Within-Year Labor Supply Responses to Year-End Incentives*

Andreas R. Kostøl† Andreas S. Myhre ‡ Matthew Merkle§ Mark Whitmeyer¶

This Version: April 30, 2026

Abstract: This paper studies how annual tax incentives translate into within-year labor supply responses when workers cannot smoothly adjust hours. We show that annual incentives can generate a within-year stopping margin, and we develop a nonparametric approach that identifies this participation component using monthly earnings data. We apply the approach to annual kinks in the Norwegian disability insurance (DI) and tax system, where the monthly data reveal whether workers stop before—rather than merely bunch—at year-end. We estimate a participation elasticity of 0.8 at the DI kink, and show that at least 60% of the annual tax response there operates through the extensive margin. Evidence elsewhere in the tax schedule is weaker and often imprecise, suggesting that the mechanism is strongest where incentives are large and labor-force attachment is weak.

Keywords: labor supply; hours constraints; tax incentives; disability insurance; bunching.
JEL codes: H20; H31; H55; J22; J26

*This project received financial support from the Research Council of Norway project 345156. This project has benefited greatly from conversations with Richard Blundell and Espen Moen, and we appreciate suggestions from participants at the IFN Stockholm Conference in 2023, the NBER Summer Institute Public Economics, and the 3rd Oslo Labor-Public Economics Workshop. The draft has previously circulated under the title "Hours Mismatch and Annual Taxes" and "Missing Work: Elasticities with Dynamic Optimization Frictions". All errors are our own. Refine.ink helped us improve the paper.

†BI Norwegian Business School; IZA. E-mail: andreas.r.kostol@bi.no
‡Statistics Norway. E-mail: andreassmyh@gmail.com
§University College London; Institute for Fiscal Studies. E-mail: m.merkle@ucl.ac.uk
¶Department of Economics, Arizona State University; E-mail: mark.whitmeyer@gmail.com

1 Introduction

Economists have long recognized that inflexible hours and fixed costs of work shape labor supply behavior (e.g., Cogan, 1981 and Chetty *et al.*, 2013). Such constraints, whether implicit or explicit, are central to understanding labor market dynamics and tax design (e.g., Hansen, 1985 and Rogerson, 1988), and optimal taxation (e.g., Saez, 2002). However, this literature has been studied largely through annual labor supply models, leaving within-year responses much less understood.[1]

In this paper, we study whether annual tax incentives generate within-year stopping responses when workers cannot smoothly adjust hours. To do so, we develop a dynamic labor supply model in which workers repeatedly decide whether to continue working as the year unfolds. The key implication is that annual taxation plus hours constraints can create a distinct within-year participation margin that is not visible in annual data alone. Unlike the standard static response to annual tax incentives, inflexible hours turn the problem into a stop-or-continue decision: some workers optimally pause once cumulative earnings cross a kink and resume when the schedule resets. That mechanism gives the monthly data independent identifying content, rather than treating them as a finer version of a standard bunching design.[2]

We interpret this stopping behavior as an extensive-margin response to annual incentives observed at the monthly level. At the same time, a monthly extensive-margin decision can also be viewed as an annual intensive-margin decision.[3] We therefore use this language mainly to clarify the timing and composition of responses within the year. More broadly, our results isolate a mechanism that matters precisely when work choices are lumpy and incentives are sharp, rather than claiming that all annual tax responses primarily reflect within-year stopping. We therefore begin with the disability insurance (DI)

[1] Keane (2011) and Saez *et al.* (2012) survey the literature on dynamic labor supply decisions and the elasticity of taxable income, respectively. A growing number of studies in public economics show that responses to incentives are often dampened by various kinds of frictions (e.g., Chetty, 2012 and Kleven & Waseem, 2013). Saez (2010) argued that these frictions explain why employees often do not respond to tax incentives, and Chetty *et al.* (2013) show that hours constraints can rationalize the divergence between Hicksian macro and micro elasticities, but not the intertemporal elasticity of labor supply: "Even accounting for indivisible labor, micro studies do not support representative-agent macro models that generate Frisch elasticities above 1" (Chetty *et al.*, 2013, p. 4).

[2] Quitting is synonymous here with pausing work. The costs associated with quitting also capture the cost of returning to work in the next tax period, or at a later date. Of course, quitting or pausing work may impose high career costs on full-time workers (Kleven *et al.*, 2025), but to some workers, ignoring the tax incentive can be costlier than quitting.

[3] In general, an extensive-margin decision over any time interval may be interpreted as an intensive-margin decision over a longer time interval, and an intensive-margin decision over any time interval may be interpreted as an extensive-margin decision over a shorter time interval.

system, where the design has its best chance of succeeding because the sample consists of marginally attached workers facing an unusually large kink. In this setting, ignoring the incentive is especially costly, because the marginal tax rate rises by 43 percentage points at the kink. We find that a temporary 62 percent reduction in net-of-tax income leads to a 53 percent fall in the participation rate, yielding an implied participation elasticity of 0.8. This result is robust across a wide set of specifications and holds for part-time workers who return to the same employer, return to a new employer, or leave the workforce when the tax resets the following year.

Our approach allows us to decompose year-end bunching into intensive- and extensive-margin responses and to estimate the share of workers who are hours-constrained. By relating the amount of "missing mass of work" implied by a participation probability function (PPF) to the "excess mass in year-end earnings" from a standard bunching estimate, we can derive the share of constrained "compliers"—i.e., workers who find it optimal to locate at or close to the kink. We find that while the missing mass of work accounts for about 2 percent of part-time employed welfare recipients, it accounts for 60 percent of "compliers."

As highlighted by Chetty (2012), theory suggests inertia when the utility cost of ignoring the incentive is low relative to the adjustment cost. To test these predictions, we then turn to the full spectrum of kinks in the tax system, which range from marginal tax changes of 3 to 25 percentage points and span the 10th to the 90th percentile of the earnings distribution. The broader tax-schedule evidence is materially weaker than the DI evidence. We detect some extensive-margin response at the lowest bracket, where moving from a tax rate of zero to 25 percent implies a labor supply response of about 5 percent and an elasticity of 0.2. At higher kinks, the estimates are smaller and often imprecise. We therefore treat the tax-schedule results as corroborative rather than as a headline empirical contribution.

The participation elasticities are largest among marginally attached workers, precisely the group for whom within-year stopping is most plausible. One interpretation is that these workers face relatively low search or re-entry costs, making temporary labor market exit and return easier to sustain within the year. Indeed, many workers stop and return to the same firm the following year, suggesting that firms can accommodate temporary exit even when hours are rigid. Career concerns also appear less important in this population, since many workers hold part-time jobs or are near retirement age. Consistent with this interpretation, the estimated participation elasticity is smaller among younger workers and among workers in larger firms, for whom career concerns may be more important.

Our paper is related to research on the implications of optimization frictions for la-

bor supply decisions. Saez (2010) discusses the role of hours frictions as one reason why workers do not bunch, and Chetty *et al.* (2011) point to aggregate factors—such as collective wage bargaining—that may attenuate the estimated micro response to incentives. Chetty (2012) formalizes the idea of optimization frictions, and Kleven & Waseem (2013) show how notches and dominated regions in budget sets can be used to identify the overall prevalence of optimization frictions.[4] Relative to this research, our approach nests the standard bunching approach of Saez (2010), in which people make a single labor supply decision, but uses monthly earnings data to study how annual responses are distributed within the year.[5] Our main contribution relative to the bunching and optimization-frictions literature is to use monthly data to decompose year-end bunching into within-year stopping and residual annual earnings adjustment. Compared with Kleven & Waseem (2013), who use annual missing and excess mass to study optimization frictions, we use monthly continuation behavior to identify the timing and composition of responses around annual kinks.

Our work also contributes to research on responses to changes in work incentives (e.g., Moffitt, 1992; Krueger & Meyer, 2002; Saez *et al.*, 2012; and Kleven & Waseem, 2013). While it is well-established that labor supply elasticities differ widely across groups or countries, understanding which margins drive those differences remains an active research field (see, e.g., Krueger & Meyer, 2002 and Chetty, 2012).[6] Our paper contributes to this line of research by unpacking how hours frictions shape labor supply elasticities, how they interact with the safety net, and why extensive-margin responses are especially important at the bottom of the income distribution. Our findings also connect to a large literature showing that disability insurance (DI) applications are counter-cyclical (e.g., Black *et al.*, 2002; Autor & Duggan, 2003; Maestas *et al.*, 2021). While most of that literature focuses on participation among potential beneficiaries, we show that the elasticity also remains high among existing beneficiaries.

Our approach also speaks to the macro-micro evidence on the elasticity of intertemporal substitution. Chetty *et al.* (2013) show that reduced-form estimates of extensive-

[4] A number of studies have extended the bunching framework to allow for adjustment costs; see, for example, Gelber *et al.* (2019), Ruh & Staubli (2019) and Zaresani (2020).

[5] Our paper is methodologically related to Saez (2003)—who uses gradual changes in annual earnings relative to locations of kinks to estimate taxable income elasticity. Similarly, our approach to estimating the counterfactual density is related to Kopczuk & Munroe (2015)—who use data only from the left of a kink, which is not affected by the introduction of a tax.

[6] A large literature studies how optimization frictions affect decisions, including labor supply responses to welfare reform with adjustment frictions (Kline & Tartari, 2016), over-withholding of tax liabilities (Jones, 2012), and information provision and take-up of the earned income tax credit (Chetty & Saez, 2013; Bhargava & Manoli, 2015). In related work, Weathers & Hemmeter (2011) and Ruh & Staubli (2019) find large labor supply responses to return-to-work programs in the US and Austrian disability insurance (DI) systems, and Zaresani (2020) finds small earnings responses to financial incentives in the Canadian DI system.

margin responses often reflect a combination of heterogeneity in reservation wages and labor market frictions such as adjustment costs and hours constraints. In our setting, interpreting the participation response as an intertemporal extensive-margin substitution elasticity requires an assumption of forward-looking behavior. Because the change in effective wages is anticipated and temporary within the tax year, the relevant margin is the retiming of work within the year, and into the next year when the schedule resets, rather than a wealth effect.

This emphasis on high-frequency "stop/continue" decisions is also conceptually related to the literature on taxi drivers' within-day labor supply and quitting decisions in response to transitory earnings opportunities. Classic evidence and interpretations in this taxi-driver literature include Camerer *et al.* (1997) and Chou (2002). Subsequent work develops and tests reference-dependent/target-based and related frameworks, and offers alternative re-analyses emphasizing intertemporal substitution and forward-looking stopping (e.g., Farber 2005, 2008, 2015; Crawford & Meng, 2011; Thakral & Tô, 2021). Ashenfelter *et al.* (2010) use permanent fare increases to study longer-run labor supply responses. Related work in cab/ride-hail settings studies pricing and the value of time (e.g., Buchholz *et al.*, 2025). A recent dynamic optimal-stopping approach that rationalizes many patterns in this literature is Buchholz *et al.* (2026). Similar to this literature, we use the timing of exit decisions to distinguish dynamic extensive-margin stopping from smooth intensive-margin adjustment.

Finally, timing matters for policy design. While tax incentives typically change at year-end, welfare systems often apply earnings tests and benefit offsets at a monthly frequency. The U.S. SSDI program, for example, determines eligibility for DI payments relative to a monthly substantial gainful activity (SGA) threshold. Our evidence is therefore relevant for the proposed "$1 for $2 offset," which would let DI recipients keep $1 of benefits for every $2 of earnings above the SGA threshold. More broadly, our findings suggest that stronger work incentives for marginalized workers with residual work capacity can generate sizable changes in part-time work and aggregate employment.

The remainder of the paper proceeds as follows. Section 2 presents a model in which workers make repeated labor supply decisions over the year. Section 3 describes the policy environment and the data. Section 4 presents the main findings, and Section 5 concludes.

2 Theory and Empirical Methodology

This section develops the paper's core distinction between smooth adjustment and within-year stopping. Our main result is sharp: mid-year stopping requires hours constraints,

whether explicit or implicit. Within a broad class of frictions, a worker who finds it optimal to stop mid-year must either be unable to choose per-period hours flexibly or must face a sufficiently high cost of work in later periods (i.e., a high marginal cost of work at 0 hours of labor supply or a sufficiently large discrete cost) that function effectively as implicit hours constraints. Other frictions do not generate participation effects. We then propose a conditional participation probability that nonparametrically identifies the effect of taxes on the within-year extensive-margin response.

