(h) - \phi(o) - z$.[42] This model can also include some firms voluntarily offering paid breaks by allowing o to be negative.

[41] Appendix C discusses a further example in which the firm and worker bargain over this week's hours. This model can attenuate the wage elasticity of chosen hours since overtime pay gives the parties opposing incentives.

[42] Note that the data observed in our sample are of hours of work $y(\mathbf{x})$ for which the worker is paid, when this differs from h. Appendix B describes how Equation 2 still holds, but for counterfactual values of hours *paid* $y = h - o$ rather than hours worked h. The bunching design lets us investigate treatment effects on paid hours, without observing off-the-clock hours or break time o.

Example 3: Complementaries between workers or weeks

Suppose the firm simultaneously chooses the hours $\mathbf{x} = (h, g)$ of two workers according to production that is isoelastic in a CES aggregate (g could also denote planned hours next week): $\pi(z, h, g) = a \cdot \left((\gamma h^\rho + g^\rho)^{1/\rho}\right)^{1+\frac{1}{\epsilon}} - z$ with γ a relative productivity shock. Let g^* denote the firm's optimal choice of hours for the second worker. Optimal h then maximizes $\pi(z, h, g^*)$ subject to $z = B(h)$, as if the firm faced a single-worker production function of $f(h) = a \cdot \left((\gamma h^\rho + g^{*\rho})^{1/\rho}\right)^{1+\frac{1}{\epsilon}}$. This function is more elastic than $a \cdot h^{1+\frac{1}{\epsilon}}$ provided that $\rho < 1 + 1/\epsilon$, attenuating the response to an increase in w implied by a given ϵ.[43] Section 4.4 discusses how complementaries affect the final evaluation of the FLSA.

B.4 Observables in the kink bunching design

Lemma 1 outlines the core consequence of Assumption CONVEX for the relationship between observed Y_i and the potential outcomes introduced in the last section:

Lemma 1 (realized choices as truncated potential outcomes). *Under Assumptions CONVEX and CHOICE, the outcome observed given the constraint* $z \geq max\{B_{0i}(\mathbf{x}), B_{1i}(\mathbf{x})\}$ *is:*

$$Y_i = \begin{cases} Y_{0i} & \text{if } Y_{0i} < k \\ k & \text{if } Y_{1i} \leq k \leq Y_{0i} \\ Y_{1i} & \text{if } Y_{1i} > k \end{cases}$$

Proof. See Appendix A. □

Lemma 1 says that the pair of counterfactual outcomes (Y_{0i}, Y_{1i}) is sufficient to pin down actual choice Y_i, which can be seen as an observation of one or the other potential outcome, or k, depending on how the potential outcomes relate to the kink point k.

Note that the "straddling" event $Y_{0i} \leq k \leq Y_{1i}$ from Lemma 1 can be written as $Y_{0i} \in [k, k + \Delta_i]$, where $\Delta_i := Y_{0i} - Y_{1i}$. Similarly, we can also write $Y_{1i} \leq k \leq Y_{0i}$ as $Y_i \in [k - \Delta_i, k]$. This forms the basic link between bunching and *treatment effects*.

Let $\mathcal{B} := P(Y_i = k)$ be the observable probability that the decision-maker chooses to locate exactly at $Y = k$. Proposition 1 gives the relationship between this bunching probability and treatment effects, which holds in a weakened form when CONVEX is replaced by WARP:

Proposition 1 (relation between bunching and Δ_i). *a) Under CONVEX and CHOICE:* $\mathcal{B} = P(Y_{0i} \in [k, k + \Delta_i])$*; b) under WARP and CHOICE:* $\mathcal{B} \leq P(Y_{0i} \in [k, k + \Delta_i])$.

Proof. See Appendix F. □

[43] This expression overstates the degree of attenuation somewhat, since h_1 and h_0 maximize $f(h)$ above for different values g^*, which leads to a larger gap between h_0 and h_1 compared with a fixed g^* by the Le Chatelier principle (Milgrom and Roberts, 1996). However h_1/h_0 still increases on net given $\rho < 1 + 1/\epsilon$.

Consider a random sample of observations of Y_i. Under i.i.d. sampling of Y_i, the distribution $F(y)$ of Y_i is identified.[44] Let $F_1(y) = P(Y_{0i} \leq y)$ be the distribution function of the random variable Y_0, and $F_1(y)$ the distribution function of Y_1. From Lemma 1 it follows immediately that $F_0(y) = F(y)$ for all $y < k$, and $F_1(y) = F(y)$ for $Y > k$. Thus observations of Y_i are also informative about the marginal distributions of Y_{0i} and Y_{1i}. Again, a weaker version of this also holds under WARP rather than CONVEX:

Proposition 2 (identification of truncated densities). *Suppose that F_0 and F_1 are continuously differentiable with derivatives f_0 and f_1, and that F admits a derivative function $f(y)$ for $y \neq k$. Under WARP and CHOICE: $f_0(y) \leq f(y)$ for $y < k$ and $f_0(k) \leq \lim_{y\uparrow k} f(y)$, while $f_1(y) \leq f(y)$ for $y > k$ and $f_1(k) \leq \lim_{y\downarrow k} f(y)$, with equalities under CONVEX.*

Proof. See Appendix F. □

As an example of how WARP alone (without CONVEX) can still be useful for identification, suppose that $\Delta_i = \Delta$ were known to be homogenous across units,[45] and $f_0(y)$ were constant across the interval $[k, k+\Delta]$, then by Propositions 1 and 2 we have that $\Delta \geq \mathcal{B}/f_0(k)$ under WARP and CHOICE.

B.5 Treatment effects in the bunching design

Proposition 1 establishes that bunching can be informative about features of the distribution of treatment effects Δ_i. This section discusses the interpretation of these treatment effects as well as some additional identification results omitted in the main text.

Unit i's treatment effect $\Delta_i := Y_{0i} - Y_{1i}$ can be thought of as the causal effect of a counterfactual change from the choice set under B_1 to the choice set under B_0. These treatment effects are "reduced form" in the sense that when the decision-maker has multiple margins of response $\mathbf{x}$ to the incentives introduced by the kink, these may be bundled together in the treatment effect Δ_i (Appendix J.5 discusses this in the setting of Best et al. 2015). This clarifies a limitation sometimes levied against the bunching design, while also revealing a perhaps under-appreciated strength. On the one hand, it is not always clear "which elasticity" is revealed by bunching at a kink, complicating efforts to identify a elasticity parameter having a firm structural interpretation (Einav et al., 2017).

On the other hand, the bunching design can be useful for ex-post policy evaluation and even forecasting effects of small policy changes (as described in Section 4.4), without committing to a tightly parameterized underlying model of choice. This provides a response to the note of caution by Einav et al. (2017), which points out that alternative structural models calibrated from the bunching-design can yield very different predictions about counterfactuals. By focusing on the counterfactuals Y_{0i} and Y_{1i}, we can specify a *particular* type of counterfactual question that can be answered robustly across a broad class of models.

[44] Note that in the overtime application sampling is actually at the firm level, which coincides with the level of decision-making units $d(i)$.

[45] One way to get homogenous treatment effects in levels in the overtime setting is to assume exponential production: $f(h) = \gamma(1 - e^{-h/\gamma})$ where $\gamma > 0$ and $h_{0it} - h_{1it} = \gamma \ln(1.5)$ for all units. The iso-elastic model instead gives homogeneous treatment effects for $log(h)$.

The "trick" of Lemma 1 is to express the observable data in terms of counterfactual choices, rather than of primitives of the utility function. The underlying utility function $u_i(z, \mathbf{x})$ is used only as an intermediate step in the logic, which only requires the nonparametric restrictions of convexity and monotonicity rather than knowing its functional form (or even what vector of choice variables **x** underly a given agent's observed value of y). This greatly increases the robustness of the method to potential misspecification of the underlying choice model.

Additional identification results for the bunching design:

While Theorem 1 of Section 4 develops the treatment effect identification result used to evaluate the FLSA, Appendix J presents some further identification results for the bunching design that are not used in this paper, which can be considered alternatives to Theorem 1. This includes re-expressing canonical results from the literature in the general framework of this section, including the linear interpolation approach of Saez (2010), the polynomial approach of Chetty et al. (2011) and a "small-kink" approximation appearing in Saez (2010) and Kleven (2016). Appendix J also discusses alternative shape constraints to bi-log-concavity, including monotonicity of densities. I also give there a result in which a lower bound to a certain local average treatment effect is identified under WARP, without requiring convexity of preferences.

The buncher ATE when Assumption RANK fails:

This section picks up from the discussion in Section 4.3, but continues with the notation of this Appendix. When RANK fails (and $p = 0$ for simplicity), the bounds from Theorem 1 are still valid under BLC of Y_0 and Y_1 for the following averaged quantile treatment effect:

$$\frac{1}{\mathcal{B}} \int_{F_0(k)}^{F_1(k)} \{Q_0(u) - Q_1(u)\} du = \mathbb{E}[Y_{0i} | Y_{0i} \in [k, k + \Delta_0^*]] - \mathbb{E}[Y_{1i} | Y_{1i} \in [k - \Delta_1^*, k]], \quad \text{(B.8)}$$

where $\Delta_0^* := Q_0(F_1(k)) - Q_1(F_1(k)) = Q_0(F_1(k)) - k$ and $\Delta_1^* := Q_0(F_0(k)) - Q_1(F_0(k)) = k - Q_1(F_0(k))$. Thus, Δ_0^* is the value such that $F_0(k + \Delta_0^*) = F_0(k) + \mathcal{B}$, and Δ_1^* is the value such that $F_1(k - \Delta_1^*) = F_1(k) - \mathcal{B}$. The averaged quantile treatment effect of Eq. (B.8) yields a lower bound on the buncher ATE, as described in Fig. B.4.

Assumption RANK and the sign of treatment effects:

Another important point regarding Assumption RANK is that it does not require $Y_{0i} \geq Y_{1i}$ for all units i. Figure B.5 shows an example in which $Y_1 = 2Y_0 - k$, so that $Y_1 < Y_0$ when $Y_0 < k$ and $Y_1 > h_0$ when $Y_0 > k$. For simplicity there is no bunching at the kink in this example, provided that Y_0 has a continuous marginal distribution around k. Note that from Lemma 1, we can write $\mathcal{B} = P(Y_1 \leq k, Y_1 \leq Y_0) - P(Y_0 \leq k, Y_1 \leq h_0)$,[46] which when combined with $\mathcal{B} = P(Y_1 \leq k) - P(Y_0 \leq k)$ (c.f. Eq. (7) in the main text) in

[46] I thank an anonymous referee for pointing out this alternative expression for $\mathcal{B}$, which holds provided that (Y_0, Y_1) is absolutely continuous.

Signing the bias when RANK fails

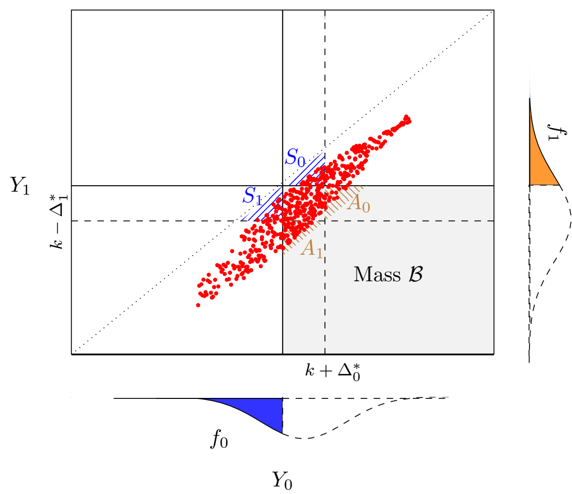

FIGURE B.4: When Assumption RANK fails, the average $\mathbb{E}[Y_{0i}|Y_{0i} \in [k, k+\Delta_0^*]]$ will include the mass in the region S_0, who are not bunchers (NE lines) but will be missing the mass in the region A_0 (NW lines) who are. This causes an under-estimate of the desired quantity $\mathbb{E}[Y_{0i}|Y_{1i} \leq k \leq Y_{0i}]$. Similarly, $\mathbb{E}[Y_{1i}|Y_{1i} \in [k-\Delta_1^*, k]]$ will include the mass in the region S_1, who are not bunchers but will be missing the mass in A_1, who are. This causes an over-estimate of the desired quantity $\mathbb{E}[Y_{1i}|Y_{1i} \leq k \leq Y_{0i}]$.