2.1 A Dynamic Model of Inflexible Work and Year-End Incentives

Consider a tax period with T sub-periods, $t = 1, \ldots, T$, discounting the future according to δ. Throughout, a worker's per-period wage is w, and the cumulative labor income is denoted $y_{t-1} = w \sum_{i=1}^{t-1} h_i$. Following Diamond (1980) and Saez (2002), we study a general tax and transfer system with negative taxes at zero labor income, and denote non-labor income $x > 0$. We assume no tax withholding, so that this cash benefit is distributed equally per month, always guaranteeing some income for consumption and allowing us to omit savings decisions from the baseline specification.[7]

Each period, given labor supply choice h_t and cumulative income y_{t-1}, a worker chooses how much to consume, c_t, subject to the constraint, $c_t \leq h_t w \left(1 - \tau \left(y_{t-1}, h_t\right)\right) + x$, where $\tau \left(y_{t-1}, h_t\right)$ gives the tax schedule (i.e., the effective marginal tax rate paid on current-period income). The tax (or equivalently, the benefit reduction) takes the following single-kinked form:

$$\tau \left(y_{t-1}, h_t\right) = \begin{cases} 0, & \text{if} \quad y_{t-1} + h_t w \leq K \\ \frac{\tau(y_{t-1} + h_t w - K)}{h_t w}, & \text{if} \quad y_{t-1} \leq K \quad \text{and} \quad y_{t-1} + h_t w > K \\ \tau, & \text{if} \quad K < y_{t-1} \end{cases},$$

where $0 < \tau < 1$ is a scalar tax rate, $K > 0$ is the location of a kink in the tax schedule, and the middle case describes the effective tax rate when only a portion of the period's earnings are in excess of the kink. We assume the tax rate is perfectly understood. In Appendix A.2, we show that there is no qualitative difference between a tax and a benefit reduction for the purposes of our identification result.

A worker draws her disutility of labor θ from some distribution, and we assume for simplicity it is drawn before making work and consumption choices. A worker of type θ's

[7] In effect, this means that over- and under-withholding is administered the following year.

per-period utility, net of labor costs is

$$u\left(c_t\right) - \theta D\left(h_t\right) - \theta \mathbf{1}_{h_t>0} l_t,$$

where $u\left(\cdot\right)$ is strictly increasing, twice continuously differentiable, and concave. The disutility of work $D(\cdot)$ is also increasing, twice continuously differentiable, strictly convex, and normalized so that $D(0) = 0$.[8] $\left(l_t\right)_{t=1}^{T}$ is a sequence of discrete "fixed costs" of working in each period t. Without savings, a worker will consume her entire income in each period; given cumulative income y_t, her per-period payoff as a function of labor supply h_t is

$$u\left(h_t w\left(1 - \tau\left(y_{t-1}, h_t\right)\right) + x\right) - \theta D\left(h_t\right) - \theta \mathbf{1}_{h_t>0} l_t.$$

We analyze this problem for two types: (i) a flexible and (ii) an hours-constrained worker. We begin by assuming that there are no fixed costs, $l_t = 0$ for all t, before adding them back in.

2.1.1 Analysis for a Flexible Worker (No Fixed Costs)

We begin with a worker who can freely choose hours. In the baseline model without savings, a worker who freely determines hours each period faces no time inconsistency and the problem is effectively static, reducing to choosing a sequence of per-period labor supply decisions, $\{h_t\}_{t=1}^{T}$.

Without tax withholding, if total desired earnings exceed the kink, there exists a period $s \in \{0, \ldots, T\}$ that is the last period before earnings strictly exceed the kink. In period $s+1$, a portion of that period's earnings becomes taxable, but only a partial increase in tax may occur, as cumulative earnings up to period s may still be strictly below the kink threshold. From period $s+2$ onward, all income is subject to taxation. If total earnings exceed the kink starting from period 1 onward, $s = 0$. If total earnings never hit strictly above the kink, $s = T$.

We assume that for all types θ, the condition $wu'\left(x\right) - \theta D'\left(0\right) > 0$ holds, ensuring that all workers find it optimal to work in at least one period within the year. The worker

[8] The leading example of such a cost is the isoelastic utility form, $\frac{\theta}{1+\frac{1}{e}} h_t^{1+\frac{1}{e}}$, where e is the elasticity of labor supply with respect to net after-tax earnings.

chooses $\{h_t\}_{t=1}^{T}$ to maximize

$$\overbrace{\sum_{t=1}^{s}\delta^{t-1}\left(u\left(h_t w+x\right)-\theta D\left(h_t\right)\right)}^{\text{periods strictly below kink}}+\overbrace{\sum_{t=s+2}^{T}\delta^{t-1}\left(u\left(h_t w\left(1-\tau\right)+x\right)-\theta D\left(h_t\right)\right)}^{\text{periods strictly above kink}}+$$
$$\overbrace{\delta^{s}\left(u\left(wh_{s+1}\left(1-\tau\right)+\tau(K-\sum_{i=1}^{s}h_i)+x\right)-\theta D\left(h_{s+1}\right)\right)}^{\text{period crossing the kink}} \tag{1}$$

This problem can be solved in two steps. First, fix $s \in \{0,\ldots,T\}$, then optimize Expression 1 subject to the constraints that

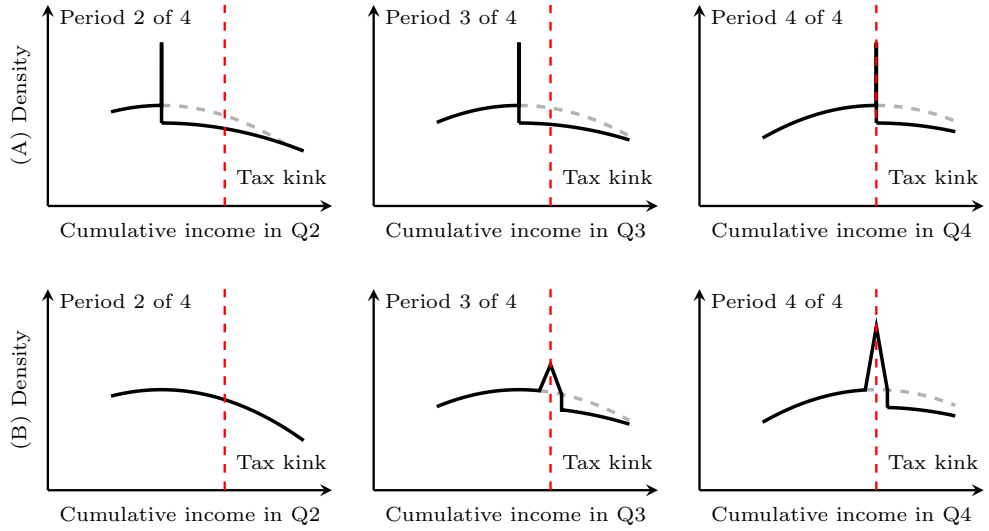
$$w\sum_{t=1}^{s}h_t \leq K, \quad \text{and} \quad w\sum_{t=1}^{s+1}h_t > K.$$

Second, given the solution to this program, optimize the choice of s. Fixing s, the regularity of the objective ensures that an optimizer of Expression 1 exists. As s can take only finitely many values, a solution to the grand (two-stage) problem exists. A special case of this problem is that in which a worker finds it optimal to have her annual earnings be exactly equal to the kink point K. In that case, the worker solves a standard "cake-eating" problem.

Unless the cost of shifting from no work to some work is prohibitively high, a flexible worker will work in every period in order to best "spread out" the convex costs. That is, with no fixed costs, a flexible worker's labor supply h_t must be strictly positive in every period t, unless $D'(0)$ is sufficiently large. If, for instance, the disutility of work at 0 hours is 0, it is optimal to work in every period. We formalize this result as Proposition 1 in Appendix A.1, and show at the end of this section that it also holds when allowing for savings.

Assuming that there exists a smooth distribution of the disutility of work, Panel (A) of Figure 1 illustrates the dynamics of the cumulative earnings distribution for flexible workers with $T = 4$. As workers find it optimal to smooth work across periods due to convex costs, some workers will also find it optimal to bunch their work hours, such that the per-period work decisions accumulate to the annual kink, as seen in the rightmost figure. This set of compliers—i.e., those who find it optimal to bunch—is constant in each period, illustrated by the fixed height of the excess mass, which moves rightward throughout the year.

Figure 1: Illustration of Two Types of Cumulative Earnings Distributions



Notes: Panel (A) illustrates a possible distribution of cumulative quarterly earnings with perfect control of hours worked, and Panel (B) illustrates it for the hours-constrained type. The solid black line denotes the realized density function in response to the tax kink, while the dashed grey line denotes the counterfactual density function that would be observed in the absence of the tax kink. The dashed vertical line denotes the location of the tax kink.

2.1.2 Analysis for an (Explicitly) Hours-Constrained Worker

For an explicitly hours-constrained worker, the key difference is that per-period hours are exogenously specified (from the worker's perspective) at $h_t = h > 0$. In principle, a worker could exert effort to adjust hours—by searching for a different job or negotiating different hours—but when this cost is sufficiently high, the only per-period choice is whether to continue working or quit. We abstract from search as a way to overcome the constraint and model the worker as choosing a stopping period$f \in \{0, 1, \ldots, T\}$, where $f = T$ means the worker does not quit. The quitting decision also comes at a cost $\psi > 0$, which captures its negative effect on the worker's income going into the next year (and other negative consequences).

To avoid trivialities, we assume that $wh < K$ and $u(hw + x) - \theta D(h) > u(x)$, ensuring that a worker finds it optimal to work at least one period. We say that a worker stops working strictly below the kink if $f < T$ and $w(f + 1)h \leq K$. That is, a worker stops working strictly below the kink if she could work at least one more period before her total income exceeded the kink. The conditions which ensure workers find it optimal to work at least one period ensure that workers never stop working strictly below the kink.[9] For some hours-constrained workers, the change in marginal tax rate makes it optimal to stop working when cumulative income crosses the kink. This can lead to mid-year stop-

[9]The difference in annual utilities from a worker stopping at $f + 1 \leq T$ and f has the same sign as $u(hw + x) - \theta D(h) - u(x)$, which is strictly positive, by assumption.

ping—when the costs of quitting are sufficiently low—because hours-constrained workers cannot freely "smooth" their hours like flexible workers can and thus reach the kink before the end of the year.

Panel (B) of Figure 1 illustrates the evolution of cumulative earnings for hours-constrained workers. It shows that hours constraints—in combination with an annual kink—induce some workers to stop working mid-year as their cumulative earnings reach the kink. In that case, the worker finds it optimal to stop working because her take-home wage, $w(1-\tau)$, is too low compared to her disutility of working, $\theta D(h)$. The figure highlights that while the estimated year-end excess mass of earnings may be the same, within-year behavior is key to understanding how hours constraints shape labor supply responses. In particular, some workers reach the desired level of cumulative earnings before period $T = 4$ and stop working mid-year, although hours constraints prevent bunching exactly at the kink, illustrated by the triangular shape of the excess mass.[10]

2.1.3 Analysis for a Flexible Worker with Fixed Costs

Naturally, it may be that there are discrete fixed costs to working. In that case, if a worker works in period t ($h_t > 0$), in addition to her smooth disutility of work $\theta D(h_t)$ she also incurs cost $\theta l_t \geq 0$. Such fixed costs may lead to mid-year stopping for flexible workers.[11] We formalize this result as Proposition 2 in Appendix A.1.

Our crucial observation is that the sufficient condition for stopping mid-year is extremely similar to the condition given in the case of an explicitly hours-constrained worker, with the fixed cost of work playing the role of the quitting cost. It is in this sense that a fixed cost is essentially an implicit hours constraint. Combining our results, flexible workers with sufficiently low marginal disutility of work at 0 and sufficiently low fixed costs of working find it uniquely optimal to work in every period. Thus, mid-year stopping must be the result of hours constraints, explicit or implicit.

2.1.4 Extensions

The model makes several simplifying assumptions: no saving, correct specification of the tax schedule, no random shocks, benefit withholding modeled as a tax, and additively separable utility $u(c_t, h_t; \theta) = u(c_t) - \theta D(h_t)$. We now briefly argue that the key re-

[10] This triangular or hump-shaped pattern would also occur if workers have a distribution of beliefs of the kink location that is centered around the true location (see e.g., Kostøl & Myhre, 2021).

[11] It may be optimal for a worker with fixed costs to stop working strictly below the kink in a mid-year period in which fixed costs are higher. The key identification assumption is that such seasonal variation in fixed costs is independent of the timing at which cumulative earnings approach the kink point.

sult—distinguishing flexible from hours-constrained workers—is robust to relaxing each assumption, deferring benefit withholding and the more general utility $u(c_t, h_t; \theta)$ to Appendix A. To simplify the exposition, we return to the no-fixed-costs specification.

Savings. Suppose that workers may save and, for the sake of contraposition, that one finds it optimal to stop working mid-year. Without loss of generality, we assume that her stopping point is after her annual income has exceeded K. Given a stock of savings $\gamma \geq 0$, and letting $V(\cdot)$ denote the worker's continuation value from carrying the stock of savings into the next year, the worker solves

$$\max_{\{c_t,\}_{t=1}^m, \{f_t\}_{t=1}^m} \sum_{t=1}^{m} \delta^{t-1} u(c_t) + V(f_m),$$

subject to $c_1 + f_1 \leq \gamma + x$ and $c_t + f_t \leq x + f_{t-1}$ for all $2 \leq t \leq m$. Let c_1^* be part of a solution to this program. As it is optimal not to work, it must be that working a little bit and consuming all of it—leaving future behavior unchanged—is suboptimal, i.e.,

$$w(1-\tau)u'(c_1^*) - \theta D'(0) \leq 0,$$

which is false as long as $D'(0)$ is sufficiently close to 0.

Mis-specification. We extend the framework to allow for misspecified workers, assuming that a fraction $\alpha \in (0,1)$ of workers mistakenly believe the kink in the benefits schedule is instead at some $\hat{K} \neq K$. Crucially, a misspecified worker is unaware of this misperception. Consequently, early in the year (for low t), such a worker behaves similarly to a correctly specified worker, except that she optimizes based on the incorrect kink. Suppose a worker learns the correct specification upon exceeding the kink: at that point, she observes a change in the tax, realizes that she is above the kink, and learns its true location. Upon being informed of the correct kink—regardless of how the information is obtained—the worker reoptimizes.