An (Y_0, Y_1) distribution with $P(Y_1 \leq Y_0) < 1$

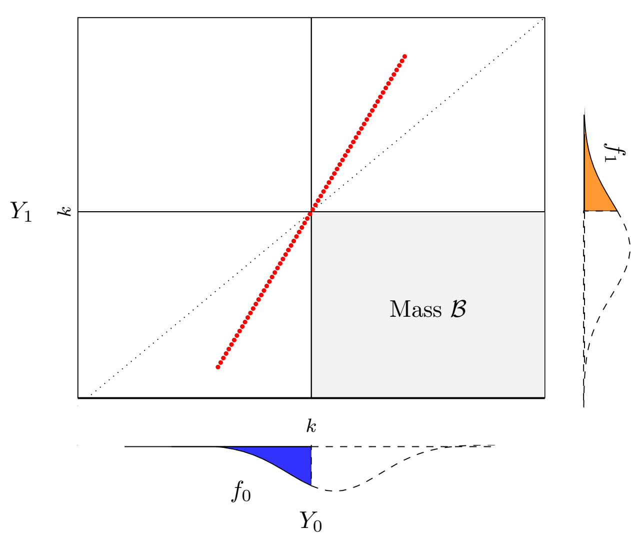

FIGURE B.5: A distribution of (Y_0, Y_1) such that $P(Y_0 > Y_1) > 0$ and $P(Y_1 > Y_0) > 0$. The thin black dots reflect the 45 degree line. Note that while there is mass on either side of the 45 degree line, there is no mass in the NW quadrant of the figure, which would violate Assumption CONVEX.

turn implies that

$$P(Y_0 \leq k, Y_1 > Y_0) = P(Y_1 \leq k, Y_1 > Y_0) \tag{B.9}$$

This implication is certainly satisfied if $Y_0 \geq Y_1$ with probability one, since then both sides are equal to zero. This is the case in for example the isoelastic model, given a positive elasticity. More generally however, Eq. (B.9) simply says that the mass above the 45 degree line in the western half of Figure B.5 is equal to the mass above the 45 degree line in the southern half of it. Any joint distribution for which there is no mass in the NW quadrant—consistent with assumption CONVEX—will satisfy (B.9), for example the case depicted in Figure B.5.

B.6 Policy changes in the bunching-design

This section presents the logic establishing Theorem 2 in the main text regarding the effects of changes to the policy generating a kink. Consider a bunching design setting in which the cost functions B_0 and B_1 can be viewed as members of family $B_i(\mathbf{x}; \rho, k)$ parameterized by a continuum of scalars ρ and k, where $B_{0i}(\mathbf{x}) = B_i(\mathbf{x}; \rho_0, k^*)$ and $B_{1i}(\mathbf{x}) = B_i(\mathbf{x}; \rho_1, k^*)$ for some $\rho_1 > \rho_0$ and value k^* of k. In the overtime setting ρ represents a wage-scaling factor, with $\rho = 1$ for straight-time and $\rho = 1.5$ for overtime:

$$B_i(y; \rho, k) = \rho w_i y - k w_i(\rho - 1) \tag{B.10}$$

where work hours y may continue to be a function $y(\mathbf{x})$ of a vector of choice variables to the firm. In this example, k controls the size of the lump-sum subsidy $kw_i(\rho - 1)$ that keeps $B_i(k; \rho, k)$ invariant as ρ is changed.

In the general setting, assume that ρ takes values in a convex subset of $\mathbb{R}$ containing ρ_0 and ρ_1, and that for any k and $\rho' > \rho$ the cost functions $B_i(\mathbf{x}; \rho, k)$ and $B_i(\mathbf{x}; \rho', k)$ satisfy the conditions of the bunching design framework from Section 4 (with the function $y_i(\mathbf{x})$ fixed across all ρ and k). That is, $B_i(\mathbf{x}; \rho', k) > B_i(\mathbf{x}; \rho, k)$ iff $y_i(\mathbf{x}) > k$ with equality when $y_i(\mathbf{x}) = k$, the functions $B_i(\cdot; \rho, k)$ are weakly convex and continuous, and $y_i(\cdot)$ is continuous. It is readily verified that Equation (B.10) satisfies these requirements with $y_i(h) = h$.[47]

For any value of ρ, let $Y_i(\rho, k)$ be agent i's realized value of $y_i(\mathbf{x})$ when a choice of $(z, \mathbf{x})$ is made under the constraint $z \geq B_i(\mathbf{x}; \rho, k)$. A natural restriction in the overtime setting that is that the function $Y_i(\rho, k)$ does not depend on k, and some of the results below will require this. A sufficient condition for $Y_i(\rho, k) = Y_i(\rho)$ is a family of cost functions that are linearly separable in k, as we have in the overtime setting with Equation (B.10), along with quasi-linearity of preferences. Quasilinearity of preferences is a property of profit-maximizing firms when z represents a cost, and is thus a natural assumption in the overtime setting.

Assumption SEPARABLE (invariance of potential outcomes with respect to k). *For all i, ρ and k, $B_i(\mathbf{x}; \rho, k)$ is additively separable between k and $\mathbf{x}$ (e.g. $b_i(\mathbf{x}, \rho) + \phi_i(\rho, k)$ for some functions b_i and ϕ_i),*

[47] As an alternative example, I construct in Appendix J.5 functions $B_i(\mathbf{x}; \rho, k)$ for the bunching design setting from Best et al. (2015). In that case, ρ parameterizes a smooth transition between an output and a profit tax, where k enters into the rate applied to the tax base for that value of ρ.

and for all i $u_i(z, \mathbf{x})$ can be chosen to be additively separable and linear in z.

Additive separability of $B_i(\mathbf{x}; \rho, k)$ in k may be context specific: in the example from Best et al. (2015) described in Appendix J.5, quasi-linearity of preferences is not sufficient since the cost functions are not additively separable in k. To maintain clarity of exposition, I will keep k implicit in $Y_i(\rho)$ throughout the foregoing discussion, but the proofs make it clear when SEPARABLE is being used.

Below I state two intermediate results that allow us to derive expressions for the effects of marginal changes to ρ_1 or k on hours. Lemma 2 generalizes an existing result from Blomquist et al. (2015), and makes use of a regularity condition I introduce in the proof as Assumption SMOOTH.[48]

Counterfactual bunchers $K_i^* = 1$ are assumed to stay at some fixed value k^* (40 in the overtime setting), regardless of ρ and k. Let $p(k) = p \cdot \mathbb{1}(k = k^*)$ denote the possible counterfactual mass at the kink as a function of k. Let $f_\rho(y)$ be the density of $Y_i(\rho)$, which exists by SMOOTH and is defined for $y = k^*$ as a limit (see proof).

Lemma 2 (bunching expressed in terms of marginal responsiveness). *Assume CHOICE, SMOOTH and WARP. Then:*

$$\mathcal{B} - p(k) \leq \int_{\rho_0}^{\rho_1} f_\rho(k)\mathbb{E}\left[-\frac{dY_i(\rho)}{d\rho}\middle| Y_i(\rho) = k\right] d\rho$$

with equality under CONVEX.

Proof. See Appendix F. □

The main tool in establishing Lemma 2 is to relate the integrand in the above to the rate at which kink-induced bunching goes away as the "size" of the kink goes to zero.

Lemma SMALL (small kink limit). *Assume CHOICE*, WARP, and SMOOTH. Then:*

$$\lim_{\rho' \downarrow \rho} \frac{P(Y_i(\rho') \leq k \leq Y_i(\rho)) - p(k)}{\rho' - \rho} = -f_\rho(k)\mathbb{E}\left[\frac{dY_i(\rho)}{d\rho}\middle| Y_i(\rho) = k\right]$$

Proof. See Appendix F. □

Note that the quantity $P(Y_i(\rho') \leq k \leq Y_i(\rho)) - p(k)$ is an upper bound on the bunching that would occur due to a kink between budget functions $B_i(\mathbf{x}; \rho, k)$ and $B_i(\mathbf{x}; \rho', k)$ (under WARP, with equality under CONVEX). As a result, Lemma SMALL shows that the uniform density approximation that has appeared in Saez (2010) and Kleven (2016) (stated in Appendix Proposition J.4) for "small" kinks becomes exact in the limit that the two cost functions approach one another. The small kink approximation says that $\mathcal{B} \approx f_\rho(k) \cdot \mathbb{E}[Y_i(\rho) - Y_i(\rho')]$, where note that treatement effects can be writtens:

$$Y_i(\rho) - Y_i(\rho') = \frac{dY_i(\rho)}{d\rho}(\rho' - \rho) + O((\rho' - \rho)^2)$$

[48]Blomquist et al. (2021) derive the special case of Lemma 2 with convex preferences over a scalar choice variable and $p = 0$, in the context of labor supply under piecewise linear taxation. I establish it here for the general bunching design model where in particular, the $Y_i(\rho)$ may depend on an underlying vector $\mathbf{x}$ which are not observed by the econometrician. I also use different regularity conditions.

By Lemma 2, we can also see that the RHS in Lemma SMALL evaluated at $\rho = \rho_1$ is equal to the derivative of bunching as ρ_1 is increased, under CONVEX.

Lemma 2 is useful for identification results regarding changes to k when it is combined with a result from Kasy (2022), which considers how the distribution of a generic outcome variable changes as heterogeneous units flow to different values of that variable in response to marginal policy changes.

Lemma 3 (continuous flows under a small change to ρ). *Under SMOOTH:*

$$\partial_\rho f_\rho(y) = \partial_y \left\{ f_\rho(y)\mathbb{E}\left[-\frac{dY_i(\rho)}{d\rho} \middle| Y_i(\rho) = y, K_i^* = 0 \right] \right\}$$

Proof. See Kasy (2022). □

The intuition behind Lemma 3 comes from the physical dynamics of fluids. When ρ changes, a mass of units will "flow" out of a small neighborhood around any y, and this mass is proportional to the density at y and to the average rate at which units move in response to the change. When the magnitude of this net flow varies with y, the change to ρ will lead to a change in the density there.

With ρ_0 fixed at some value, let us index observed Y_i and bunching $\mathcal{B}$ with the superscript $[k, \rho_1]$ when they occur in a kinked policy environment with cost functions $B_i(\cdot; \rho_0, k)$ and $B_i(\cdot; \rho_1, k)$. Lemmas 2 and 3 together imply Theorem 2 (see Appendix A for proof).*Note:* Assumption SEPARABLE is only necessary for Items 1-2 in Theorem 2, Item 3 holds without it and with $\frac{\partial Y_i(\rho,k)}{\partial \rho}$ replacing $\frac{dY_i(\rho)}{d\rho}$.

C Incorporating workers that set their own hours

This section considers the robustness of the empirical strategy from Section 4 to a case where some workers are able to choose their own hours. In this case, a simple extension of the model leads to the bounds on the buncher ATE remaining valid, but it is only directly informative about the effects of the FLSA among workers who have their hours chosen by the firm. In this section I follow the notation from the main text where h_{it} indicate the hours of worker i in week t.

Suppose that some workers are able to choose their hours each week without restriction ("worker-choosers"), and that for the remaining workers ("firm-choosers") their employers set their hours. In general we can allow who chooses hours for a given worker to depend on the period, so let $W_{it} = 1$ indicate that i is a worker-chooser in period t. Additionally, we continue to allow conterfactual bunchers for whom counterfactual hours satisfy $h_{0it} = h_{1it} = 40$, regardless of who chooses them. I replace Assumption CONVEX from Section 4 to allow agents to *either* dislike pay (firm-choosers), or like pay (worker-choosers):

Assumption CONVEX* (convex preferences, monotonic in either direction). *For each i, t and function $B(\mathbf{x})$, choice is $(z_{Bi}, \mathbf{x}_{Bi}) = argmax_{z,\mathbf{x}}\{u_i(z, \mathbf{x}) : z \geq B(\mathbf{x})\}$ where $u_i(z, \mathbf{x})$ is strictly increasing in z, if $W_{it} = 1$, strictly decreasing in z, if $W_{it} = 0$, and satisfies $u_i(\theta z + (1-\theta)z^*, \theta\mathbf{x} + (1-\theta)\mathbf{x}^*) > \min\{u_i(z, \mathbf{x}), u_i(z^*, \mathbf{x}^*)\}$ for any $\theta \in (0, 1)$ and points $(z, \mathbf{x})$, $(z^*, \mathbf{x}^*)$ such that $y_i(\mathbf{x}) \neq k$ and $y_i(\mathbf{x}^*) \neq k$.*

For generality, I here use weaker notion of convexity of preferences from Assumption CONVEX in Appendix B. It is implied by strict quasi-concavity of $u_i(z, \mathbf{x})$.

Note: This setup is general enough to also allow a stylized bargaining-inspired model in which choices maximize a weighted sum of quasilinear worker and firm utilities. For example, suppose that for any pay schedule $B(h)$:

$$h = \underset{h}{\text{argmax}}\, \beta\,(f(h) - z) + (1-\beta)(z - \nu(h)) \quad with \quad z = B(h) \tag{C.11}$$

where $f(h) - z$ is firm profits with concave production f, $z - \nu(h)$ is worker utility with a convex disutility of labor $\nu(h)$, and $\beta \in [0, 1]$ governs the weight of each party in the negotiation (this corresponds to Nash bargaining in which outside options are strictly inferior to all h for both parties, and utility is log-linear in z). Rearranging the maximand of Equation (C.11) as $(1-2\beta)z + \{\beta f(h) - (1-\beta)\nu(h)\}$, we can observe that this setting delivers outcomes as-if chosen by a single agent with quasi-concave preferences, as $\beta f(h) - (1-\beta)\nu(h)$ is concave. For Assumption CONVEX from Section 4 to hold with the assumed direction of monotonicity in pay z, we would require that $\beta > 1/2$ for all worker-firm pairs: informally, that firms have more say than workers do in determining hours. However the more general CONVEX* holds regardless of the distribution of β over worker-firm pairs. If $\beta_{it} < 1/2$, paycheck it will look like a worker-chooser, and if $\beta_{it} > 1/2$ paycheck it will look like a firm-chooser.

In the generalized model of CONVEX*, bunching is prima-facie evidence that firm-choosers exist, because there is no prediction of bunching among worker-choosers provided that potential outcomes are continuously distributed. By contrast, k is a "hole" in the worker-chooser hours distribution: intuitively, if a worker is willing to work 40 hours then they will also find it worthwhile to work more, given the sudden increase in their wage. Indeed under regularity conditions all of the data local to 40 are from firm-choosers (or counterfactual bunchers). To make this claim precise, assume that for worker-choosers, hours are the only margin of response (i.e. their utility depends on $\mathbf{x}$ only thought $y(\mathbf{x})$), and let $IC_{0it}(y)$ and $IC_{1it}(y)$ be the worker's indifference curves passing through h_{0it} and h_{1it}, respectively. I assume these indifference curves are twice Lipschitz differentiable, and let $M_{it} := \sup_y \max\{|IC''_{0it}(y)|, |IC''_{1it}(y)|\}$, where IC'' indicates second derivatives.

Proposition 3. *Suppose that the joint distribution of h_{0it} and h_{1it} admits a continuous density conditional on $K^*_{it} = 0$, and that for any worker-chooser IC_{0it} and IC_{1it} are differentiable with M_{it}/w_{it} having bounded support. Then, under CHOICE and CONVEX*:*

- $P(h_{it} = k \text{ and } K^*_{it} = 0) = P(h_{1it} \le k \le h_{0it} \text{ and } K^*_{it} = 0 \text{ and } W_{it} = 0)$
- $\lim_{h\uparrow k} f(h) = P(W_{it} = 0) \lim_{h\uparrow k} f_{0|W=0}(h)$
- $\lim_{h\downarrow k} f(h) = P(W_{it} = 0) \lim_{h\downarrow k} f_{1|W=0}(h)$

Proof. See Appendix K. □

The first bullet of Proposition 3 says that all active bunchers are also firm-choosers, and have potential outcomes that straddle the kink. The second and third bullets state that the density of the data as hours approach 40 from either direction is composed only of worker-choosers. This result on density limits requires the stated regularity condition on M_{it}/w_{it}, which prevents worker indifference curves from becoming too close to themselves featuring a kink (plus a requirement that straight-time wages w_{it} be bounded away from zero).

Given the first item in Proposition 3, the buncher ATE introduced in Section 4 only includes firm-choosers:

$$\mathbb{E}[h_{0it} - h_{1it}|h_{it} - 40, K^*_{it} = 0] = \mathbb{E}[h_{0it} - h_{1it}|h_{it} - 40, K^*_{it} = 0, W_{it} = 0]$$

Accordingly, I assume rank invariance among the firm-chooser population only:

Assumption RANK* (near rank invariance and counterfactual bunchers). *The following are true:*

1. $P(h_{0it} = k) = P(h_{1it} = k) = p$
2. $Y_{it} = k$ *iff* ($h_{0it} \in [k, k + \Delta^*_0]$ *and* $W_{it} = 0$) *iff* ($h_{1it} \in [k - \Delta^*_1, k]$ *and* $W_{it} = 0$), *for some* Δ^*_0, Δ^*_1

where p continues to denote $P(K^*_{it} = 1)$.

We may now state a version of Theorem 2 that conditions all quantities on $W = 0$, provided that we assume bi-log concavity of h_0 and h_1 conditional on $W = 0$ and $K = 0$.

Theorem 1* (bi-log-concavity bounds on the buncher ATE, with worker-choosers). *Assume CHOICE, CONVEX* and RANK* hold. If both h_{0it} and h_{1it} are bi-log concave conditional on the event ($W_{it} = 0$ and $K^*_{it} = 0$), then:*

$$\mathbb{E}[h_{0it} - h_{1it}|h_{it} = k, K^*_{it} = 0] \in \left[\Delta^L_k, \Delta^U_k\right]$$

where

$$\Delta^L_k = g(F_{0|W=0,K^*=0}(k), f_{0|W=0,K^*=0}(k), \mathcal{B}^*) + g(1 - F_{1|W=0,K^*=0}(k), f_{1|W=0,K^*=0}(k), \mathcal{B}^*)$$

and

$$\Delta^U_k = -g(1 - F_{0|W=0,K^*=0}(k), f_{0|W=0,K^*=0}(k), -\mathcal{B}^*) - g(F_{1|W=0,K^*=0}(k), f_{1|W=0,K^*=0}(k), -\mathcal{B}^*)$$

where $\mathcal{B}^* = P(h_{it} = k|W_{it} = 0, K^*_{it} = 0)$.The bounds are sharp.

Proof. See Appendix K. □

Identification with worker-choosers

Theorem 1* does not immediately yield identification of the buncher-ATE bounds Δ^L_k and Δ^U_k, as we need to estimate each of the arguments to the function g. As shown in the proof of Theorem 1*, the bounds can be rewritten in terms of p, the identified quantities $\mathcal{B}$, $P(W_{it} = 0)\lim_{y\uparrow k} f_{0|W=0}(y)$ and $P(W_{it} = 0)\lim_{y\uparrow k} f_{1|W=0}(y)$, and two unidentified probabilities: $P(h_{0it} < k \text{ and } h_{it} = h_{0it} \text{ and } W_{it} = 1)$ and $P(Y_{1it} > k \text{ and } h_{it} = h_{1it} \text{ and } W_{it} = 1)$.

To illustrate the unidentified quantities, Figure C.6 depicts an example of a joint distribution of (h_0, h_1) that includes worker-choosers and satisfies Assumption RANK*. The x-axis is h_0, and the y-axis is h_1, with the solid lines indicating 40 hours and the dotted diagonal line depicting $h_1 = h_0$. The dots show a hypothetical joint-distribution of the potential outcomes, with the (red) dots south of the 45-degree line representing firm-choosers, and the (blue and orange) points above representing worker-choosers. Blue x's indicate worker-choosers who choose their value of h_0, while orange circles indicate worker-choosers who choose their value of h_1. The red dot at $(40, 40)$ represents a mass of counterfactual bunchers.

A distribution with worker-choosers

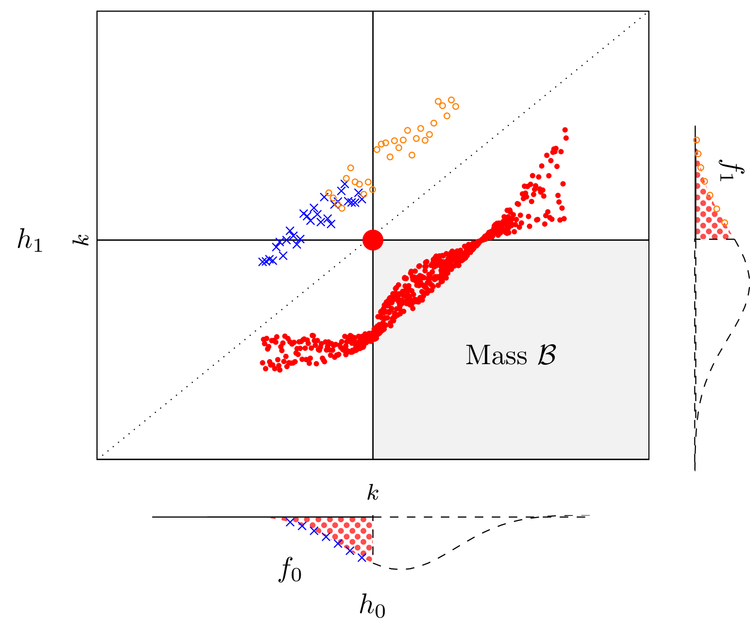

FIGURE C.6: The joint distribution of (h_{0it}, h_{1it}), for a distribution including worker-choosers and satisfying assumption RANK*, cf. Figure 5. See text for description.