The key question is whether reoptimization can induce a worker to stop working. To address this, we consider the problem of a worker whose misspecification is corrected. One possibility is that the worker becomes aware of her misspecification before exceeding the kink. In this case, her problem is identical to that of a correctly specified worker, with the only difference being that she has a strictly positive endowment (her accumulated earnings) and a shorter time horizon (as fewer months remain in the year). The other possibility is that the worker becomes aware of her misspecification only after exceeding the

kink. Would she then stop working mid-year? The answer is no, as long as the marginal disutility of work at zero hours is close to zero. Moreover, following the logic of Lemma 1 in Appendix A.1—which we use in formalizing our main result as Proposition 1—it is never optimal for a misspecified worker to stop strictly before reaching the kink.

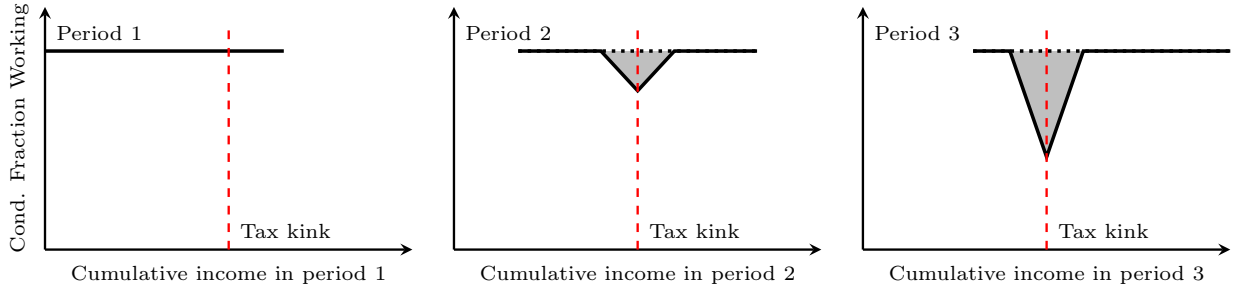
Preference shocks. Finally, extending the framework to include preference shocks, shifting the fixed disutility cost of working throughout the year—for example, due to the presence of young children or health shocks—is straightforward. This extension is empirically relevant, as we shall see that many workers quit in any given month. The main assumption we must invoke is that such preference shocks evolve smoothly over time, and are not correlated with the timing that cumulative earnings enters the vicinity of kinks or notches in the budget constraints.

2.2 Empirical Methodology and Identification

The theory establishes a relationship between the work decision and the distance from cumulative income to the kink in the participation tax rate (PTR): the closer a worker is to the kink, the more likely she is to quit. We operationalize this using a participation probability function (PPF), $p_{e,t}$, which gives the probability of working again by year-end, conditional on working in the current period, as a function of cumulative income. Here, a cumulative income bin is denoted e at within-year time period t, and the lower and upper bound of the response region are denoted L and U.

We denote the PPF as a solid line in Figure 2. For illustrative purposes, the endpoints of the PPF shift rightward over time, reflecting that the support of cumulative earnings expands as the year progresses. Because few individuals have earned above the annual kink early in the year, there is little or no distortion to the PPF in the first period. The distortion becomes visible in the second period and grows into the third as more workers cross the kink and stop working in response. This distortion creates a response region around the kink, denoted by the shaded triangle. The PPF in the response regions falls below the counterfactual PPF, denoted $\tilde{p}_{e,t}$, which captures the expected participation rate in the absence of the kink and is illustrated by the dashed line. As highlighted in the preceding section, this distortion of the PPF near the kink must be the result of hours-constraints, either explicit or implicit.

Figure 2: Illustration of Missing Mass



Notes: This figure displays an example of the actual and counterfactual participation probability functions (PPFs) in the bunching region around the tax kink for period $t = 1$, $t = 2$, and $t = 3$ when $T = 4$. The actual PPF is denoted by the solid black line and represents the probability of working in at least one period between $t+1$ and T conditional on having worked in t. The counterfactual PPF is denoted by the dashed black line and represents the estimated conditional probability of working in the absence of the tax kink. The shaded triangle denotes the response region and the total area of the shaded triangle—the total difference between the actual and counterfactual PPFs—is the mass of missing work, m_t.

Using the counterfactual PPF, the participation response at the kink is identified from the following expression

$$\Delta p = E[p_{e,t} - \tilde{p}_{e,t} | e = K],$$

which in the illustration is equal to the vertical distance between the solid line (actual PPF) and the dotted line (counterfactual PPF). Similarly, the total amount of missing mass of work, m_t, can be estimated from the response region as the following expression,

$$m_t = -\frac{\sum_{e=L}^{U} (p_{e,t} - \tilde{p}_{e,t}) M_{e,t}}{\tilde{p}_{K,t} \tilde{M}_{K,t}},$$

which is equal to the weighted integral over the shaded triangle, where the weights are given by the observed frequency $M_{e,t}$ in each cumulative earnings bin, divided by the counterfactual mass at the kink, $\tilde{p}_{K,t}\tilde{M}_{K,t}$ as in the standard bunching estimation approach. The total number of earnings bins in the response region is given by $(U - L)$.

2.2.1 Identifying the Constrained Share of Compliers

Next, we want to understand the distribution of constrained optimizers, relative to the overall excess mass at the end of the year. The idea of the bunching approach is similar to the dynamic framework we have presented, but assumes individuals make one decision of how many hours to supply in a year. As in our framework, there exists a smooth distribution of the disutility parameter θ, which ensures a smooth distribution of earnings. Only individuals with initial labor earnings in the range $[K, K + \Delta z]$ will find it optimal to bunch exactly at (or near) the kink in response to the introduction of a new tax bracket

with marginal tax rate τ. This behavior generates an excess mass of earnings among compliers at the end of the tax period, denoted b_T.

We can decompose the role of the participation margin in shaping the excess mass by estimating the share of hours-constrained compliers, α. We do so by relating the cumulative missing mass to the excess mass in the following way

$$E[\alpha_t \mid e = K] = \frac{m_{n,t}}{b_{n,T}} = \frac{m_t}{b_T} \cdot \frac{\tilde{p}_{K,t}\tilde{M}_{K,t}}{\tilde{N}_{K,T}},$$

where $m_{n,t}$ is the number of missing workers at time t, $b_{n,T}$ is the number of year-end bunchers, $\tilde{p}_{K,t}\tilde{M}_{K,t}$ is the counterfactual mass at the kink in period t, and $\tilde{N}_{K,T}$ is the counterfactual frequency of individuals locating their period T earnings at the kink. The first equality defines α_t in levels as the number of missing workers relative to the number of year-end bunchers. The second equality rewrites this object using the normalized missing-mass and bunching estimators.

2.2.2 Implementation

We estimate the PPF using a histogram estimator with equally sized bins. The counterfactual PPF is obtained by fitting a constant and a linear slope to data left of the response region. Although the theory implies a flat line, our specification allows for a slope. We then extrapolate linearly from the lower bound of the response region to the kink or the upper bound. For the year-end excess mass of earnings, we allow a more flexible counterfactual following the polynomial approach of Chetty *et al.* (2011) to estimate the excess mass of earnings at year-end.[12]

Standard errors are obtained by block-bootstrapping over individuals with 500 replications, keeping all observations of each resampled individual to account for serial correlation. Within each bootstrap sample, we re-estimate all parameters and define the standard

[12]We follow standard practice and assume a flexible polynomial fitted to the histogram of earnings—excluding the bunching region—which pins down the counterfactual histogram height at the kink $\hat{N}_K$ (see, e.g., Chetty *et al.*, 2011). Specifically, we fit the following function

$$N_j = \sum_{i=0}^{p} \beta_i \left(z_j\right)^i + \sum_{k=z^L}^{z^U} \gamma_k \mathbf{1}\left(z_j = k\right) + \varepsilon_j, \tag{2}$$

where N_j is the number of workers in earnings bin j, z_j is the earnings level, z^L is the lower bound of the bunching region—below K, when people face optimization frictions or make forecast errors—and z^U is the upper bound of the bunching region. Both upper and lower bounds can typically be verified by visual inspection. Flexibly controlling for earnings in this range with the dummies $\gamma_k \mathbf{1}\left(z_j = k\right)$ allows us to estimate the polynomial function without considering the behavioral response to the kink.

error as the standard deviation across replications.

2.3 A Reduced Form Approximation of the Participation Elasticity

The estimated participation elasticity reflects a mixture of two types. Workers constrained by fixed costs or suboptimal hours schedules exhibit a non-zero extensive-margin response, while workers who choose hours freely adjust only at the intensive margin (hence, $\triangle p/\tilde{p} = 0$). Consequently, the observed elasticity equals

$$\varepsilon = \alpha \cdot \frac{\triangle p/\tilde{p}}{\triangle(1-PTR)/(1-PTR)} + (1-\alpha)\cdot 0, \tag{3}$$

in which the weight α can be adjusted to account for the prevalence of hours constraints. To obtain an elasticity for the population of fully (explicitly or implicitly) constrained workers, we divide the "observed" elasticity by the estimated share (α). Another important issue is that this elasticity only captures hours constraints that are binding from below—i.e., those who would like fewer hours than their employer can offer.[13]

3 Policy Environment and Data

This section describes the policy environment and data.

3.1 Policy Environment

Our empirical analysis relies on two institutional features of the Norwegian setting. The first is the policy environment: The tax system contains multiple annual kinks spanning the earnings distribution, and the safety net allows eligible workers to combine cash benefits with earned income, subject to sharp benefit offsets. Chetty (2012) shows that large incentive changes are more likely to induce a response because the welfare cost of ignoring them is high. We therefore begin with Disability Insurance (DI), the largest transfer program besides retirement pension, which allows recipients to combine benefits with earned income. The key advantage of this setting is a large kink where marginal tax rates change by approximately 43 percentage points, at an income level equivalent to about 6 hours of work per week.

[13] Our approach is unable to speak to the role of having multiple employers to reach a desired amount of consumption and leisure.

Disability Insurance and Benefit Offset The Norwegian DI program provides partial earnings replacement to workers aged 18–67 who cannot work due to a lasting impairment. Like other European DI programs, it distinguishes between full and partial disability awards, the latter permitting beneficiaries to use residual work capacity. The working-age DI recipiency rate has been roughly 10 percent throughout the 2000s, with annual spending around $10 billion, or 3 percent of GDP (NAV, 2018). Benefits replace about 66 percent of past earnings up to a ceiling.[14] Once awarded DI, recipients keep their DI receipt until the general retirement age, which is 67 years.

Benefits are paid monthly, as in other countries. As in the US system, fully disabled beneficiaries can earn up to a substantial gainful activity (SGA) threshold without any reduction in benefits. However, while the US Social Security DI system uses a monthly earnings threshold, the Norwegian system applies an annual threshold (see, e.g., Kostøl & Mogstad, 2014 for details). During our sample period, the Norwegian SGA amounts to $8,000 annually; earnings up to that threshold are subject only to income taxation. Unlike the US system, which removes all benefits once monthly earnings exceed the threshold, the Norwegian system phases benefits out gradually. When annual earnings (w) exceed the annual SGA threshold, benefits are reduced at rate τ for every dollar in gross earnings above the threshold.[15] Exceeding the SGA does not affect DI eligibility; however, if annual earnings exceed 80 percent of their pre-disability earnings (PDE), DI benefits are not paid for that year.[16] The relationship between earnings and DI benefits (B) can be summarized as follows, where B_0 is the uncapped DI benefits before any reductions:

[14]The level of DI benefits received is based on an individual's previous earnings and calculates an individual's average indexed annual earnings (AIE). Past wages are indexed to present wages using a deflator equal to the average wage growth in the economy. For awards before 2015, years with the lowest earnings are excluded, and the proportion of income replaced falls with the level of previous earnings. After 2015, AIE is calculated using the highest average income in three of the last five years prior to disability onset. Only income up to a ceiling which is slightly higher than the average wage in the population is included in the calculation of AIE. If AIE is sufficiently low, minimum benefits are provided. The system can also grant partial awards, allowing beneficiaries to use a larger fraction of their residual work capacity.

[15]There are two ways in which the benefit phase-out is administered. If individuals expect to earn above the SGA amount, they can report their year-end earnings expectations directly to the social security administration (SSA), which leads to immediate adjustments to benefits. This is equivalent to tax withholding. The other is by ex post adjustments, allowing for implicit borrowing with repayment the following calendar year. Since 2015, employers report income monthly to the SSA, and the DI system starts phasing out benefits once recipients' cumulative monthly earnings exceed the annual SGA amount. Before 2015, these ex post adjustments were made after the DI program received information from the tax authorities, which was typically around August in the following calendar year.