Observed to the the econometrician is the point mass at 40 as well as the truncated marginal distributions depicted at the bottom and the right of the figure, respectively. The observable $P(h_{it} \leq h)$ for $h < 40$ doesn't exactly identify $P(h_{0it} \leq h)$ because some blue x's are missing: these are worker-choosers for whom $h_1 > 40 > h_0$ and choose to work overtime at their h_1 value. Thus they show up in the data at $h > 40$ even thoug they have $h_0 < 40$. Similarly, some orange circles do not appear in the observations above 40: these are worker-choosers for whom $h_1 > 40 > h_0$ and choose to work their h_0 value, not working overtime. The probabilities $P(h_{it} < 40 \text{ and } W_{it} = 0)$ and $P(h_{it} > 40 \text{ and } W_{it} = 0)$ can thus only be estimated with some error, with the size of the error depending on the mass of worker-choosers in the northwest quadrant of Figure C.6. However, in practice this has little impact on the results, as the bounds Δ_k^L and Δ_k^U are not very sensitive to the values of the CDF inputs $F_{0|W=0,K^*=0}(k)$ and $F_{1|W=0,K^*=0}(k)$. The bounds mostly depend on the density estimates and the size of the bunching mass, given their empirical values. Thus Theorem 1* still partially identifies the buncher LATE among firm-choosers, to a good approximation.

However, a further caveat of Theorem 1* is worth mentioning. An evaluation of the FLSA would ideally account for worker-choosers (who are working longer hours as a result of the policy) when averaging treatment effects. However, the proportion of worker-choosers and the size of their hours increases are not identified. Using the buncher ATE to estimate the overall ex-post effect of the FLSA—as described in Section 4.4—may overstate its overall average net hours reduction. However, the survey evidence mentioned in Section 2 suggests that the set of worker-choosers is relatively small, mitigating this concern.

D Motivating the bi-log-concavity assumption

D.1 As extrapolation of an upper bound

The polynomial estimation approaches of Saez (2010) and Chetty et al. (2011) can be thought of as extrapolating the exact curve of a polynomial fit to the observable distribution to point identify ϵ, in the iso-elastic model. An alternative is to extrapolate features of the observed density without extrapolating it's exact functional form. For example, Bertanha et al. (2023) proposes computing the maximum derivative of the density of $\ln(h)$ for $h \neq k$ and assuming that this Lipshitz bound also holds across the missing region depicted in 4. Similarly, Blomquist et al. (2021) propose bounding the *level* of the density of $\ln(h)$ to be within the convex hull of the left and right limits of the density of $\ln(h)$, expanded by a specified constant σ. Taking $\sigma = 1$ nests the non-parametric shape constraint of imposing monotonicity of the density of $\ln(h)$.

The logic of verifying BLC of h_0 to the left of the kink to motivate BLC of h_0 across the unobserved region $[k, k + \Delta_0^*]$(and analagously for h_1, looking to the right of the kink) as demonstrated in Figure 1, can be described in similar terms. Focusing on the case with no counterfactual bunchers ($p = 0$) for simplicity, Theorem 1 assumes that $\ln F_0(h)$ and $\ln(1 - F_0(h))$ are both concave on the interval $[k, k + \Delta_0^*]$. Assuming this CDF is twice differentiable, this is equivalent to:

$$\sup \left\{ \max\{d^2/dh^2 \ln F_0(h), d^2/dh^2 \ln(1 - F_0(h))\} : h \in [k, k + \Delta_0^*] \right\} \leq 0 \tag{D.12}$$

If h_0 is in fact BLC to the left of the kink, then

$$\sup \left\{ \max\{d^2/dh^2 \ln F_0(h), d^2/dh^2 \ln(1 - F_0(h))\} : h < k \right\} \leq 0$$

and hence a sufficient condition for (D.12) is that

$$\begin{aligned} &\sup \left\{ \max\{d^2/dh^2 \ln F_0(h), d^2/dh^2 \ln(1 - F_0(h))\} : h \in [k, k + \Delta_0^*] \right\} \\ &\qquad \leq sup \left\{ \max\{d^2/dh^2 \ln F_0(h), d^2/dh^2 \ln(1 - F_0(h))\} : h < k \right\} \end{aligned}$$

Similarly, if BLC of h_1 is verified for values to the right of the kink, then a sufficient condition for the assumption required by Theorem 1 is that

$$\sup\left\{\max\{d^2/dh^2\ln F_1(h), d^2/dh^2\ln(1-F_1(h))\} : h \in [k-\Delta_1^*, k]\right\}$$
$$\leq sup\left\{\max\{d^2/dh^2\ln F_1(h), d^2/dh^2\ln(1-F_1(h))\} : h > k\right\}$$

In this way, the BLC assumptions made by Theorem 1 can be thought of as extrapolating the extreme value of a property (or properties) of the distribution of F_d from a region in which that property is observed, to an unobserved region corresponding to the bunchers. While Blomquist et al. (2021) extrapolates the max/min levels of the density right next to the kink, and Bertanha et al. (2023) the magnitude of it's derivative across all point away from the kink, a sufficient condition for my result in Theorem 1 is to extrapolate the maximum of the second derivative of both $\ln F_d$ and $\ln(1-F_d)$, for each of $d \in \{0, 1\}$.

D.2 Bi-log-concavity in terms of hazard functions

The partial identification result of Theorem 1 hinges on the assumption that the distribution of counterfactual hours h_{0it} and h_{1it} (among units it that are not counterfactual bunchers) are both bi-log-concave (BLC). In this section, I decompose this assumption into two parts and describe how each part arises naturally as a property of the distribution of working hours.

Consider a random variable with CDF $F(h)$ admitting of a density $f(h)$. BLC is equivalent to the following:

1. the hazard rate function $f(h)/(1-F(h))$ is (weakly) increasing in h
2. the reverse hazard rate function $f(h)/F(h)$ is (weakly) decreasing in h

These can be derived by observing that the derivative of $\log F(h)$ is $f(h)/F(h)$ and that $-f(h)/(1-F(h))$ is the derivative of $\log(1-F(h))$—see Dümbgen et al. (2017) for details. Note that while BLC is introduced in Section 4 as a property applying to the whole support of a random variable, the Theorem 1 bounds on the buncher ATE only in face require these properties to hold for h_0 on the interval $[k, k+\Delta_0^*]$ and on the interval $[k-\Delta_1^*, k]$ for h_1 (as described in the proof).

The property of an increasing hazard rate arises in reliability theory, which often models the aging properties of a system over time. Consider a very simple model in which workers continue working until they "fail" at some stochastic number H of work hours. The hazard rate $f(h)/(1-F(h))$ then captures the probability that the worker fails after h hours given that they have not failed yet ($H > h$). That the instantaneous probability of failure for a system of age h would be increasing in h is a natural notion of *wear* (e.g. of a machine), and is often referred to as the *increasing failure rate* or IFR property (Barlow et al., 1996). While we might view the above as a model of worker fatigue that manifests as a dichotomous notion of "failure", the next section shows how IFR also emerges in a more realistic model in which worker productivity declines gradually and stochastically over time.

The second aspect of bi-log-concavity is that the reversed hazard rate $f(h)/F(h)$ is weakly decreasing in h, referred to as a *decreasing reverse hazard rate* or DRHR. Block et al. (1998) show that any non-negative random variable must be DRHR at least somewhere in its support. The BLC assumption made by Theorem 1 requires something stronger: that counterfactual hours be DRHR (in addition to being IFR) across a *particular* region near the kink. Like IFR, one can characterize DRHR in terms of failure times: DRHR holds iff the time $h - H$ that has elapsed by some moment h since failure at H (given that $H \leq h$) is increasing in h in the sense of stochastic dominance (Gupta and Nanda, 2001). This is an intuitive property, but could fail to hold if the density of H increases too rapidly at some h. The model in the next section provides primitive conditions that rule this out.[49]

While the above considerations may lend plausibility to the IFR and DRHR properties of counterfactual working hours h_0 and h_1 by giving them intuitive interpretations, they fall short of providing explicit sufficient "economic" conditions for them both. The next section does so, by modeling the working hours as chosen optimally according to a worker productivity that is generated by an underlying process with an assumed Markovian structure.

D.3 A model of hours with stochastic shocks to productivity

Recall Equation (3) of Section 4, which provides intuition for firms' optimal choices of hours in terms of the marginal product of an hour of labor from unit it. In a model in which firms maximize the net revenue $\pi_{it}(h) = f_{it}(h) - B_{kit}(h)$ from worker i in week t, where $f_{it}(\cdot)$ is a revenue production function with respect to hours, then counterfactual choices can be written as $h_{0it} = MPH_{it}^{-1}(w_{it})$ and $h_{1it} = MPH_{it}^{-1}(1.5w_{it})$. Here $MPH_{it}(h) = \frac{d}{dh}f_{it}(h)$ can be thought of as worker i's instantaneous hourly productivity at hour h within the week, and w_{it} is their straight wage.

Within a model of this form, we can motivate BLC of h_{1it} and h_{0it} among a set of ex-ante identical workers that experience different realizations of a common stochastic process generating the function f. Assume these workers share a straight wage $w_{it} = w$, and are not counterfactual bunchers in the language of Section 4.3 (thus conditioning on $K_{it}^* = 0$ will be kept implicit). Consider a single fixed week t which I suppress for now in the notation.

All workers have a common productivity $MPC_i(0) = p_0$ when they are "fresh" and have not yet worked any hours this week. At each moment in continuous time, a worker's hourly productivity either stays the same or drops by a discrete amount. Let $\{p_j\}_{j=0,1,...}$ be a decreasing sequence that denotes hourly productivity after j productivity drops. This function of j is assumed common to all workers i.

We'll see that bi-log-concavity of $MPH_i^{-1}(w)$ for any w then follows when the timing of these drops has a simple Markovian structure. In particular, assume that the probability of j increasing by one in a small timespan around h hours depends only on j and is independent of h and the past trajectory of productivity. This is a reasonable assumption if what matters for the future evolution of worker fatigue is that worker's current level of fatigue, rather than how many hours they have been working so far *per-se*.

[49] Another way to motivate the DRHR property of work hours is to note that if there are T total hours in a week, the number of non-work hours L is $T - H$, and to observe that H is DRHR if and only if L is IFR. Thus DRHR of H can be interpreted as saying that the failure rate of "leisure" is increasing: the probability that L lies in a infinitesimal neighborhood of ℓ, given that $L > \ell$, is an increasing function of ℓ.

Since MPH_i is weakly decreasing in h for all i, we can define an inverse MPH function as $MPH_i^{-1}(w) = \inf\{h : MPH_i(h) \leq w\}$. The RHS of this expression is referred to as a *first-passage time*, a random variable whose distribution is often of interest in the reliability theory literature. We can understand the first passage time $MPH_i^{-1}(w)$ as the first time h that a worker's fatigue j has accumulated to j^*, the smallest j such that $p_j \leq w$. (i.e. it is no longer profitible for the worker to continue working at wage w).

Kijima (1998) shows that if a continuous-time Markov chain on the positive integers can only increase or decrease by one unit at a time, the distribution of first passage times from zero to any given level j^* satisfies both IFR and DRHR, and is hence BLC.[50] Recall that roughly speaking, these properties mean that the density of first passage times can neither rise nor fall too abruptly at any one point h. To get some intuition for this result, let $j_i(h)$ be the number of productivity drops worker i has received in the first h hours of work. Then the (time homogeneous) Markov property implies that transitions into state j^* satisfy:

$$P(j_i(h+s) = j^* | j_i(h) = j^* - 1) = s \cdot \lambda_{j^*-1} + o(s)$$

for some set of rate parameters λ_{j^*}, and any h. Since $j_i(h)$ must first pass through $j^* - 1$ to arrive at j^*, this implies that the density of first passage times evaluated at h is equal to $P(j_i(h) = j^* - 1) \cdot \lambda_{j^*-1}$. Since the factor λ_{j^*-1} is common to all h, the density of first-passage times can only have a "spike" or a "hole" at h if the function $P(j_i(h) = j^* - 1)$ has a corresponding spike or hole at h. But the Markov structure constrains the form of $P(j_i(h) = j^* - 1)$ in a way that rules this out (see e.g. Taylor and Karlin (1994)).

While results of Kijima (1998) show that the first passage times in this model satisfy both components of BLC, Keilson (1971) calculates the distribution of first-passage times explicitly. When a continuous-time Markov chain on the integers cannot increase or decrease by more than one unit at a time, it is referred to as a "birth-death" processes. Keilson (1971) shows that first passage times to j^* for birth-death processes are distributed as a convolution of j^* exponential densities. The resulting density is log-concave, a special case of BLC.[51]

Now let us bring back the index t for the week of paycheck unit it. In the spirit of Section 4 one might view the above as a model of *scheduled* hours that the firm chooses given worker at the beginning of week t, if the firm is aware of that worker's realization of the productivity process. This might be reasonable if the workers' production function f_{it} is the same each week t, and the firm is able to quickly learn it upon hiring the worker. Alternatively, we can view the above model as describing shocks to productivity that are not yet

[50]Kijima (1998) predates the introduction of the term *bi-log-concavity* by Dümbgen et al. (2017), so he does not use this term.

[51]Technically, the results of both Kijima (1998) and Keilson (1971) require there to be a non-zero probability of the fatigue process decreasing by one unit from any state j, with some transition rate $\mu_j > 0$. The above model is instead a "pure-birth" process in which fatigue only ever increases. We can obtain the desired result with $\mu_j = 0$ however by considering a sequence of Markov processes characterized by downward transition rates $\mu_j^{(n)} > 0$ where $\lim_{n \to \infty} \mu_j^{(n)} = 0$. Since CDFs of corresponding first passage times $h^{(n)}$ are pointwise continuous functions of $\mu_j^{(n)}$, and BLC is preserved under convergence in distribution (Saumard, 2019), it follows that the distribution of hours in the pure-birth model is BLC. A similar construction can be used to accommodate productivity with continuous rather than discrete support, viewing the continuous diffusion process for productivity as the limit of a sequence of birth-death processes. See Keilson (1971) for details.

revealed to firms at the beginning of the week. At each moment *of each week*, workers receive a possible shock to productivity and the firm decides whether to keep the worker engaged in labor or withdraw them for the wee (after withdrawal productivity resets to p_0 for the beginning of week $t + 1$). This simple optimal stopping problem admits of the same general solution Eq. (2) considered before, since productivity within the week is declining with probability one.[52]

In practice, few workers work for a single spell during a given week. However, the above model can also be construed as applying to hours h_d within a single "shift" of work occuring on day d. Suppose that after a worker is withdrawn from labor for the day, they rest and productivity resets to p_0 on day $d + 1$. Owing to the Markovian property of productivity, the length of each spell h_d within a week will be independent of the others, and the total hours for the week $h = h_1 + h_2 + \dots h_7$ is distributed as a convolution of log concave densities. Such a convolution is itself log-concave (Saumard and Wellner, 2014), and hence BLC.

E Additional empirical information and results

E.1 Sample restrictions

Beginning with the initial sample described in Column (2) of Table 1, I keep paychecks from workers who are paid on a weekly basis, and condition on paychecks that contain a record of positive hours for work, vacation, holidays, or sick leave, totaling fewer than 80 hours in a week.[53] I also drop observations from California, which has a daily overtime rule that is binding for a significant number of workers, and could confound the effects of the weekly FLSA rule.

Furthermore, I focus on hourly workers. While the data include a field for the employer to input a salary, there is no guarantee that employers actually use this feature in the payroll software. Therefore, I use a combination of sampling restrictions to ensure I remove all non-hourly workers from the sample. First, I drop workers that ever have a salary on file with the payroll system. Second, I only keep workers at firms for whom *some* workers have a salary on file, the assumption being that employers either don't use the feature at all or use it for all of their salaried employees. I also drop paychecks from workers for whom hours are recorded as 40 in every week that they appear in the data,[54] as it is possible that these workers are simply coded as working 40 hours despite being paid on a salary basis. I also drop workers who never receive overtime pay.

[52]Hence, facing a weakly convex pay schedule, the firm never has incentive to keep the worker engaged in labor beyond the point at which their hourly productivity first dips below the marginal hourly wage. This leads directly to Equation (2). In this setting, h_{0it} is understood as the hours that the firm would choose if the worker were paid their straight wage for all hours, but faced the same realization of stochastic productivity decline this week as they actually do (and similarly for h_{1it})

[53]This restriction removes about 2% of the sample after the other restrictions. While a genuine 80 hour workweek is possible, I consider these observations to likely correspond to two weeks of work despite the worker's pay frequency being coded as weekly.

[54]For the purposes of this restriction, I count the "40 hours" event as occurring when either hours of work for pay or total hours of pay (including non-work pay like vacation) is equal to 40.

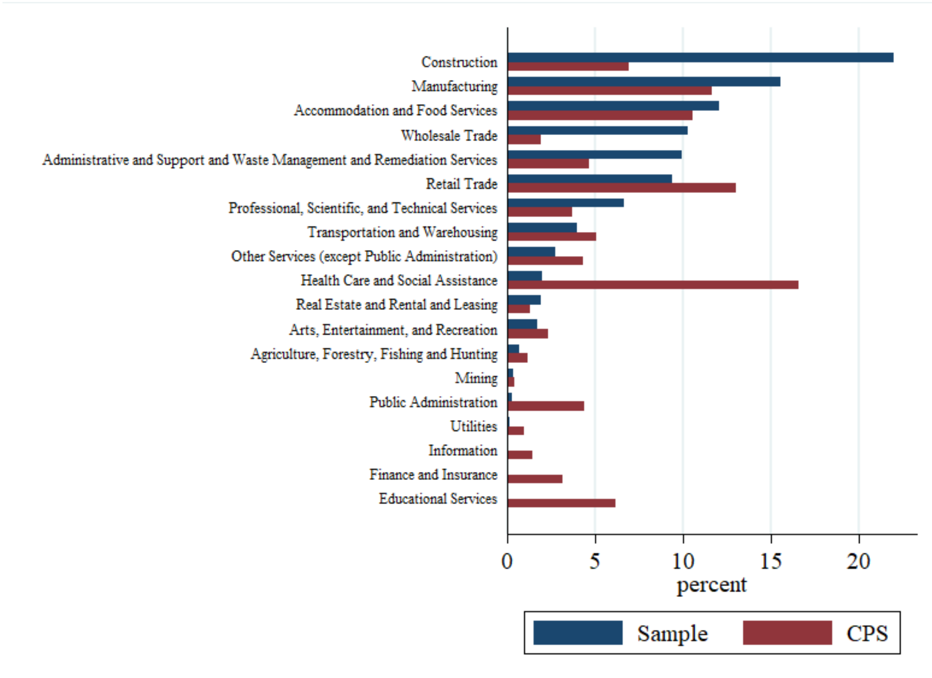

FIGURE E.7: Industry distribution of estimation sample collapsed to the worker level, compared with the CPS sample described in Section 3.

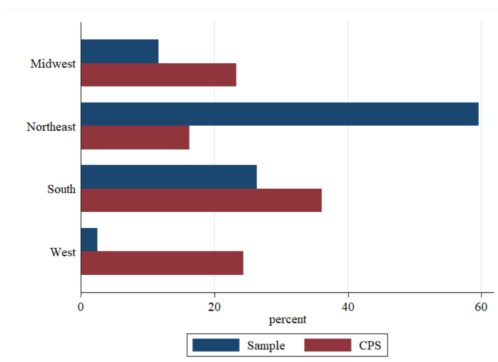

FIGURE E.8: Geographical distribution of estimation sample collapsed to the worker level, compared with the CPS sample described in Section 3.

E.2 Additional figures and tables

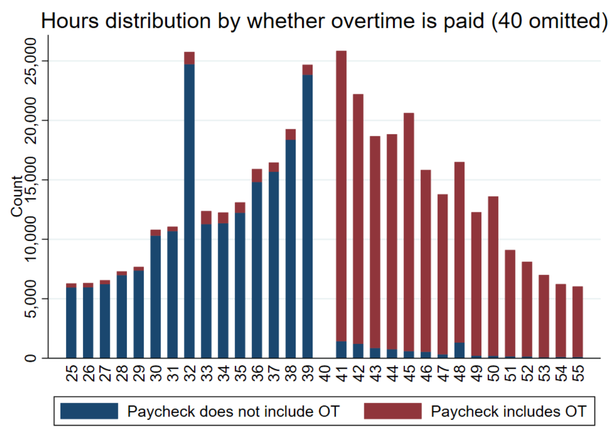

FIGURE E.9: Histogram of hours worked pooling all paychecks in sample, with one hour bins, with 40 omitted. Blue mass in the stacks indicate that the paycheck included no overtime pay, while red indicates that the paycheck does include overtime pay. The figure suggests two things: i) compliance with the FLSA overtime mandate is high; and ii) there is little overtime pay voluntarily offered to workers in weeks when they have less than 40 hours of work.

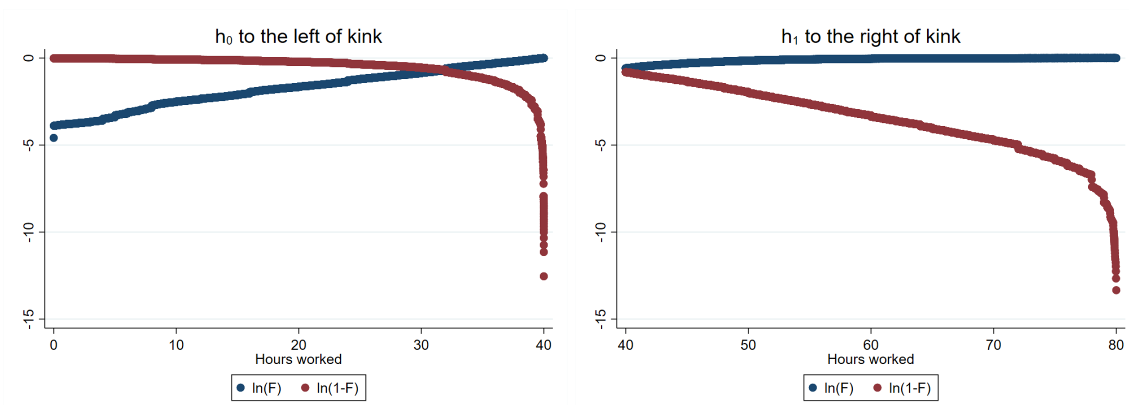

FIGURE E.10: Validating the assumption of bi-log-concavity away from the kink. The left panel plots estimates of $\ln F_0(h)$ and $\ln(1 - F_0(h))$ for $h < 40$, based on the empirical CDF of observed hours worked. Similarly the right panel plots estimates of $\ln F_1(h)$ and $\ln(1 - F_1(h))$ for $h > k$, where I've conditioned the sample on $Y_i < 80$. Bi-log-concavity requires that the four functions plotted be concave globally.

Industry	Avg. OT hours	OT % hours	OT % pay	Industry share
Accommodation and Food Services	2.37	0.06	0.11	0.08
Administrative and Support	5.69	0.13	0.18	0.08
Agriculture, Forestry, Fishing and Hunting	3.76	0.11	0.15	0.00
Arts, Entertainment, and Recreation	3.87	0.10	0.13	0.00
Construction	3.09	0.07	0.10	0.20
Educational Services	1.83	0.05	0.07	0.00
Finance and Insurance	0.31	0.00	0.01	0.00
Health Care and Social Assistance	4.59	0.12	0.12	0.02
Information	1.67	0.04	0.06	0.00
Manufacturing	3.37	0.08	0.11	0.18
Mining	2.26	0.07	0.12	0.00
Other Services	2.61	0.06	0.09	0.02
Professional, Scientific, and Technical Services	2.91	0.07	0.10	0.06
Public Administration	2.36	0.05	0.08	0.00
Real Estate and Rental and Leasing	2.85	0.07	0.09	0.02
Retail Trade	2.83	0.07	0.10	0.08
Transportation and Warehousing	5.24	0.12	0.17	0.04
Utilities	3.80	0.08	0.11	0.00
Wholesale Trade	5.15	0.11	0.14	0.10
Total Sample	3.55	0.08	0.12	1.00

TABLE E.1: Overtime prevalence by industry in the sample, including average number of OT hours per weekly paycheck, % OT hours among hours worked, % pay for hours work going to OT, and industry share of total hours in sample.