[16]Pre-disability earnings (PDE) are defined as the individual's normal annual full-time earnings prior to the disability date and are determined by the SSA. For individuals with limited earnings histories, PDE are set to statutory minimum levels. DI recipients are able to exceed 80 percent of PDE for up to five consecutive years and keep their DI receipt.

$$B = \begin{cases} B_0 & if \quad w \leq SGA \\ B_0 - \tau(w - SGA) & if \quad SGA < w \leq 0.8 \cdot PDE \\ 0 & if \quad w > 0.8 \cdot PDE \end{cases} \tag{4}$$

where the implicit tax-rate τ is equal to the replacement rate of DI benefits which is approximately 66 percent for most recipients. Specifically, $\tau = B_0/PDE$. Both earnings and DI benefits are subject to regular income taxes. Accounting for income taxes, the kink more than doubles the effective marginal tax from, on average, 35 to 78 percent, where the individual slope change varies with the replacement rate τ.

Table 1: Kinks in the Income Tax Schedule

Kink point:	\$6,667	\$9,851	\$11,789	\$27,612	\$73,407	\$118,080
mtr	25	8.2	23.6	35.2	44.2	47.2
$\triangle mtr$	25	-16.8	15.4	11.6	9	3

Notes: This table describes the kinks in the schedule of marginal rates for income tax in 2015. The kink locations are reported in the first row in 2015 dollars (USD/NOK = 7.5). The marginal tax rate which applies to income beginning at the kink point is listed in the second row, and the change in the marginal tax rate applied to income to the left and right of the kink—in percentage points—is reported in the third row.

Income Tax Schedule The Norwegian income tax schedule contains several kinks, implying discrete changes in marginal tax rates at specific earnings thresholds. Table 1 summarizes the kink points in the income tax schedule in our sample period and reports the associated marginal tax rates (MTRs) as well as the change in the MTR at each kink point. A notable kink occurs at annual earnings of approximately \$6,667, where the MTR increases by 25 percentage points. Around \$9,851, the MTR falls by 16.8 percentage points, before increasing again by 15.4 percentage points at \$11,789. The income tax schedule then features additional kink points over a wide range of earnings where MTRs increase more gradually, with changes ranging from 3 to 11.6 percentage points.

3.2 Data

The second key feature is the monthly administrative data, which cover every employment relationship—including earnings, contracted hours, hourly wages, and bonuses—and let us distinguish genuine work stopping from relabeling or self-employment reporting. Employers report earnings and cash benefits paid, and the data are updated continuously as new employer files are received. We link these records to monthly files on DI recipients provided by the SSA—which contain the level of DI benefits, pre-DI earnings (AIE), the

date of award, whether the award was for partial or full DI benefits, and the age, gender, and household information of recipients. Finally, we link the sample to Statistics Norway registers on educational attainment, annual earnings, and cash transfers from other welfare programs.

Table 2: Summary Statistics

	A. Workforce		B. DI Recipients	
	Mean	SD	Mean	SD
Column:	(1)	(2)	(3)	(4)
Age	41.0	(12.8)	50.0	(11.6)
Fraction female	.47		.54	
Monthly earnings ($)	5,343	(3,162)	971	(1,247)
Regular pay ($)	3,869	(3,216)	328	(855)
Variable pay ($)	1,062	(1,830)	498	(836)
Fraction regular pay	.70		.25	
Fraction hourly pay	.27		.43	
Fraction fee-based/other pay	.03		.33	
Contracted hours per week	33.4	(12.2)	12.7	(15.3)
Fraction full time	.73		.19	
Fraction part time	.25		.49	
Fraction no contracted hours	.01		.32	
Years of schooling	12.1	(4.1)	10.2	(2.7)
Observations <Individuals>	26,371,841	2,672,063	655,369	47,009

Notes: This table provides summary statistics for the overall population and DI recipients. Panel A reports statistics for the overall (non-disabled) workforce in 2015, restricted to individuals aged 18–66. Panel B reports statistics for DI recipients with DI receipt in 2015 and full DI award prior to 2015. Age and years of schooling are measured in 2015. Outcome variables in Panel A are measured in 2015, while outcome variables in Panel B are measured in 2015–2017. Earnings variables are measured in 2015 dollars (NOK/USD = 7.5).

Table 2 presents summary statistics for the overall workforce (Panel A) and DI recipients (Panel B). Most earnings in the general workforce come from regular pay, while DI recipients earn primarily through variable income alongside DI cash benefits. DI recipients are also older and less educated than the general workforce.

An important feature of the Norwegian administrative data is that all income components except self-employment income are third-party reported (e.g., by employers and financial intermediaries). We restrict attention to employment income to avoid self-employment income, which is reported only annually. Coverage and reliability are rated as exceptional by international assessments (see, e.g., Atkinson *et al.*, 1995). Since administrative data

are a matter of public record, there is no attrition due to non-response or non-consent by individuals or firms, and individuals can only exit these datasets due to natural attrition (i.e., death or out-migration).

4 Evidence

This section asks whether the data exhibit the stopping behavior implied by the model and, if so, how much of the year-end response it explains. We first study the DI kink, where the theory has its best chance of succeeding, and then turn to the income-tax schedule as an external-validity check.

4.1 Responses to the Benefit Offset Kink

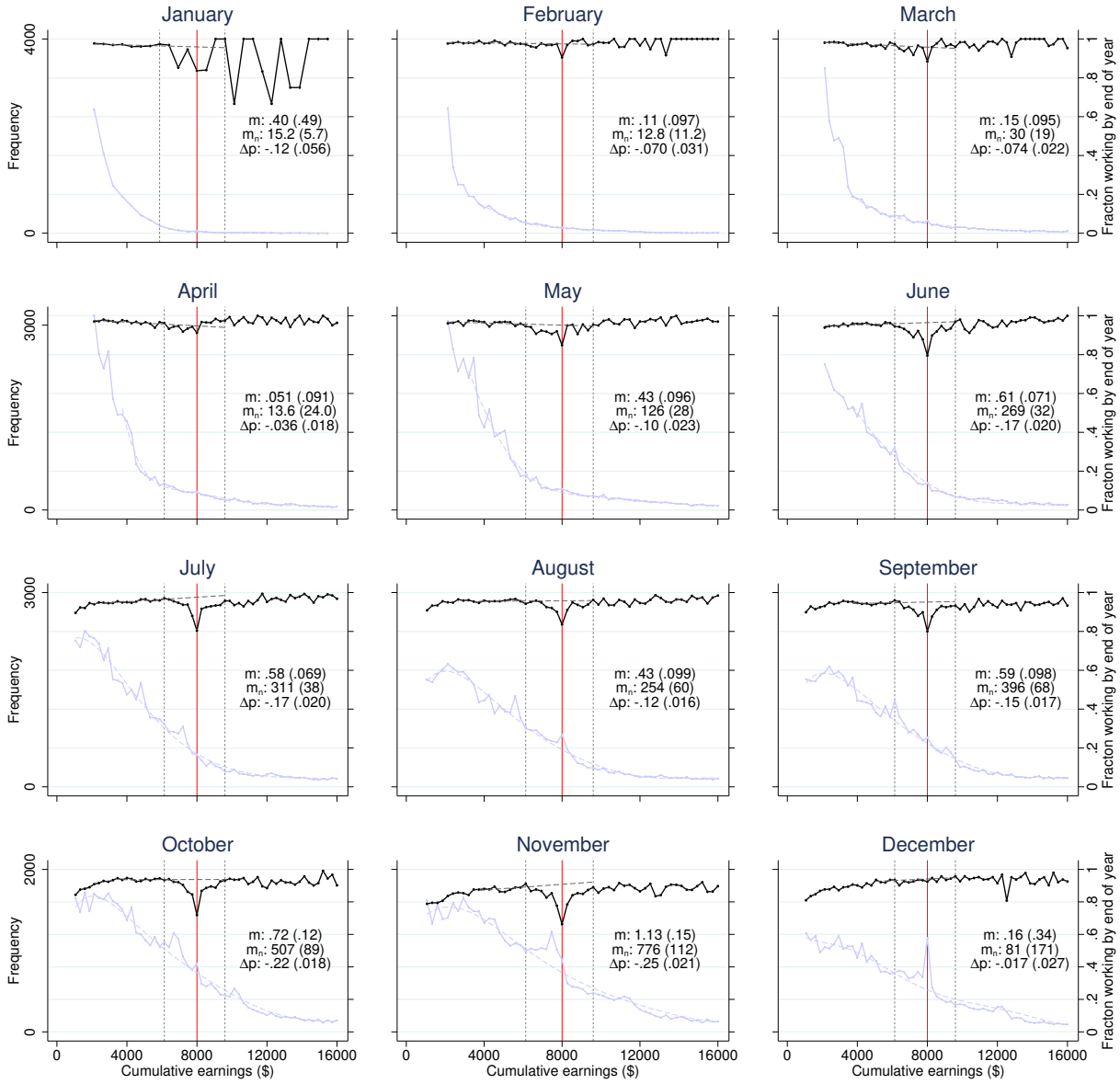
We begin with the DI system, where the benefit offset kink is large and salient. The sample includes all DI recipients with some labor income who receive full benefits at the start of 2015.[17] We study responses to the kink at \$8,000, and measure outcomes during 2015–2017.

4.1.1 Participation Responses

A virtue of our approach is that the conditional work probabilities can be plotted directly. The theory predicts that missing mass around the kink signals hours constraints limiting responses to tax incentives. We test this by estimating the PPF for each month: the fraction of workers active in a given month who work again before year-end, as a function of cumulative earnings through that month. Figure 3 plots these probabilities with solid black lines. We estimate counterfactual work probabilities in each bin using linear extrapolation of the observations to the left of the excluded region and plot the fitted PPF using dashed lines. In the background, we plot the cumulative earnings distributions using solid blue lines, along with the counterfactual earnings distribution using a dashed blue line. The counterfactual earnings distribution is estimated using a seventh-degree polynomial, omitting bins in the excluded region indicated by the vertical dashed lines.

[17] These restrictions are due to a lack of information on the exact location of the kink among the partially disabled and because recipients awarded benefits after this date faced a different set of work incentives. Additionally, we exclude awards from after 2015—as they were subject to a kink of about \$5,000—and those aged above 66—due to eligibility for old-age retirement benefits beginning at age 67.

Figure 3: Missing Mass by Month



Notes: These figures display the observed cumulative earnings distributions for the sample of DI recipients and fitted counterfactual distributions using a 7th degree polynomial as described in Equation 2 for each calendar month (left axes). The figures further display the observed fractions of recipients who work by the end of the calendar year (p) for each earnings bin (using a binwidth of \$267 = 2,000 NOK) and the estimated counterfactual probabilities ($\tilde{p}$) estimated by a linear extrapolation ($i = 1$) using observations to the left of the kink as described in Equation 2 (right axes). The figure for December uses work by the end of the following calendar year as its outcome, following otherwise identical calculations. Missing work (m) is defined as the total weighted difference between p and $\tilde{p}$ in the excluded region around the kink relative to $\tilde{p}$ at the kink, and is weighted by the estimated counterfactual frequency. The excluded region is indicated by the dashed vertical lines. Missing work frequency (m_n) is defined as the total weighted difference between p and $\tilde{p}$ in the excluded region, weighted by the estimated counterfactual frequency. $\triangle p$ is defined as the difference between p and $\tilde{p}$ at the \$8,000 kink. For the full details of the estimated parameters, see Section 2.2. The sample consists of DI recipients aged 18–66 who receive full (100 percent) benefits, have some labor market earnings during the calendar year, and who were awarded DI before 2015. Standard errors (in parentheses) are calculated by bootstrap using 500 replications.

As seen in the top row of Figure 3, very few individuals have earnings close to the

annual $8,000 kink in the first months. Moreover, we see that the PPF estimates are noisy, particularly for cumulative earnings above the kink. In each figure, we report the missing mass, the corresponding frequency of individuals stopping work, and the change in participation rates at the kink, with bootstrapped standard errors in parentheses. Appendix Figure OA.1 reports the dynamics of point estimates with 95 percent confidence bands. The missing mass and quitting behavior are not statistically significantly different from zero in the first four months.

From May onward, the pattern becomes economically and statistically clear. Recipients whose cumulative earnings are near the kink are significantly less likely to continue working than those below or above it, indicating that the annual work incentives induce some recipients to stop as their cumulative earnings approach $8,000. The gap widens later in the year, making this the paper's cleanest evidence of a within-year stopping response.

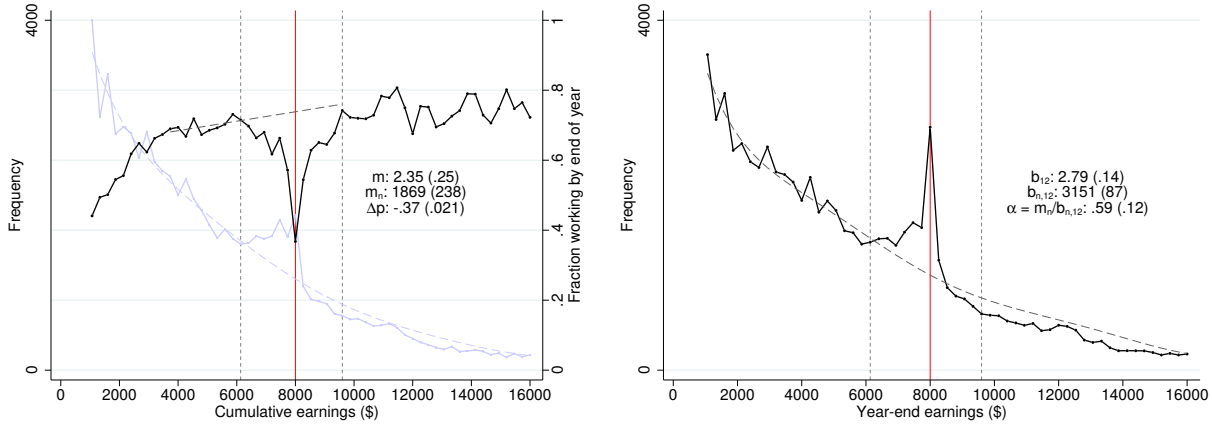
The participation response is small and borderline insignificant through April. From May onward, the fraction who stop when earnings are near $8,000 rises steadily, reaching approximately 0.25 by November. In December—when there are no remaining months in which to work—we instead compute the share of individuals who work again by the end of the following calendar year as a function of cumulative income in the current year. This serves as a placebo test and confirms that the temporary incentive to stop working is driving the participation response, rather than unobserved characteristics of individuals with earnings near the kink.