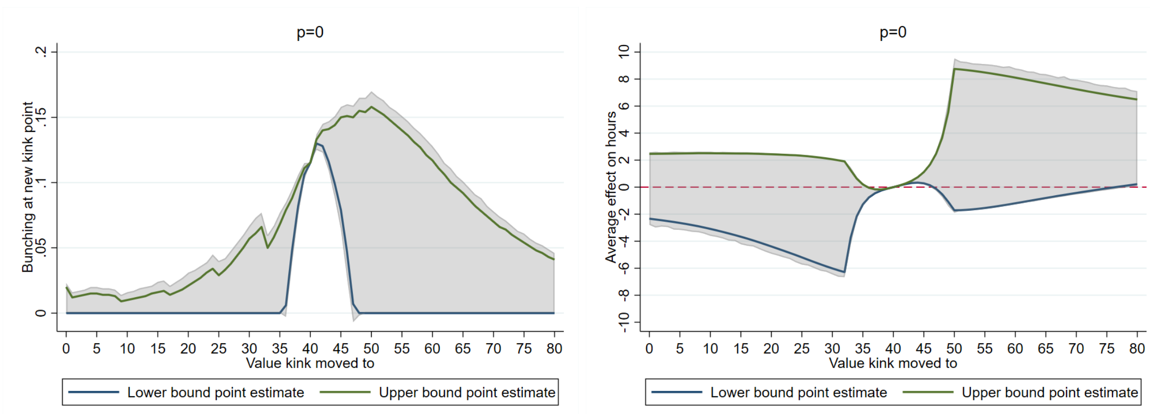

FIGURE E.11: Estimates of the bunching (left panel) and average effect on hours (right panel) were k changed to any value from 0 to 80, assuming $p = 0$. Pointwise bootstrapped 95% confidence intervals, cluster bootstrapped by firm, are shaded gray. Bounds are not informative far from 40. These estimates do not account for adjustment to straight-time wages, so should be viewed as quantifying short-run responses.

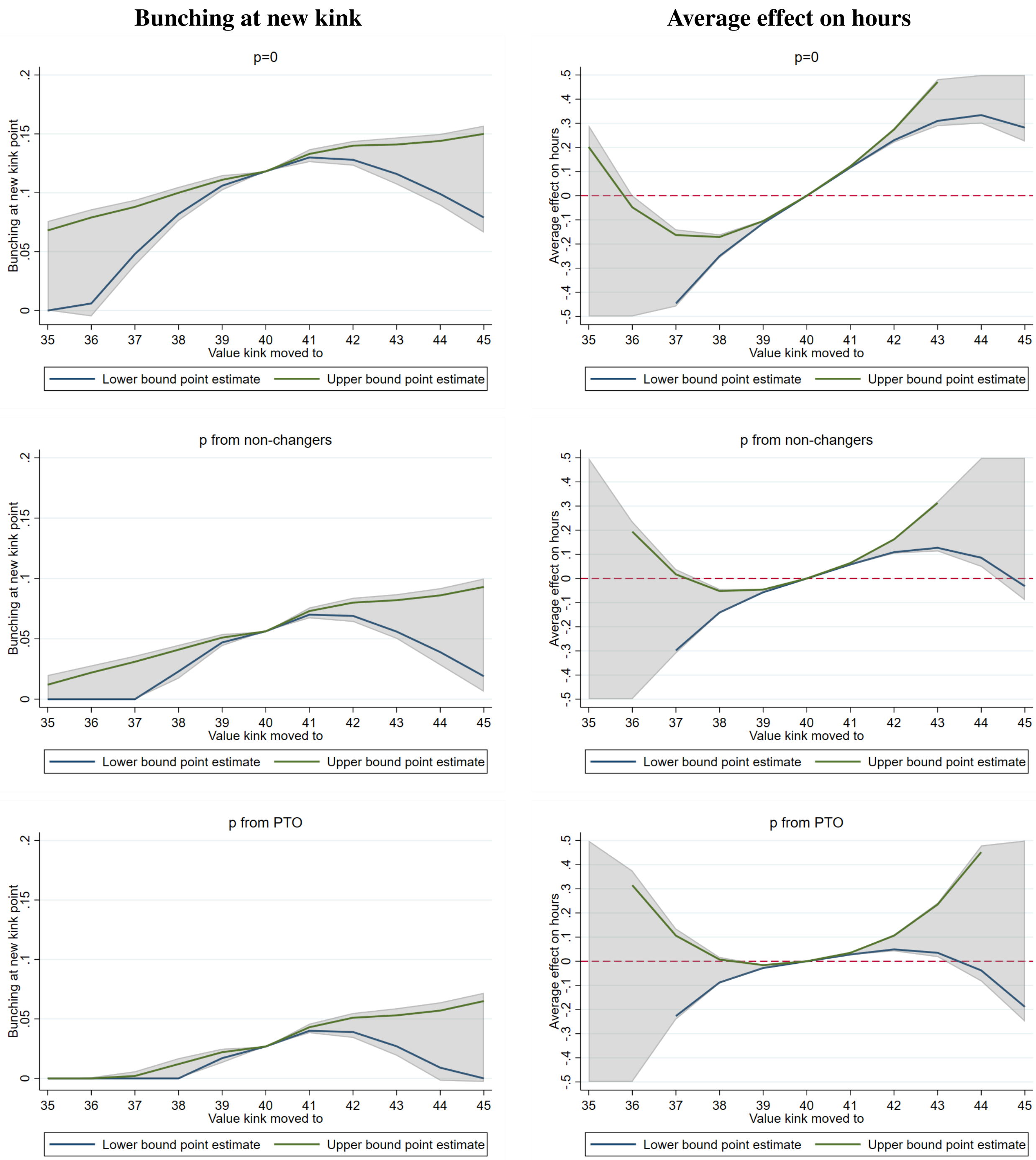

FIGURE E.12: Bounds for the bunching that would exist at standard hours k if it were changed from 40 (left panel), as well as for the impact on average hours (right panel). Bounds of the effect on hours are clipped to the interval $[-0.5, 0.5]$ for visibility. Pointwise bootstrapped 95% confidence intervals, cluster bootstrapped by firm, are shaded gray.

E.3 A test of the Trejo (1991) model of straight-time wage adjustment

One way to assess the role of the wage rigidity reported in Table 2 is to test directly whether straight-time wages and hours are plausibly related *at the weekly level* according to Equation (1). Given the kink in Eq. (1), we can perform such a test using the wage and hours reported on each paycheck, while making only weak differentiability assumptions on unobservables for identification.

Suppose that for some subset of units it, wages are actively adjusted to the hours they work according to Equation (1), in order to target some total earnings z_{it}. Denote the corresponding units by a latent variable $A_{it} = 1$. These units may come from workers with limited variation in their schedules in those weeks in which $h_{it} = h_i^*$ for some typical hours h_i^* according to which their wages were set by Eq. (1) at hiring. $A_{it} = 1$ units might instead have dynamic wages that adjust to week-by-week variation in their hours h_{it}. Let $A_{it} = 0$ denote units for whom the worker's wage is determined in some other way. Let $q(h) = P(A_{it} = 1|h_{it} = h)$ denote the proportion of these two groups at various points in the hours distribution. An extreme version of the fixed-job model of Trejo (1991) for example, would have $q(h) = 1$ for all h.

By the law of iterated expectations and some algebra we have that:

$$\mathbb{E}\left[\ln w_{it}|h_{it} = h\right] = q(h)\left\{\mathbb{E}\left[\ln z_{it}|h_{it} = h, A_{it} = 1\right] - \ln\left(h + 0.5(h - 40)\mathbb{1}(h \geq 40)\right)\right\} \\ - (1 - q(h))\mathbb{E}\left[\ln w_{it}|h_{it} = h, A_{it} = 0\right]$$

The middle term above introduces a kink in the conditional expectation of the log of straight-time wages with respect to hours. If we assume that $\mathbb{E}\left[\ln z_{it}|h_{it} = h, A_{it} = 1\right]$, $\mathbb{E}\left[\ln w_{it}|h_{it} = h, A_{it} = 0\right]$ and $q(h)$ are all continuously differentiable in h, then the magnitude of this kink identifies $q(40)$, the proportion of active wage responders local to $h = 40$:

$$\lim_{h \downarrow 40} \frac{d}{dh}\mathbb{E}\left[\ln w_{it}|h_{it} = h\right] - \lim_{h \uparrow 40} \frac{d}{dh}\mathbb{E}\left[\ln w_{it}|h_{it} = h\right] = -\frac{1}{2} \cdot \frac{q(40)}{40}$$

These continuous differentiability assumptions are reasonable, if wage setting according to Equation (1) is the only force introducing non-smoothness in the relationship between wages and hours at 40. For instance, we assume that production technologies do not have any special features at 40 hours that would cause the distribution of target earnings levels z_{it} among the $A_{it} = 1$ units to itself have a kink around $h_{it} = 40$.

Figure E.13 reports the results of fitting separate local linear functions to the CEF of log wages on either side of $h = 40$. We can reject the hypothesis that the fixed-job model applies to all employees at all times, near 40. However, the data appear to be consistent with a proportion $q(40) \approx 0.25$ of all paychecks close to 40 hours reflecting an hours/wage relationship governed by Equation (1). This can be rationalized by straight-wages being updated intermittently to reflect expected or anticipated hours, which vary in practice quite a bit between pay periods.

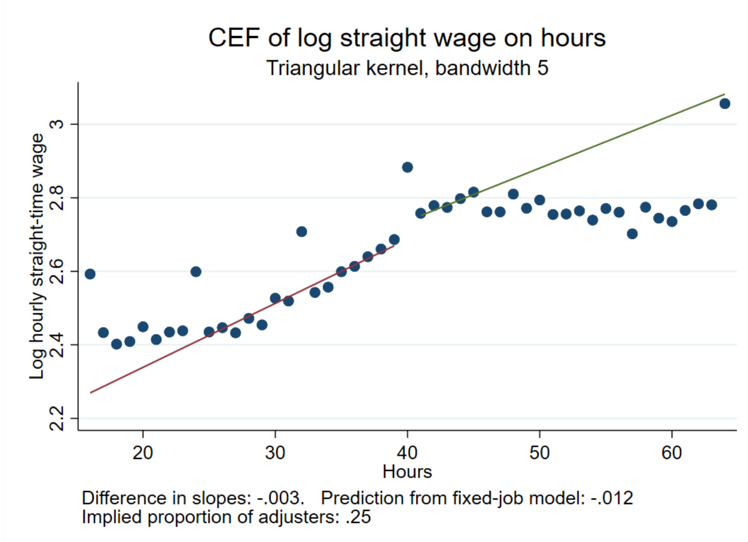

FIGURE E.13: A test of the fixed-jobs model presented in Trejo (1991), based on the magnitude of the kink in the conditional expectation of log wages with respect to hours at 40 (see above). Regression lines fit on each side with a uniform kernel within 25 hours of the 40. This figure closely resembles Figure 5 of Bick et al. (2022) which uses CPS data for hourly workers.