Aggregating these monthly responses to an annual equivalent requires addressing two issues. First, some workers may quit and restart multiple times within a year. Second, naively pooling all monthly observations would double-count individuals who stop at the kink mid-year and do not resume until January, potentially biasing the missing mass upward.

We address this by using only the last observation for each individual (through November) if they work in multiple months. This captures the individual's final optimization decision within the year and ensures that each person is counted once. We report our main findings of annualized participation responses in Figure 4. Panel (a) shows the observed and counterfactual cumulative earnings distributions (solid blue and dashed lines) alongside the observed and counterfactual continuation probabilities (solid and dashed black lines). There are clear signs of extensive-margin responses: Recipients with cumulative earnings near the kink are markedly less likely to continue working than those below or above it. The estimated missing mass is 2.35—meaning that the total missing work near the kink corresponds to 2.35 times the counterfactual work probability at the kink—and

roughly 1,869 individuals stop working due to the benefit offset.

Figure 4: Missing Mass of Work and Excess Mass of Earnings at DI Benefit Kink



(a) Missing Mass (b) Year-End Excess Mass

Notes: Figure (a) displays the observed cumulative earnings distribution for the sample of DI recipients and a fitted counterfactual using a 7th degree polynomial as described in Equation 2 (left axis). The figure further displays the observed fraction of recipients who work by the end of the calendar year (p) for each earnings bin (using a binwidth of \$267 = 2,000 NOK) and the estimated counterfactual probability ($\tilde{p}$) estimated by a linear extrapolation ($i = 1$) using observations to the left of the kink as described in Equation 2 (right axis). Missing work (m) is defined as the total weighted difference between p and $\tilde{p}$ in the excluded region around the kink relative to the weighted counterfactual $\tilde{p}$ at the kink. The excluded region is indicated by the dashed vertical lines. Missing work frequency (m_n) is defined as the total weighted difference between p and $\tilde{p}$ in the excluded region. $\triangle p$ is defined as the difference between p and $\tilde{p}$ at the \$8,000 kink. Figure (b) displays the observed year-end earnings distribution and a fitted counterfactual using a 7th degree polynomial as described in Equation 2 (left axis). Year-end bunching (b_{12}) is defined as the total difference between the observed and counterfactual distribution in the excluded region around the kink relative to the counterfactual at the kink. Similarly, the number of year-end bunchers ($b_{n,12}$) is defined as the total difference between the observed and counterfactual distribution in the excluded region. For the full details of the estimated parameters, see Section 2.2. The sample consists of DI recipients aged 18–66 who receive full (100 percent) benefits, have some labor market earnings during the calendar year, and who were awarded DI before 2015. Standard errors (in parentheses) are calculated by bootstrap using 500 replications.

Panel A of Table 3 reports the estimated participation responses and implied elasticity. Relative to the counterfactual continuation probability of 0.74, the decline corresponds to a 50 percent drop in participation at the kink, yielding a participation elasticity of approximately 0.80 with respect to the net-of-tax participation rate. The estimates are highly significant, indicating large extensive-margin responses to the benefit offset.

Panel B fits a flexible polynomial to the observed continuation rate, omitting the excluded region. The results are robust: The estimated participation response at the kink is 0.35 and the implied elasticity is 0.78.

Panel C examines whether the results are driven by permanent labor force exits. Restricting the sample to individuals employed in the subsequent year (using data through 2017, limited to 2015–2016 observations), the participation response is 0.44—larger than the baseline of 0.39—indicating that the estimates are not disproportionately driven by permanent exits.

Table 3: Estimates of the Labor Force Participation Elasticity

	Extensive margin response ($\triangle p$)		Labor force participation ($\tilde{p}$)		Participation elasticity (ϵ)		Obs
Column:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A: Work by end of year (2015–2017)	-.37	(.02)	.74	(.02)	.80	(.03)	93,975
B: Specification checks							
i) Fitted polynomial	-.35	(.02)	.72	(.01)	.78	(.03)	93,975
C: Work by end of year (2015–2016)	-.39	(.02)	.75	(.02)	.84	(.04)	68,288
i) Working next year	-.44	(.02)	.82	(.02)	.85	(.03)	50,791
a) In same firm	-.45	(.02)	.84	(.02)	.86	(.03)	46,162
b) In different firm	-.36	(.06)	.77	(.04)	.76	(.09)	10,573
ii) Not working next year	-.20	(.05)	.32	(.04)	.98	(.15)	17,497

Notes: This table displays estimates of labor force participation responses ($\triangle p$) and the corresponding participation elasticity wrt. the (net-of) participation tax rate (ϵ). $\triangle p$ is defined as the difference between the observed probability of working by the end of the calendar year (p) and the estimated counterfactual probability ($\tilde{p}$) at the \$8,000 kink. The participation elasticity (ϵ) is the elasticity of labor force participation wrt. the (net-of) participation tax rate as defined in Equation 3. In Panels A and C, $\tilde{p}$ is estimated by a linear extrapolation ($i = 1$) using observations to the left of the kink. In Panel B, $\tilde{p}$ is estimated by a flexible polynomial as described in Equation 2 using a 7th degree polynomial. The sample consists of DI recipients aged 18–66 who receive full (100 percent) benefits, have some labor market earnings during the calendar year, and who were awarded DI before 2015. Standard errors (in parentheses) are calculated by bootstrap using 500 replications.

The next rows show that the response is similar for workers who return to the same firm but somewhat lower for firm-switchers, consistent with lower adjustment costs for the latter group. Workers who do not return to employment the following year show a smaller participation response (approximately 0.20), consistent with higher quitting costs. However, their baseline participation rate is also much lower (0.32), reflecting weak labor force attachment, which mechanically yields a higher implied elasticity.

4.1.2 Population Shares

Next, we relate our estimates of missing mass to the year-end excess mass of earnings around the kink. We find clear evidence of bunching at the kink by the end of the calendar year. Panel (b) in Figure 4 displays the observed year-end earnings distribution around the kink and the estimated counterfactual probability function, which is estimated using a seventh-degree polynomial, omitting observations in the excluded region. Moreover, we estimate that the excess mass is 2.79, meaning that there are almost three times as many excess individuals with earnings at (or just around) the kink as predicted by the polynomial function (i.e., almost four times the total predicted mass). This estimate is bootstrapped 500 times, and the standard error shows that the excess mass is highly

significant. Weighted by the frequency in the bin, the excess mass amounts to 3,151 individuals. By combining this with our estimate of the number of "missing workers" from Panel (a), we find that the share of constrained individuals is 1,869/3,151 = 0.59, with a standard error of approximately 0.12. Our estimates indicate that more than half of the excess mass of earnings is due to an extensive margin response, rather than an intensive margin response.

Table 4: Missing Mass and Prevalence of Hours Constraints: Kink in DI Program ($8,000)

	Missing mass (m)		Year-end bunching (b_{12})		Fraction hours constrained (α)		Obs
Column:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A: Baseline (2015–2017)	2.35	(.25)	2.79	(.13)	.59	(.12)	93,975
B: Working next year (2015–2016)	2.88	(.23)	2.93	(.17)	.77	(.12)	50,791
i) In same firm	2.91	(.22)	3.01	(.17)	.78	(.12)	46,162
ii) In different firm	3.00	(.62)	2.09	(.34)	1.02	(.44)	10,573

Notes: This table displays the estimated missing work, year-end bunching, and fraction hours-constrained for the sample of DI recipients. Missing work (m) is defined as the total weighted difference between p and $\tilde{p}$ in the excluded region around the kink relative to the weighted counterfactual $\tilde{p}$ at the kink. $\triangle p$ is defined as the difference between p and $\tilde{p}$ at the $8,000 kink. Year-end bunching (b_{12}) is defined as the total difference between the observed and counterfactual year-end earnings distribution in the excluded region around the kink relative to the estimated counterfactual at the kink. Fraction hours constrained (α) is defined as the fraction of missing man-months (m_n) relative to the number of year-end bunchers ($b_{n,12}$). For the full details of the estimated parameters, see Section 2.2. The sample consists of DI recipients aged 18–66 who receive full (100 percent) benefits, have some labor market earnings during the calendar year, and who were awarded DI before 2015. Standard errors (in parentheses) are calculated by bootstrap using 500 replications.

Table 4 reports our estimates of missing mass and the fraction of hours-constrained workers. Panel A presents the baseline estimates shown in Figure 4, with our baseline estimate of the fraction of hours-constrained workers at 59 percent. We report the estimated frequencies in Appendix Table OA.1.

The estimates remain robust when restricting the sample to individuals who only temporarily stop working at the same firm, as shown in row (i) of Panel B. Consistent with theory, we find that quitting work and starting at a new firm the following year is associated with more binding hours constraints. As shown in row (ii) of Panel B, our estimate is effectively 100 percent, though we note that these estimates are considerably noisier. We also estimate a higher fraction of hours-constrained workers—approximately 77 percent—when conditioning on workers who remain in the labor force in the following year, suggesting that the responses are indeed driven by individuals who temporarily exit the labor market rather than those who leave permanently.

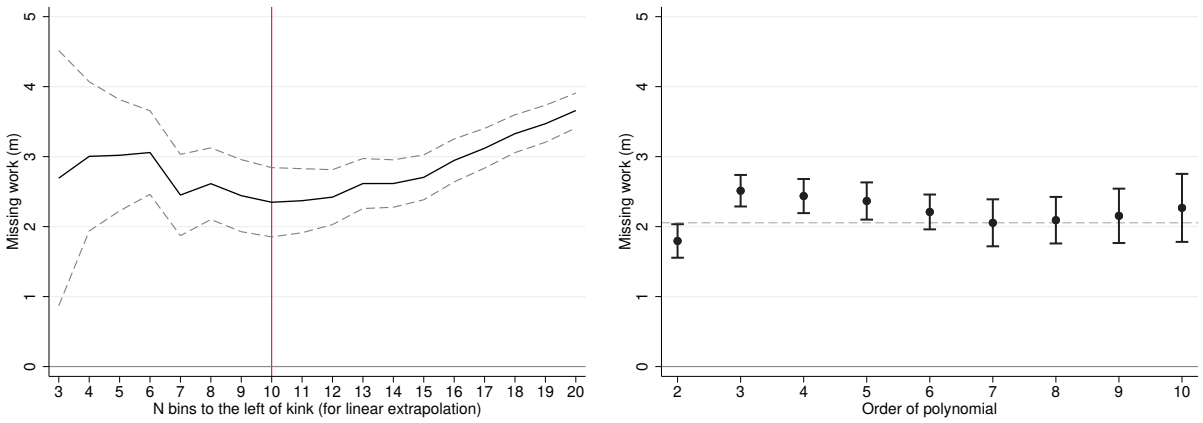
Overall, our evidence shows that the extensive margin is important among marginally attached workers. The substantial missing mass in the earnings distribution, combined

with high estimates of the fraction of hours-constrained workers, suggests that many individuals stop working mid-year in response to the strong benefit offset incentive and inflexible work arrangements.

4.1.3 Robustness

We first verify that our findings are robust to alternative specifications of the counterfactual participation rate. We start by showing that our estimates are quantitatively similar when using a flexible polynomial (omitting observations in the excluded region) to estimate the counterfactual labor force participation rate. We report these findings together with the baseline specification in Appendix Figure OA.2. Next, we change the number of bins we use to the left of the excluded region using the linear extrapolation method. This exercise allows the slope of the counterfactual probability and the estimated responses to differ. The results are reported in Panel (a) of Figure 5. Reassuringly, the estimates are similar. Panel (b) shows that our estimates also remain quantitatively similar if we use the polynomial method and change the order of the polynomial used to fit the counterfactual PPF.

Figure 5: Sensitivity to Counterfactual Specification in Missing Mass Estimation



(a) Linear Extrapolation: Sensitivity (b) Polynomial Method: Sensitivity

Notes: Figure (a) displays the estimated missing work (m) for the sample of DI recipients where the counterfactual labor force participation ($\tilde{p}$) is estimated by a linear extrapolation ($i = 1$) as described in Equation 2 by using different numbers of bins to the left of the kink to estimate $\tilde{p}$ (indicated by the horizontal axis). The red vertical line indicates our main specification using 10 bins to the left of the kink to estimate $\tilde{p}$. Similarly, Figure (b) displays the difference between observed labor force participation (p) and an estimated counterfactual $\tilde{p}$ which is estimated by a n-th degree polynomial omitting bins in the excluded region as described in Equation 2, where the degree of the polynomial (n) is indicated by the horizontal axis. The dashed horizontal line indicates our main specification using a 7th degree polynomial to estimate $\tilde{p}$. Both specifications use a 7th degree polynomial to estimate the counterfactual cumulative earnings density (in order to estimate m). For full details of the estimated parameters, see Section 2.2. The sample consists of DI recipients aged 18–66 who receive full (100 percent) benefits, have some labor market earnings during the calendar year, and who were awarded DI before 2015. Standard errors (indicated by the dashed lines and vertical bars) are calculated by bootstrap using 500 replications.

We also assess the sensitivity of estimates of the excess mass of year-end bunching.

In Panel (a) of Appendix Figure OA.3, we show that estimates fluctuate between 2 and 4 when we change the order of the polynomial used to estimate the counterfactual density of year-end cumulative earnings. However, our estimate of the fraction of hours constrained workers (α) are qualitatively similar when varying the other orders of polynomials as shown in Panel (b), with point estimates ranging between about 0.45–0.75. While these estimates are not statistically significant different from each other, all estimates remain significantly different from zero.

Next we evaluate how differential stopping behavior throughout the year may influence our results. First, some individuals may engage in intermittent employment throughout the year. It is important to note that our approach is agnostic to such stop-start behavior; the key consideration is whether the stopping rate increases as cumulative income enters the response region. Second, some individuals may prefer to start working in February and continue until December, rather than begin in January and stop in December. Such behavior may be rationalized by calendar-month variation in quitting costs—for example, if it is easier to find a job starting in February than in January. While our PPF approach is unable to capture this behavior, this limitation does not introduce bias into our estimates of the participation response. However, it may imply that we are underestimating the share of individuals facing hours constraints.

4.2 Worker and Firm Heterogeneity

We next examine heterogeneity among marginally attached workers. Table 5 reports the participation elasticity, missing mass, excess mass, and the implied fraction of hours-constrained workers for different subgroups. The participation elasticity is substantially larger among older workers than younger ones, consistent with younger workers facing higher career costs of quitting.

Differences across other subgroups are less stark. There are no significant differences by educational attainment or type of work, although the estimates suggest women are less likely to face hours constraints—possibly because fixed costs of work are a more important barrier for them, consistent with evidence that women experience greater workplace flexibility (e.g., Goldin, 2014). We also find some evidence that the response is larger in smaller firms, possibly because larger firms offer more flexibility or career opportunities, raising the cost of quitting.

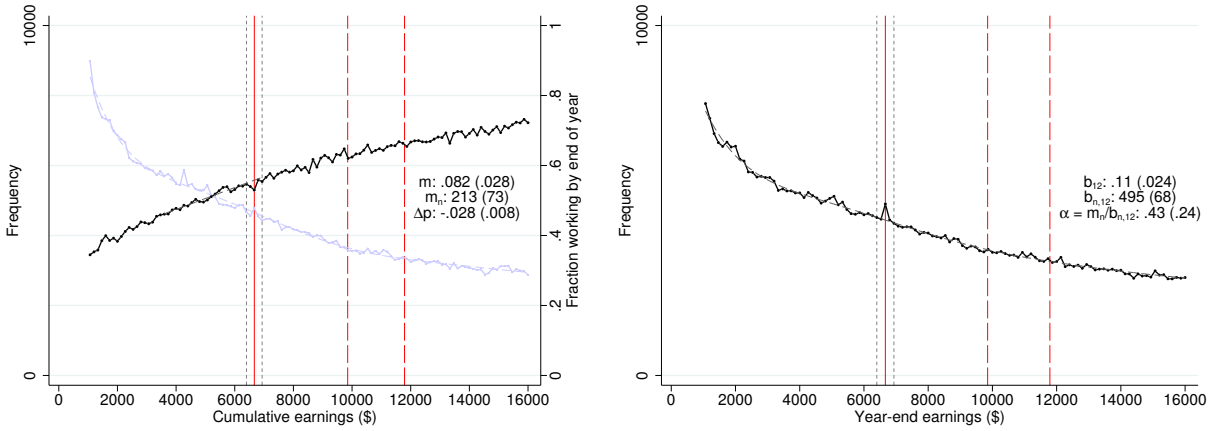
Table 5: Heterogeneity in Missing Mass

	Participation elasticity (ϵ)		Missing Mass (m)		Year-end bunching (b_{12})		Fraction hours constrained (α)		Obs
Column:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
All working DI recipients	.80	(.03)	2.35	(.25)	2.79	(.13)	.59	(.12)	93,975
Age 18–49	.57	(.05)	1.14	(.28)	1.08	(.19)	.76	(.40)	35,128
Age 50–66	.89	(.04)	3.00	(.31)	3.73	(.19)	.56	(.12)	58,847
Male	.85	(.04)	2.58	(.29)	2.72	(.16)	.64	(.14)	43,748
Female	.75	(.05)	2.12	(.34)	2.86	(.18)	.55	(.16)	50,227
No high school	.78	(.05)	2.21	(.26)	2.22	(.20)	.71	(.17)	48,771
High school or more	.83	(.04)	2.57	(.34)	3.41	(.18)	.53	(.13)	45,204
Regular pay	.79	(.05)	2.02	(.35)	3.01	(.26)	.59	(.19)	20,459
Some variable pay	.80	(.04)	2.75	(.33)	3.01	(.18)	.59	(.13)	60,666
Firm characteristics:									
% full-time work: > median	.87	(.04)	3.23	(.31)	3.52	(.21)	.67	(.14)	46,988
% full-time work: < median	.70	(.05)	1.43	(.31)	2.06	(.20)	.48	(.17)	46,987
N employees: ≤ 20	.91	(.04)	3.20	(.43)	4.30	(.21)	.53	(.14)	34,649
N employees: 21–100	.71	(.06)	1.90	(.32)	1.31	(.22)	1.06	(.42)	33,791
N employees: >100	.66	(.07)	1.84	(.44)	2.08	(.25)	.63	(.27)	25,535

Notes: This table displays estimated missing work, year-end bunching, and the fraction hours constrained for different subgroups of DI recipients. For the full details of the estimated parameters, see Section 2.2. The sample consists of DI recipients aged 18–66 who receive full (100 percent) benefits, have some labor market earnings during the calendar year, and who were awarded DI before 2015. Standard errors (in parentheses) are calculated by bootstrap using 500 replications.

Implied Elasticity of Intertemporal Substitution. Having established a strong participation response and that roughly half of compliers face hours constraints, we compute the elasticity in the limiting case where all workers are constrained. Dividing the participation elasticity by the fraction constrained yields an elasticity of 1.35 for the overall sample. If workers are fully aware of the kink location, this estimate can be interpreted as an intertemporal substitution elasticity; if some workers are inattentive, it is a lower bound. This interpretation is consistent with the finding that most workers return to work the following year. The implied elasticity varies across groups: It is higher for older workers, those with at least a high school diploma, and workers in smaller firms.

Figure 6: Graphical Representation of Hours Constraints: First Tax Kink ($6,667)



(a) Missing Mass (b) Year-End Bunching (b_T)

Notes: Figure (a) displays the observed cumulative earnings distribution for the workforce sample (non-DI) and a fitted counterfactual using a 7th degree polynomial as described in Equation 2 (left axis). The figure further displays the observed fraction of workers who work by the end of the calendar year (p) for each earnings bin (using a binwidth of $133 = 1,000 NOK) and the estimated counterfactual probability ($\hat{p}$) estimated by a linear extrapolation ($i = 1$) using observations to the left of the kink as described in Equation 2 (right axis). Missing work (m) is defined as the total weighted difference between p and $\hat{p}$ in the excluded region around the $6,667 kink relative to $\hat{p}$ at the kink, and is weighted by the estimated counterfactual frequency. The excluded region is indicated by the dashed vertical lines. Missing work frequency (m_n) is defined as the total weighted difference between p and $\hat{p}$ in the excluded region, weighted by the estimated counterfactual frequency. $\triangle p$ is defined as the difference between p and $\hat{p}$ at the $6,667 kink. Figure (b) displays the observed year-end earnings distribution and a fitted counterfactual using a 7th degree polynomial as described in Equation 2 (left axis). Year-end bunching (b_{12}) is defined as the total difference between the observed and counterfactual distribution in the excluded region around the $6,667 kink relative to the counterfactual at the kink. Similarly, the number of year-end bunchers ($b_{n,12}$) is defined as the total difference between the observed and counterfactual distribution in the excluded region. In both figures, the red dashed lines indicate (other) kinks in the income tax schedule at $9,851 and $11,789, respectively. For the full details of the estimated parameters, see Section 2.2. The sample consists of the overall workforce (non-DI) aged 18–66 years with some labor market earnings in 2015. Earnings at round numbers in multiples of 1,000 NOK are excluded. Standard errors (in parentheses) are calculated by bootstrap using 500 replications.

4.3 Responses to Kinks in the Income Tax Schedule

We now turn to the income tax system, presenting evidence across the full range of kinks in 2015—from the bottom to the top of the earnings distribution, with marginal tax changes ranging from 3 to 25 percentage points. This exercise is best viewed as a scope test of the mechanism rather than as a second headline result. Figure 6 plots the PPF in Panel (a) and the year-end earnings distribution in Panel (b) for the lowest kink. The counterfactual is estimated as before using a linear trend, excluding the three bins closest to the kink.

At the lowest kink, at $6,667, we find evidence of some missing mass of work. As illustrated by Panel (a) of Figure 6, about 5 percent stop working at the kink. Relating this to the height of the counterfactual distribution at the kink, we estimate an unadjusted participation elasticity of about 0.2. This estimate is reported in the first row of Table 6.

Panel (b) of Figure 6 shows the year-end earnings distribution and the estimated excess mass. Combining the missing and excess mass implies that roughly 43 percent of individuals at this kink are hours-constrained. Dividing the participation elasticity of 0.2

by 0.43 yields an elasticity of 0.47 in the limiting case where all workers are constrained. This is somewhat higher than the estimate of Card & Hyslop (2005) for Canadian welfare recipients, though they consider extensive-margin decisions over a longer time horizon than in our setting.

Table 6: Estimates From The Tax System

	mtr	$\triangle mtr$	Participation elasticity (ϵ)		Missing mass (m)		Year-end bunching (b_{12})	
Column:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Kink point:								
\$6,667	25	25	.20	(.06)	.082	(.029)	.115	(.024)
\$9,851	8.2	-16.8	-.08	(.06)	.020	(.023)	-.059	(.024)
\$11,789	23.6	15.4	.01	(.09)	.017	(.027)	.043	(.025)
\$27,612	35.2	11.6	-.05	(.06)	-.001	(.019)	.009	(.028)
\$73,407	44.2	9	.06	(.02)	.016	(.006)	.021	(.023)
\$118,080	47.2	3	-.12	(.11)	-.008	(.012)	.142	(.055)
Placebo kinks:								
100,000 NOK	23.6	0	-.020	(.010)	.001	(.029)	-.010	(.025)
150,000 NOK	23.6	0	.016	(.007)	.000	(.024)	.087	(.027)
200,000 NOK	23.6	0	.010	(.008)	-.025	(.020)	.016	(.030)

Notes: This table displays the estimated participation elasticity, missing work, and year-end bunching for the workforce-sample (non-DI) for each kink in the income tax schedule and a set of placebo kinks. The participation elasticity (ϵ) is the elasticity of labor force participation wrt. the (net-of) participation tax rate as defined in Equation 3. Missing work (m) is defined as the total weighted difference between p and $\tilde{p}$ in the excluded region around each kink point relative to $\tilde{p}$ at the kink, and is weighted by the estimated counterfactual cumulative earnings frequency. Year-end bunching (b_{12}) is defined as the total difference between the observed and counterfactual year-end earnings distribution in the excluded region around each kink relative to the estimated counterfactual at the kink. For the full details of the estimated parameters, see Section 2.2. The sample consists of the overall workforce (non-DI) aged 18–66 years with some labor market earnings in 2015. Earnings at round numbers in multiples of 1,000 NOK are excluded. Standard errors (in parentheses) are calculated by bootstrap using 500 replications.

Turning to the remaining kinks of the tax system, we plot the participation probability and year-end earnings distribution in Appendix Figure OA.4 and report our findings in Table 6. We find that there are less clear patterns of missing mass of work or excess mass in earnings at the three intermediate kinks between \$9,851 and \$27,612. At the fifth, or second-highest, kink at \$73,407, the fraction continuing to work is nearly flat across the kink, suggesting that the counterfactual can be estimated by fitting a straight line from either side. Comparing the observed drop at the kink to this fitted counterfactual, we find some evidence of missing mass. Although both the missing mass and participation effects are small, they are statistically distinguishable from zero. However, noisy estimates of excess bunching in year-end earnings limit our ability to draw strong conclusions about the overall prevalence of hours constraints. Taken together, the estimates support the view

that career concerns are more salient across much of the earnings distribution, and that the cost of quitting outweighs the cost of not optimizing.

5 Conclusion

In this paper, we study whether annual tax incentives generate within-year stopping responses when workers cannot smoothly adjust hours. We develop a theoretical framework showing that binding hours constraints—whether explicit or implicit—can induce mid-year employment exits as cumulative taxable income enters a higher tax bracket. The conceptual contribution is that monthly participation data can distinguish this stopping mechanism from a standard smooth-hours response to the same annual incentive.