E.4 Details of the employment effect calculation

Taking my prefered estimate that hourly workers work approximately 1/3 of an hour less per week on average because of the rule, hours per worker are reduced by just under 1%. If we assume the same sized effect occurs for covered salaried workers, and ignore scale effects of the overtime rule on the total number of labor hours in FLSA-eligible jobs, this suggests employment among such jobs is 1% higher than it would be without the overtime premium. This serves as an upper bound, since overall total hours worked may decrease due to overtime regulation.

Following Hamermesh (1993), assume that the percentage change in employment decomposes as:

$$\Delta \ln E|_{EH} - \eta \cdot \Delta \ln LC \cdot \frac{\eta}{\alpha - \eta} \tag{E.13}$$

where η is constant-output demand elasticity for labor, α is a labor supply elasticity. Following Hamermesh (1993) I use $\Delta \ln LC = 0.7\%$ based on Ehrenberg and Schumann (1982), calibrated assuming that 80% of labor costs come from wages with overtime representing 2% of total hours. $\Delta \ln E|_{EH}$ is the quantity implied by my estimates: the percentage change in employment that would occur were the total number of worker-hours EH unchanged. Taking a preferred estimate of the average effect of the FLSA as reported in Table 4 to be about 1/3 of an hour, I use a value of $\Delta \ln E|_{EH} = \frac{1/3}{40} \approx 0.9\%$.

"Best-guess" values for the other parameters used by Hamermesh, 1993 are $\eta = -0.3$ and $\alpha = 0.1$, based on a review of empirical estimates. This yields 0.17 percentage points for the substitution term $\eta \cdot \Delta \ln LC \cdot \frac{\eta}{\alpha - \eta}$, suggesting that the effect of the FLSA is attenuated from roughly 0.87 percentage points to about a 0.70 percentage point net increase in employment. I assume that the FLSA overtime rule applies

		η		
		-0.15	-0.3	-0.5
	0	0.76	0.64	0.50
α	0.1	0.80	**0.70**	0.56
	0.5	0.85	0.79	0.68

TABLE E.2: Back-of-the-envelope employment effects based on the average reduction in hours estimated via the bunching design and Equation (E.13), as a function of the demand elasticity for labor (rather than capital) η, and labor supply elasticity α. The bold entry reflects "best-guess" values of η and α.

to a total of 100 million workers, based on 80 million hourly workers combined with an estimated 20 million covered salary workers Kimball and Mishel (2015). Assuming the same percentage increase in employment applies to hourly workers and covered salary workers, the above estimate corresponds to 700,000 jobs created. Generating a negative overall employment response by assuming higher substitution to capital requires $\eta = -1.25$, well outside of empirical estimates.

F Additional proofs

F.1 Proof of Propositions 1 and 2

Consider Proposition 1. Item i) in the proof of Lemma 1 establishes that under CHOICE and WARP $Y_i = k$ implies $Y_{1i} \leq k \leq Y_{0i}$, since taking contrapositives we have that ($Y_i \geq k$ and $Y_i \leq k$) implies $Y_{1i} \leq k \leq Y_{0i}$. We have also seen from item ii) that under CHOICE and CONVEX $Y_{1i} \leq k \leq Y_{0i}$ also implies $Y_i = k$, thus $Y_{1i} \leq k \leq Y_{0i}$ and $Y_i = k$ are equivalent. Note that by adding $\Delta_i = Y_{0i} - Y_{1i}$ to both sides of the inequality $Y_{1i} \leq k$, we have that $Y_{0i} \leq k + \Delta_i$. Combining with the other inequality that $Y_{0i} \geq k$, we can thus rewrite the event $Y_{1i} \leq k \leq Y_{0i}$ as $Y_{0i} \in [k, k + \Delta_i]$ (or equivalently to $Y_{1i} \in [k - \Delta_i, k]$). We thus have that $\mathcal{B} \leq P(Y_{0i} \in [k, k + \Delta])$ under CHOICE and WARP, and that $\mathcal{B} = P(Y_i = k) = P(Y_{1i} \leq k \leq Y_{0i})$ under CHOICE and CONVEX.

Now consider Proposition 2. By item i) in the proof of Proposition 1, we have that under WARP and CHOICE $Y_{0i} \leq k \implies Y_i = Y_{0i}$. Thus, for any $\delta > 0$ and $y < k$: $Y_{0i} \in [y - \delta, y]$ implies that $Y_i \in [y - \delta, y]$ and hence $P(Y_{0i} \in [y - \delta, y]) \leq P(Y_i \in [y - \delta, y])$. This implies that $f_0(y) - f(y) = \lim_{\delta \downarrow 0} \frac{P(Y_{0i} \in [y-\delta, y]) - P(Y_i \in [y-\delta, y])}{\delta} \leq 0$, i.e. that $f(y) \geq f_0(y)$. An analogous argument holds for Y_1, where we consider the event $Y_{1i} \in [y, y + \delta]$ any $y > k$. Under CONVEX instead of WARP, the inequalities are all equalities, by Lemma 1.

F.2 Proof of Lemma 2

Let $\Delta_i^k(\rho, \rho') := Y_i(\rho, k) - Y_i(\rho', k)$ for any $\rho, \rho' \in [\rho_0, \rho_1]$ and value of k.

Assumption SMOOTH (regularity conditions). *The following hold:*

1. *$P(\Delta_i^k(\rho,\rho') \le \Delta, Y_i(\rho,k) \le y)$ is twice continuously differentiable at all $(\Delta, y) \neq (0, k^*)$, for any $\rho, \rho' \in [\rho_0, \rho_1]$ and k.*

2. *$Y_i(\rho, k) = Y(\rho, k, \epsilon_i)$, where ϵ_i has compact support $E \subset \mathbb{R}^m$ for some m. $Y(\cdot, k, \cdot)$ is continuously differentiable on all of $[\rho_0, \rho_1] \times E$, for every k.*

3. *there possibly exists a set $\mathcal{K}^* \subset E$ such that $Y(\rho, k, \epsilon) = k^*$ for all $\rho \in [\rho_0, \rho_1]$ and $\epsilon \in \mathcal{K}^*$. The quantity $\mathbb{E}\left[\frac{\partial Y_i(\rho,k)}{\partial \rho}\middle| Y_i(\rho,k) = y, \epsilon_i \notin \mathcal{K}^*\right]$ is continuously differentiable in y for all y including k^*.*

In the remainder of this proof I keep k be implicit in the functions $Y_i(\rho,k)$ and $\Delta_i^k(\rho,\rho')$, as it will remained fixed. Item 1 of SMOOTH excludes the point $(0, k^*)$ on the basis that we may expect point masses at $Y_i(\rho) = k^*$, as in the overtime setting. Following Section 4, item 3 imposes that all such "counterfactual bunchers" have zero treatment effects, while also introducing a further condition that will be used later in Lemma 3. Let K_i^* be an indicator for $\epsilon_i \in \mathcal{K}^*$ and denote $p = P(K_i^* = 1)$. Item 1 implies that the density $f_{\Delta(\rho,\rho'),Y(\rho)}(\Delta, y)$ is continuous in y whenever $y \neq k^*$ or $\Delta \neq 0$, so I define $f_{\Delta(\rho,\rho'),Y(\rho)}(\Delta, k^*) = \lim_{y\to k^*} f_{\Delta(\rho,\rho'),Y(\rho)}(\Delta, y)$ for any ρ, ρ' and Δ. Similarly, we can define the marginal density $f_\rho(y)$ of $Y_i(\rho)$ at k^* to be $\lim_{y\to k^*} f_\rho(y)$ for any ρ.

By item 1 of Assumption SMOOTH, the marginal $F_\rho(y) := P(Y_i(\rho) \le y)$ is differentiable away from $y = k$ with derivative $f_\rho(y)$. From the proof of Theorem 1 it follows that $\mathcal{B} \le F_{\rho_1}(k) - F_{\rho_0}(k) + p(k)$ with equality under CONVEX, and thus:

$$\begin{aligned}\mathcal{B} - p(k) \le F_{\rho_1}(k) - F_{\rho_0}(k) &= \int_{\rho_0}^{\rho_1} \frac{d}{d\rho} F_\rho(k) d\rho = \int_{\rho_0}^{\rho_1} \lim_{\delta\downarrow 0} \frac{F_{\rho+\delta}(k) - F_\rho(k)}{\delta} d\rho \\ &= \int_{\rho_1}^{\rho_0} \lim_{\delta\downarrow 0} \frac{P(Y_i(\rho+\delta) \le k \le Y_i(\rho)) - p(k)}{\delta} d\rho = \int_{\rho_1}^{\rho_0} f_\rho(k)\mathbb{E}\left[-\frac{Y_i(\rho)}{d\rho}\middle| Y_i(\rho) = k\right] d\rho\end{aligned}$$

where the third equality applies the identity $1 = P(Y_{0i} \le k) + P(Y_i(\rho) \le k \le Y_i(\rho+\delta)) + P(Y_{1i} > k)$ under CHOICE and WARP (this follows from item i) of the proof of Lemma 1) to the pair of choice constraints $B(\rho)$ and $B(\rho+\delta)$, noting that $P(Y_i(\rho) < k) = F_\rho(k) - p(k)$. The final equality uses Lemma SMALL.

F.3 Proof of Lemma SMALL

Throughout this proof we let f_W denote the density of a generic random variable or random vector W_i, if it exists. Write $\Delta_i(\rho,\rho') = \Delta_i(\rho,\rho',\epsilon_i)$ where $\Delta_i(\rho,\rho',\epsilon) := Y(\rho,\epsilon) - Y(\rho',\epsilon)$.

$$
\begin{aligned}
\lim_{\rho'\downarrow\rho} \frac{P(Y_i(\rho) \le k \le Y_i(\rho')) - p(k)}{\rho'-\rho} &= \lim_{\rho'\downarrow\rho} \frac{P(Y_i(\rho) \in [k, k+\Delta(\rho,\rho')_i]) - p(k)}{\rho'-\rho} \\
&= \lim_{\rho'\downarrow\rho} \frac{P(Y_i(\rho) \in (k, k+\Delta(\rho,\rho')_i])}{\rho'-\rho} \\
&= \lim_{\rho'\downarrow\rho} \frac{1}{\rho'-\rho} \int_0^\infty d\Delta \int_k^{k+\Delta} dy \cdot f_{\Delta(\rho,\rho'),Y(\rho)}(\Delta, y) \\
&= \lim_{\rho'\downarrow\rho} \int_0^\infty d\Delta \int_k^{k+\Delta} dy \cdot \frac{f_{\Delta(\rho,\rho'),Y(\rho)}(\Delta,k) + (y-k) r_{\Delta(\rho,\rho'),Y(\rho)}(\Delta,k,y)}{\rho'-\rho}
\end{aligned}
\tag{F.14}
$$

where we have used that by item 1 the joint density of $\Delta_i(\rho,\rho')$ and $Y_i(\rho)$ exists for any ρ,ρ' and is differentiable and $r_{\Delta(\rho,\rho'),Y(\rho)}$ is a first-order Taylor remainder term satisfying

$$
\lim_{y\downarrow k} |r_{\Delta(\rho,\rho'),Y(\rho)}(\Delta,y)| = |r_{\Delta(\rho,\rho'),Y(\rho)}(\Delta,k)| = 0
$$

for any Δ.