We develop a nonparametric approach to identify within-year participation responses using monthly work decisions and cumulative income. Applied to annual tax kinks in the Norwegian tax and transfer system, the strongest evidence comes from DI recipients, for whom we estimate a participation elasticity of 0.8 and find that approximately 60% of optimizers are constrained. We also find suggestive evidence that these responses are larger among workers in small firms and older workers. Outside that setting, broader tax variation yields much smaller and less precise responses, so we view this paper's main empirical contribution as documenting the mechanism where incentives are large and labor-force attachment is weak.

More broadly, the results highlight the value of monthly data for understanding how annual incentives affect the timing of work decisions within the year. They also suggest that within-year participation responses are most likely to matter in environments with large incentives and workers whose labor supply is organized around lumpy work choices rather than smooth hour-by-hour adjustment.

Appendix A: Theory

This appendix provides proofs and extensions of the benchmark theory in the main text.

A.1 Proof

Proposition 1. *If a flexible worker's marginal disutility of work at* 0, $D'(0)$, *is sufficiently low, and there are no fixed costs, it is uniquely optimal to work in every period.*

Proposition 1 follows from two lemmas.

Lemma 1. *If her total earnings hit strictly below the kink, it is optimal for the flexible worker to work every month.*

Proof. If total earnings hit strictly below the kink, a necessary condition at the optimum is

$$\delta^{t-1}\left[wu'\left(h_t w + x\right) - \theta D'\left(h_t\right)\right] = 0,$$

which is time invariant (as we can just divide out by the discount factor). Accordingly, if total earnings hit strictly below the kink, it is optimal for the worker to work all year. □

Lemma 2. *If a flexible worker's total earnings hit above the kink, it is optimal to stop mid-year only if her marginal disutility of work at* 0, $D'(0)$, *is strictly positive.*

Proof. There are two cases. First, let a flexible worker's total earnings hit the kink exactly, i.e., $w\sum_{i=1}^{T} h_i = K$. We will show that $h_t > 0$ for all t. Suppose for the sake of contradiction not: Suppose first there exists some t and $t' > t$ such that $h_t = 0$ but $h_{t'} = h > 0$. We perturb the worker's payoff as follows. The derivative of

$$u\left(h_t w + x\right) - \theta D\left(h_t\right) + \delta^{t'-t}\left[u\left((h - h_t)w + x\right) - \theta D\left(h - h_t\right)\right],$$

with respect to h_t, evaluated at $h_t = 0$, is

$$wu'\left(x\right) - \theta D'\left(0\right) - \delta^{t'-t}\left[wu'\left(hw + x\right) - \theta D'\left(h\right)\right] > 0,$$

provided $D'(0)$ is sufficiently low, contradicting the optimality of $h_t = 0$. Second, suppose there exists some t' such that $h_t = 0$ for all $t > t'$. Pick an arbitrary $t > t'$. The derivative of

$$u\left(h_t w\left(1 - \tau\right) + x\right) - \theta D\left(h_t\right),$$

with respect to h_t, evaluated at $h_t = 0$, is

$$w(1-\tau)u'(x) - \theta D'(0) > 0,$$

provided $D'(0)$ is sufficiently low, contradicting the optimality of $h_t = 0$.

The second case is that in which a flexible worker's total earnings strictly exceed the kink. Let s ($s \in \{0, \ldots, T-1\}$) denote the last period before earnings strictly exceed the kink:

$$s \coloneqq \max \left\{ t \in \{0, \ldots, T-1\} \ : \ w \sum_{i=1}^{t} h_i \leq K \right\}.$$

By an identical perturbation argument as above, and by the definition of s, it must be the case that $h_t > 0$ for all $t \leq s+1$. But then the derivative of the objective with respect to h_t for any $t \geq s+2$, evaluated at $h_t = 0$, is

$$\delta^{t-1} \left(w(1-\tau) u'(x) - \theta D'(0) \right),$$

which is weakly less than 0 if and only if

$$\theta D'(0) \geq w(1-\tau) u'(x).$$

As the right hand side of this inequality is strictly positive, the result follows. □

Let h^* be the unique solution to

$$wu'(h^* w + x) - \theta D'(h^*) = 0.$$

Then,

Proposition 2. *If there exists $t < T$ such that $wth^* \leq K - wh^*$ and $u(h^* w + x) - \theta D(h^*) - \theta l_{t'} \leq u(x)$ for all $t' > t$, then it is optimal to stop mid-year. If $u(h^* w + x) - \theta D(h^*) - \theta l_T < u(x)$, it is uniquely optimal to stop mid-year*

Proof. The worker's utility from working in the final period is bounded above by

$$u(h^* w + x) - \theta D(h^*) - \theta l_T \underset{(<)}{\leq} u(x),$$

by assumption. Thus, it is (uniquely) optimal to stop in period $T-1$ or earlier. □

A.2 Benefits Reduction

In the text, we remarked that there is little qualitative difference between a tax and a benefits reduction. We now elaborate on this comment. We assume that the tax is implemented via reduction in future benefits as follows. Let the worker be eligible for a nominal annual benefit of Tx, and let

- $y_t = \sum_{i=1}^{t} h_i w$ denote cumulative labor income through period t,
- x_t denote the benefit paid in period t,
- X_t denote cumulative benefits paid through t,
- K denote the maximum cumulative labor income before benefit reduction begins, and
- $\tau \in (0,1)$ denote the claw-back rate applied to income above K.

Define the allowable annual income at time t, A_t:

$$A_t = \begin{cases} y_t + Tx, & \text{if} \quad y_t \leq K \\ K + Tx + (y_t - K)(1-\tau), & \text{if} \quad K < y_t \end{cases}.$$

For all $j \geq t$, given cumulative benefits paid up to that point, X_{t-1}, the per-period benefit x_j is

$$x_j = \frac{A_t - y_t - X_{t-1}}{T - t + 1}.$$

Observe that for all t such that $y_t \leq K$, $x_t = x$. Moreover, it is easy to see that when viewing the tax and benefits schedule from an annual perspective, this setup is identical to that in the text. However, there are subtle differences in our model with month-by-month labor supply decisions. In particular, the direct tax takes place "instantaneously", whereas the benefit reduction is spread out over the subsequent months. This, of course, only makes it more attractive for a flexible worker to work every period as this spreading diminishes the effective penalty to working (due to the concavity of the utility function and impatience).

It is clear that, for a flexible worker, Lemma 1 still holds exactly (the tax and benefits-reduction specifications are identical below the kink). There is also an analog of Lemma 2:

Lemma 3. *If a flexible worker's total earnings hit above the kink and her marginal disutility of work at* 0 *is* 0 *(or sufficiently low), it is never optimal for her to stop mid-year.*

Proof. It is sufficient to prove that the result holds under the condition

$$D'(0) < \frac{w(1-\tau)u'(x)}{\theta T},$$

which, in particular, covers the case $D'(0) = 0$. Suppose for the sake of contradiction that $h_r = 0$ for some r, even though $y_T > K$.

First, consider the case where $h_r = 0$ and $y_{r-1} < K$ (i.e., the zero-work period is before cumulative income hits the kink). Since $y_T > K$, there exists some $j > r$ with $h_j > 0$; specifically, let $j := \min\{i > r : h_i > 0\}$. Choose $\bar{\varepsilon} > 0$ such that $\bar{\varepsilon} < h_j$ and $y_{r-1} + w\bar{\varepsilon} < K$. For $\varepsilon \in (0, \bar{\varepsilon})$, define

$$h_r^\varepsilon = \varepsilon, \qquad h_j^\varepsilon = h_j - \varepsilon, \qquad \text{and} \qquad h_i^\varepsilon = h_i \text{ for } i \notin \{r, j\}.$$

That is, we simply reallocate ε hours of labor from period j to r.

Observe that $h_i^\varepsilon = 0$ for $i = r+1, \ldots, j-1$, so $y_i^\varepsilon = y_{r-1} + w\varepsilon < K$ for all $i = r, \ldots, j-1$. Accordingly, $x_i^\varepsilon = x$ for all $i = r, \ldots, j-1$. Moreover, $y_j^\varepsilon = y_j$, and because $X_{j-1}^\varepsilon = X_{j-1}$, we have $x_j^\varepsilon = x_j$, hence $X_j^\varepsilon = X_j$. Thus, the worker's continuation utility from period $j+1$ onward is unchanged. The right derivative of the worker's value at $\varepsilon = 0$ is

$$\left.\frac{dU}{d\varepsilon}\right|_{\varepsilon=0^+} = \delta^{r-1}\left[wu'(x) - \theta D'(0)\right] - \delta^{j-1}\left[wu'(c_j) - \theta D'(h_j)\right],$$

where $c_j := h_j w + x_j$. Next, we prove that the consumption in the period where work resumes, c_j, exceeds the consumption in the period where no work occurs (i.e., the per-period benefit before cumulative income exceeds the kink, x).

Claim 1. $c_j > x$.

Proof. If $y_j \leq K$, then $x_j = x$ and $c_j = x + h_j w > x$. If $y_j > K$, then $y_{j-1} < K$ implies $X_{j-1} = (j-1)x$, so

$$x_j = x - \frac{\tau(y_j - K)}{T - j + 1},$$

and, therefore,

$$c_j = x + h_j w - \frac{\tau(y_j - K)}{T - j + 1} > x,$$

because $y_j - K < h_j w$ and $\tau < 1$. □

As $c_j > x$, $u'(x) \geq u'(c_j)$. Since $h_j > 0$ and D is strictly convex, $D'(h_j) > D'(0)$. Moreover, our assumed bound on $D'(0)$ is sufficient to ensure that $wu'(x) - \theta D'(0) > 0$.

Therefore

$$\left.\frac{dU}{d\varepsilon}\right|_{\varepsilon=0^+} = \left(\delta^{r-1} - \delta^{j-1}\right)\left[wu'(x) - \theta D'(0)\right] + \delta^{j-1}\left[w\left(u'(x) - u'(c_j)\right) + \theta\left(D'(h_j) - D'(0)\right)\right] > 0,$$

a contradiction. We conclude that the zero-work period cannot occur with $y_{r-1} < K$.

It remains to rule out zero-work periods with $y_{r-1} \geq K$. To do so, we will make repeated use of the following observation:

Claim 2. Take any period s with $y_{s-1} \geq K$. If $y_{s-1} > K$, then for any local perturbation of h_s, $\frac{\partial x_i}{\partial h_s} = -\frac{\tau w}{T-s+1}$ for all $i = s, \ldots, T$.[1]

Proof. For $i = s$, this follows directly from the above-kink formula for x_s. If it holds up to some $i \geq s$, then

$$\frac{\partial X_i}{\partial h_s} = -(i - s + 1)\frac{\tau w}{T - s + 1},$$

and since

$$x_{i+1} = \frac{Tx - \tau(y_{i+1} - K) - X_i}{T - i},$$

we obtain

$$\frac{\partial x_{i+1}}{\partial h_s} = \frac{-\tau w - \frac{\partial X_i}{\partial h_s}}{T - i} = -\frac{\tau w}{T - s + 1}.$$

□

Suppose first that $h_i = 0$ for all $i \geq t$ for some t with $y_{t-1} \geq K$, and let $n := T - t + 1$. Directly,

$$\begin{aligned}
\left.\frac{\partial U}{\partial h_t}\right|_{h_t=0^+} &= \delta^{t-1}\left[w\left(1 - \frac{\tau}{n}\right)u'(c_t) - \theta D'(0)\right] - \frac{\tau w}{n}\sum_{i=t+1}^{T}\delta^{i-1}u'(c_i) \\
&= \delta^{t-1}\left[wu'(x_t)\left(1 - \frac{\tau}{n} - \frac{\tau}{n}\sum_{k=1}^{n-1}\delta^k\right) - \theta D'(0)\right] \\
&\geq \delta^{t-1}\left[w(1-\tau)u'(x) - \theta D'(0)\right] > 0
\end{aligned}$$

where the first equality uses Claim 2; the second equality uses the fact that $c_i = x_t$ for all $i \geq t$, as no further earnings occur and thus the remaining benefit stock is paid evenly over the remaining n periods; and the weak inequality follows because $\sum_{k=1}^{n-1}\delta^k \leq n - 1$ means that the term in round brackets is at least $1 - \tau$, and because $y_t = y_{t-1} > K$ so that $x_t \leq x$ and, thus, $u'(x_t) \geq u'(x)$. Once again, our assumed bound on $D'(0)$ ensures that

[1] If $y_{s-1} = K$, we understand this as a right derivative at $h_s = 0$, and as an ordinary derivative in a neighborhood of any interior point $h_s > 0$.

this final expression is strictly positive. This contradicts optimality, and we conclude that there cannot be a singular period of zero-work after the kink.