I now show that the whole term corresponding to this remainder is zero. First, note that:

$$
\begin{aligned}
\left|\lim_{\rho'\downarrow\rho} \int_0^\infty d\Delta \int_k^{k+\Delta} dy \cdot \frac{(y-k) r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta,y)}{\rho'-\rho}\right| &= \lim_{\rho'\downarrow\rho} \left| \int_0^\infty d\Delta \int_k^{k+\Delta} dy \cdot \frac{(y-k) r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta,y)}{\rho'-\rho} \right| \\
&\le \lim_{\rho'\downarrow\rho} \int_0^\infty d\Delta \int_k^{k+\Delta} dy \cdot \left| \frac{(y-k) r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta,y)}{\rho'-\rho} \right| \\
&\le \lim_{\rho'\downarrow\rho} \int_0^\infty d\Delta \frac{\Delta}{\rho'-\rho} \int_k^{k+\Delta} dy \cdot \left| r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta,y) \right|
\end{aligned}
$$

where I've used continuity of the absolute value function and the Minkowski inequality. Define $\xi(\rho,\rho') = \sup_{\epsilon\in E} \Delta(\rho,\rho',\epsilon)$. The strategy will be show that $\lim_{\rho'\downarrow\rho} \xi(\rho,\rho') = 0$, and then since $r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta,y) = 0$ for any $\Delta > \xi(\rho,\rho')$ and all y (since the marginal density $f_{\Delta(\rho,\rho')}(\Delta)$ would be zero for such Δ). With $\xi(\rho,\rho')$ so-defined:

$$
\begin{aligned}
\text{RHS of above} &\le \lim_{\rho'\downarrow\rho} \int_0^{\xi(\rho,\rho')} d\Delta \frac{\xi(\rho,\rho')}{\rho'-\rho} \int_k^{k+\xi(\rho,\rho')} dy \cdot \left| r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta,y) \right| \\
&= \lim_{\rho'\downarrow\rho} \frac{\xi(\rho,\rho')}{\rho'-\rho} \cdot \lim_{\rho'\downarrow\rho} \int_0^{\xi(\rho,\rho')} d\Delta \int_0^{\xi(\rho,\rho')} dy \cdot \left| r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta, k+y) \right|
\end{aligned}
\tag{F.15}
$$

where in the second step I have assumed that each limit exists (this will be demonstrated below). Let us first consider the inner integral of the above: $\int_k^{k+\xi(\rho,\rho')} dy \cdot \left| r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta,y) \right|$, for any Δ. Supposing that $\lim_{\rho'\downarrow\rho} \xi(\rho,\rho') = 0$, it follows that this inner integral evaluates to zero, by the

Leibniz rule and using that $r_{\Delta_i(\rho,\rho'),Y_i(\rho)}(\Delta, k) = 0$. Thus the entire second limit is equal to zero.

Now I prove that $\lim_{\rho'\downarrow\rho} \xi(\rho, \rho') = 0$ and that $\lim_{\rho'\downarrow\rho} \frac{\xi(\rho,\rho')}{\rho'-\rho}$ exists. First, note that continuous differentiability of $Y(\rho, \epsilon_i)$ implies $Y_i(\rho)$ is continuous for each i so $\lim_{\rho'\downarrow\rho} \Delta_i(\rho, \rho') = 0$ point-wise in ϵ. We seek to turn this point-wise convergence into uniform convergence over ϵ, i.e. that $\lim_{\rho'\downarrow\rho} \sup_{\epsilon\in E} \Delta(\rho, \rho', \epsilon) = \sup_{\epsilon\in E} \lim_{\rho'\downarrow\rho} \Delta(\rho, \rho', \epsilon) = \sup_{\epsilon\in E} 0 = 0$. The strategy will be to use equicontinuity of the sequence and compactness of E. Consider any such sequence $\rho_n \overset{n}{\to} \rho$ from above, and let $f_n(\epsilon) := Y(\rho, \epsilon) - Y(\rho_n, \epsilon)$ and $f(\epsilon) = \lim_{n\to\infty} f_n(\epsilon) = 0$. Equicontinuity of the sequence $f_n(\epsilon)$ says that for any $\epsilon, \epsilon' \in E$ and $e > 0$, there exists a $\delta > 0$ such that $||\epsilon - \epsilon'|| < \delta \implies |f_n(\epsilon) - f_n(\epsilon')| < e$.

This follows from continuous differentiability of $Y(\rho, \epsilon)$. Let $M = \sup_{\rho\in[\rho_0,\rho_1],\epsilon\in E} |\nabla_{\rho,\epsilon} Y(\rho, \epsilon)|$. M exists and is finite given continuity of the gradient and compactness of $[\rho_0, \rho_1] \times E$. Then, for any two points $\epsilon, \epsilon' \in E$ and any $\rho \in [\rho_0, \rho_1]$:

$$|Y(\rho, \epsilon) - Y(\rho, \epsilon')| = \left|\int_{\epsilon'}^{\epsilon} \nabla_\epsilon Y(\rho, \epsilon) \cdot \mathbf{d}\epsilon\right| \leq \int_{\epsilon'}^{\epsilon} |\nabla_\epsilon Y(\rho, \epsilon) \cdot \mathbf{d}\epsilon| \leq M \int_{\epsilon'}^{\epsilon} ||\mathbf{d}\epsilon|| \leq M||\epsilon - \epsilon'||$$

where $\mathbf{d}\epsilon$ is any path from ϵ to ϵ' and I have used the definition of M and Cauchy-Schwarz in the second inequality. The existence of a uniform Lipschitz constant M for $Y(\rho, \epsilon)$ implies a uniform equicontinuity of $Y(\rho, \epsilon)$ of the form that for any $e > 0$ and $\epsilon, \epsilon' \in E$, there exists a $\delta > 0$ such that $||\epsilon - \epsilon'|| < \delta \implies \sup_{\rho\in[\rho_0,\rho_1]} |Y(\rho, \epsilon) - Y(\rho, \epsilon')| < e/2$, since we can simply take $\delta = e/(2M)$. This in turn implies that whenever $||\epsilon - \epsilon'|| < \delta$:

$$\begin{aligned}|Y(\rho, \epsilon) - Y(\rho_n, \epsilon) - \{Y(\rho, \epsilon') - Y(\rho_n, \epsilon')\}| &= |Y(\rho, \epsilon) - Y(\rho, \epsilon') - \{Y(\rho_n, \epsilon) - Y(\rho_n, \epsilon')\}| \\ &\leq |Y(\rho, \epsilon) - Y(\rho, \epsilon')| + |Y(\rho_n, \epsilon) - Y(\rho_n, \epsilon')| \leq e,\end{aligned}$$

our desired result. Together with compactness of E, equicontinuity implies that $\lim_{n\to\infty} \sup_{\epsilon\in E} f_n(\epsilon) = \sup_{\epsilon\in E} \lim_{n\to\infty} f_n(\epsilon) = 0$.

We apply an analogous argument for $\lim_{\rho'\downarrow\rho} \frac{\xi(\rho,\rho')}{\rho'-\rho}$, where now $f_n(\epsilon) = \frac{Y(\rho,\epsilon)-Y(\rho_n,\epsilon)}{\rho_n-\rho}$. For this case it's easier to work directly with the function $\frac{Y(\rho,\epsilon)-Y(\rho_n,\epsilon)}{\rho_n-\rho}$, showing that it is Lipschitz in deviations of ϵ uniformly over $n \in \mathbb{N}, \epsilon \in E$.

$$\begin{aligned}\left|\frac{Y(\rho, \epsilon) - Y(\rho_n, \epsilon)}{\rho_n - \rho} - \frac{Y(\rho, \epsilon') - Y(\rho_n, \epsilon')}{\rho_n - \rho}\right| &= \frac{1}{\rho_n - \rho}\left|\int_{\epsilon'}^{\epsilon} \nabla_\epsilon Y(\rho, \epsilon) \cdot \mathbf{d}\epsilon - \int_{\epsilon'}^{\epsilon} \nabla_\epsilon Y(\rho_n, \epsilon) \cdot \mathbf{d}\epsilon\right| \\ &\leq \frac{1}{\rho_n - \rho}\left(\left|\int_{\epsilon'}^{\epsilon} \nabla_\epsilon Y(\rho, \epsilon) \cdot \mathbf{d}\epsilon\right| + \left|\int_{\epsilon'}^{\epsilon} \nabla_\epsilon Y(\rho_n, \epsilon) \cdot \mathbf{d}\epsilon\right|\right) \\ &\leq \frac{2M}{\rho_n - \rho}\int_{\epsilon'}^{\epsilon} ||\mathbf{d}\epsilon|| \leq \frac{2M}{\rho_n - \rho}||\epsilon - \epsilon'||\end{aligned}$$

This implies equicontinuity of $\frac{Y(\rho,\epsilon)-Y(\rho_n,\epsilon)}{\rho_n-\rho}$ with the choice $\delta = e(\rho_n - \rho)/(2M)$. As before, equicontinuity and compactness of E allow us to interchange the limit and the supremum, and thus:

$$\lim_{n\to\infty} \frac{\xi(\rho,\rho_n)}{\rho_n-\rho} = \lim_{n\to\infty} \frac{\sup_{\epsilon\in E}\{Y(\rho,\epsilon)-Y(\rho_n,\epsilon)\}}{\rho_n-\rho} = \lim_{n\to\infty}\sup_{\epsilon\in E} \frac{Y(\rho,\epsilon)-Y(\rho_n,\epsilon)}{\rho_n-\rho}$$
$$= \sup_{\epsilon\in E}\lim_{n\to\infty} \frac{Y(\rho,\epsilon)-Y(\rho_n,\epsilon)}{\rho_n-\rho} = \sup_{\epsilon\in E}\frac{\partial Y(\rho,\epsilon)}{\partial \rho} := M' < \infty$$

where finiteness of M' follows from it being defined as the supremum of a continuous function over a compact set. This establishes that the first limit in Eq. (F.15) exists and is finite, completing the proof that it evaluates to zero.

Given that the second term in Eq. (F.14) is zero, we can simplify the remaining term as:

$$\lim_{\rho'\downarrow\rho}\frac{P(Y_i(\rho)\le k\le Y_i(\rho'))-p(k)}{\rho'-\rho} = \lim_{\rho'\downarrow\rho}\frac{1}{\rho'-\rho}\int_0^\infty f_{\Delta(\rho,\rho'),Y(\rho)}(\Delta,k)\Delta d\Delta$$
$$= f_\rho(k)\lim_{\rho'\downarrow\rho}\frac{1}{\rho'-\rho}P(\Delta_i(\rho,\rho')\ge 0|Y_i(\rho)=k)\cdot\mathbb{E}\left[\Delta_i(\rho,\rho')|Y_i(\rho)=k,\Delta_i(\rho,\rho')\ge 0\right]$$
$$= f_\rho(k)(k)\lim_{\rho'\downarrow\rho}\frac{1}{\rho'-\rho}\mathbb{E}\left[\Delta_i(\rho,\rho')|Y_i(\rho)=k,\Delta_i(\rho,\rho')\right] = f_\rho(k)(k)\mathbb{E}\left[\lim_{\rho'\downarrow\rho}\frac{\Delta_i(\rho,\rho')}{\rho'-\rho}\middle| Y_i(\rho)=k\right]$$
$$= f_\rho(k)\mathbb{E}\left[-\frac{Y_i(\rho)}{d\rho}\middle| Y_i(\rho)=k\right]$$

where I have used that treatment effects must be positive at the kink (see Lemma POS in Appendix J.4) and then the dominated convergence theorem. To see that we may use it, note that $\frac{dY_i(\rho)}{d\rho} = \frac{\partial Y(\rho,\epsilon_i)}{\partial\rho} < M$ uniformly over all $\epsilon_i\in E$, and $\mathbb{E}\left[M|\,Y_i(\rho)=k\right] = M < \infty$.