Finally, suppose that there is a post-kink block of zeros $h_a = \cdots = h_{b-1} = 0$ with $h_b > 0$. Denote

$$t := b - 1, \qquad \text{and} \qquad m := T - t + 1.$$

Then $h_t = 0$, $h_{t+1} > 0$, and $y_{t-1} \geq K$. Using Claim 2,

$$g_t := \left.\frac{\partial U}{\partial h_t}\right|_{h_t=0^+} = \delta^{t-1}\left[w\left(1-\frac{\tau}{m}\right)u'(c_t) - \theta D'(0)\right] - \frac{\tau w}{m}\sum_{i=t+1}^{T}\delta^{i-1}u'(c_i).$$

Since no earnings occur in period t, we have $c_t = x_t \leq x$, with equality if $y_t = K$. Moreover,

$$x_{t+1} = x_t - \frac{\tau w h_{t+1}}{m-1},$$

so

$$c_{t+1} = x_{t+1} + wh_{t+1} = c_t + wh_{t+1}\left(1 - \frac{\tau}{m-1}\right) > c_t.$$

By concavity of u, $u'(c_t) \geq u'(c_{t+1})$.

Also, as $h_{t+1} > 0$, the period $t+1$ choice is characterized by the FOC (where we applied Claim 2):

$$\delta^t\left[w\left(1-\frac{\tau}{m-1}\right)u'(c_{t+1}) - \theta D'(h_{t+1})\right] - \frac{\tau w}{m-1}\sum_{i=t+2}^{T}\delta^{i-1}u'(c_i) = 0.$$

Since $D'(h_{t+1}) \geq 0$, this implies (multiplying both sides by $(m-1)/m$):

$$\frac{\tau w}{m}\sum_{i=t+2}^{T}\delta^{i-1}u'(c_i) \leq \delta^t w\frac{m-1-\tau}{m}u'(c_{t+1}).$$

Thus,

$$\begin{aligned}
-\frac{\tau w}{m}\sum_{i=t+1}^{T}\delta^{i-1}u'(c_i) &= -\frac{\tau w}{m}\sum_{i=t+2}^{T}\delta^{i-1}u'(c_i) - \frac{\tau w}{m}\delta^t u'(c_{t+1}) \\
&\geq -\delta^t w\frac{m-1-\tau}{m}u'(c_{t+1}) - \frac{\tau w}{m}\delta^t u'(c_{t+1}) \\
&= -\delta^t w\frac{m-1}{m}u'(c_{t+1}).
\end{aligned}$$

Substituting this bound into the expression for g_t delivers

$$g_t \geq \delta^{t-1}\left[w\left(1-\frac{\tau}{m}\right)u'(c_t)-\theta D'(0)\right]-\delta^t w\frac{m-1}{m}u'(c_{t+1}).$$

Using $u'(c_t)\geq u'(c_{t+1})$ and $\delta\leq 1$,

$$g_t \geq \delta^{t-1}\left(w\left[\left(1-\frac{\tau}{m}\right)-\delta\frac{m-1}{m}\right]u'(c_t)-\theta D'(0)\right)\geq \delta^{t-1}\left[\frac{1-\tau}{m}wu'(c_t)-\theta D'(0)\right].$$

Finally, $c_t\leq x$, so $u'(c_t)\geq u'(x)$, and our assumed bound on $D'(0)$ ensures $\frac{1-\tau}{T}wu'(x)>\theta D'(0)$, therefore,

$$g_t \geq \delta^{t-1}\left[\frac{1-\tau}{m}wu'(x)-\theta D'(0)\right]\geq \delta^{t-1}\left[\frac{1-\tau}{T}wu'(x)-\theta D'(0)\right]>0,$$

again contradicting optimality. We thus conclude that there cannot be a block of zero-work periods after the kink.

Having exhausted every possibility, we conclude that $h_i>0$ for every $i=1,\ldots,T$. □

A.3 Relaxing Additive Separability

In the text, we specify the following period t utility:

$$u\left(h_t w\left(1-\tau\left(y_{t-1},h_t\right)\right)+x\right)-\theta D\left(h_t\right).$$

We now verify that the result—flexible workers do not stop mid-year—does not depend on this functional form. Indeed, suppose a worker's period t utility is instead

$$u\left(c_t,h_t;\theta\right),$$

which we assume to be

- continuously differentiable in both arguments;
- strictly increasing in consumption and decreasing in labor: $\frac{\partial u}{\partial c}(c,h;\theta)>0$ and $\frac{\partial u}{\partial h}(c,h;\theta)\leq 0$; and
- concave.

Then, letting

$$U_t=u\left(h_t w\left(1-\tau\left(y_{t-1},h_t\right)\right)+x,h_t;\theta\right),$$

we have an analog of Lemmas 1 and 2:

Proposition 3. *For a flexible worker, if a worker's marginal disutility of work at* 0 *is* 0 *for all* $c \leq x$*, it is not optimal for her to stop mid-year.*

Proof. Suppose earnings hit above the kink and consider a period beyond the kink. Then the worker's marginal utility at $h_t = 0$ is

$$\left.\frac{dU_t}{dh_t}\right|_{h_t=0} = w(1-\tau) \cdot \frac{\partial u}{\partial c}(x, 0; \theta) + \frac{\partial u}{\partial h}(x, 0; \theta).$$

We omit the arguments for the other periods (crossing and pre-crossing), as the logic is identical. □

References

ASHENFELTER, ORLEY, DORAN, KIRK, & SCHALLER, BRUCE. 2010. A Shred of Credible Evidence on the Long-run Elasticity of Labour Supply. *Economica*, **77**(308), 637–650.

ATKINSON, A., RAINWATER, L., & SMEEDING, T.M. 1995. Income Distributions in OECD countries: Evidence from the Luxembourg Income Study. *OECD Publications and Information Center*.

AUTOR, D. H., & DUGGAN, M. G. 2003. The rise in the disability rolls and the decline in unemployment. *Quarterly Journal of Economics*, **118**(1), 157–205.

BHARGAVA, SAURABH, & MANOLI, DAYANAND. 2015. Psychological frictions and the incomplete take-up of social benefits: Evidence from an IRS field experiment. *The American Economic Review*, **105**(11), 3489–3529.

BLACK, DAN, DANIEL, KERMIT, & SANDERS, SETH. 2002. The Impact of Economic Conditions on Participation in Disability Programs: Evidence from the Coal Boom and Bust. *The American Economic Review*, **92**(1), 27–50.

BUCHHOLZ, NICHOLAS, DOVAL, LAURA, KASTL, JAKUB, MATEJKA, FILIP, & SALZ, TOBIAS. 2025. Personalized Pricing and the Value of Time: Evidence From Auctioned Cab Rides. *Econometrica*, **93**(3), 929–958.

BUCHHOLZ, NICHOLAS, SHUM, MATTHEW, & XU, HAIQING. 2026. Rethinking Reference Dependence: Wage Dynamics and Optimal Taxi Labor Supply. *Journal of Political Economy*, Forthcoming.

CAMERER, COLIN, BABCOCK, LINDA, LOEWENSTEIN, GEORGE, & THALER, RICHARD. 1997. Labor Supply of New York City Cabdrivers: One Day at a Time. *The Quarterly Journal of Economics*, **112**(2), 407–441.

CARD, DAVID, & HYSLOP, DEAN R. 2005. Estimating the Effects of a Time-Limited Earnings Subsidy for Welfare-Leavers. *Econometrica*, **73**(6), 1723–1770.

CHETTY, RAJ. 2012. Bounds on elasticities with optimization frictions: A synthesis of micro and macro evidence on labor supply. *Econometrica*, **80**(3), 969–1018.

CHETTY, RAJ, & SAEZ, EMMANUEL. 2013. Teaching the tax code: Earnings responses to an experiment with EITC recipients. *American Economic Journal: Applied Economics*, **5**(1), 1–31.

CHETTY, RAJ, FRIEDMAN, JOHN, OLSEN, TORE, & PISTAFERRI, LUIGI. 2011. Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records. *Quarterly Journal of Economics*, **126**, 749–804.

CHETTY, RAJ, GUREN, ADAM, MANOLI, DAY, & WEBER, ANDREA. 2013. Does indivisible labor explain the difference between micro and macro elasticities? A meta-analysis of extensive margin elasticities. *NBER macroeconomics Annual*, **27**(1), 1–56.

CHOU, YUAN K. 2002. Testing Alternative Models Of Labour Supply: Evidence From Taxi Drivers In Singapore. *The Singapore Economic Review*, **47**(1), 17–47.

COGAN, JOHN F. 1981. Fixed Costs and Labor Supply. *Econometrica: Journal of the Econometric Society*, 945–963.

CRAWFORD, VINCENT P., & MENG, JUANJUAN. 2011. New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income. *American Economic Review*, **101**(5), 1912–1932.

DIAMOND, PETER. 1980. Income taxation with fixed hours of work. *Journal of Public Economics*, **13**(1), 101–110.

FARBER, HENRY S. 2005. Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers. *Journal of Political Economy*, **113**(1), 46–82.

FARBER, HENRY S. 2008. Reference-Dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers. *American Economic Review*, **98**(3), 1069–1082.

FARBER, HENRY S. 2015. Why You Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers*. *The Quarterly Journal of Economics*, **130**(4), 1975–2026.

GELBER, ALEXANDER M, JONES, DAMON, & SACKS, DANIEL W. 2019. Estimating Earnings Adjustment Frictions: Method and Evidence from the Earnings Test. *American Economic Journal: Applied Economics*.

GOLDIN, CLAUDIA. 2014. A grand gender convergence: Its last chapter. *American economic review*, **104**(4), 1091–1119.

HANSEN, GARY D. 1985. Indivisible labor and the business cycle. *Journal of monetary Economics*, **16**(3), 309–327.

JONES, DAMON. 2012. Inertia and overwithholding: explaining the prevalence of income tax refunds. *American Economic Journal: Economic Policy*, **4**(1), 158–185.

KEANE, MICHAEL P. 2011. Labor Supply and Taxes: A Survey. *Journal of Economic Literature*, **49**(4), 961–1075.

KLEVEN, HENRIK, KREINER, CLAUS THUSTRUP, LARSEN, KRISTIAN, & SØGAARD, JAKOB EGHOLT. 2025. Micro vs macro labor supply elasticities: The role of dynamic returns to effort. *American economic review*.

KLEVEN, HENRIK J., & WASEEM, MAZHAR. 2013. Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan. *Quarterly Journal of Economics*, **128**, 669–723.

KLINE, PATRICK, & TARTARI, MELISSA. 2016. Bounding the labor supply responses to a randomized welfare experiment: A revealed preference approach. *American Economic Review*, **106**(4), 972–1014.

KOPCZUK, WOJCIECH, & MUNROE, DAVID. 2015. Mansion tax: the effect of transfer taxes on the residential real estate market. *American economic Journal: economic policy*, **7**(2), 214–57.

KOSTØL, ANDREAS R, & MYHRE, ANDREAS S. 2021. Labor supply responses to learning the tax and benefit schedule. *American Economic Review*, **111**(11), 3733–66.

KOSTØL, ANDREAS RAVNDAL, & MOGSTAD, MAGNE. 2014. How Financial Incentives Induce Disability Insurance Recipients to Return to Work. *American Economic Review*, **104**(2), 624–55.

KRUEGER, ALAN B, & MEYER, BRUCE D. 2002. Labor supply effects of social insurance. *Handbook of public economics*, **4**, 2327–2392.

MAESTAS, NICOLE, MULLEN, KATHLEEN J., & STRAND, ALEXANDER. 2021. The effect of economic conditions on the disability insurance program: Evidence from the great recession. *Journal of Public Economics*, **199**, 104410.

MOFFITT, ROBERT. 1992. Incentive Effects of the U.S. Welfare System: A Review. *Journal of Economic Literature*, **30**(1), 1–61.

NAV. 2018. Nav-ytelsene frem mot 2060 - En forenklet analyse av store penger. *NAV-rapport nr. 1-2018*.

ROGERSON, RICHARD. 1988. Indivisible labor, lotteries and equilibrium. *Journal of monetary Economics*, **21**(1), 3–16.

RUH, PHILIPPE, & STAUBLI, STEFAN. 2019. Financial incentives and earnings of disability insurance recipients: Evidence from a notch design. *American Economic Journal: Economic Policy*, **11**(2), 269–300.

SAEZ, EMMANUEL. 2002. Optimal income transfer programs: intensive versus extensive labor supply responses. *The Quarterly Journal of Economics*, **117**(3), 1039–1073.

SAEZ, EMMANUEL. 2003. The effect of marginal tax rates on income: a panel study of bracket creep. *Journal of Public Economics*, **87**(5-6), 1231–1258.

SAEZ, EMMANUEL. 2010. Do taxpayers bunch at kink points? *American economic Journal: economic policy*, **2**(3), 180–212.

SAEZ, EMMANUEL, SLEMROD, JOEL, & GIERTZ, SETH H. 2012. The elasticity of taxable income with respect to marginal tax rates: A critical review. *Journal of economic literature*, **50**(1), 3–50.

THAKRAL, NEIL, & TÔ, LINH T. 2021. Daily Labor Supply and Adaptive Reference Points. *American Economic Review*, **111**(8), 2417–2443.

WEATHERS, ROBERT R, & HEMMETER, JEFFREY. 2011. The impact of changing financial work incentives on the earnings of Social Security Disability Insurance (SSDI) beneficiaries. *Journal of Policy Analysis and Management*, **30**(4), 708–728.

ZARESANI, AREZOU. 2020. Adjustment Costs and Incentives to Work: Evidence from a Disability Insurance Program. *Journal of Public Economics*, **188**(104223).

Online Appendix

Appendix B: Additional Results

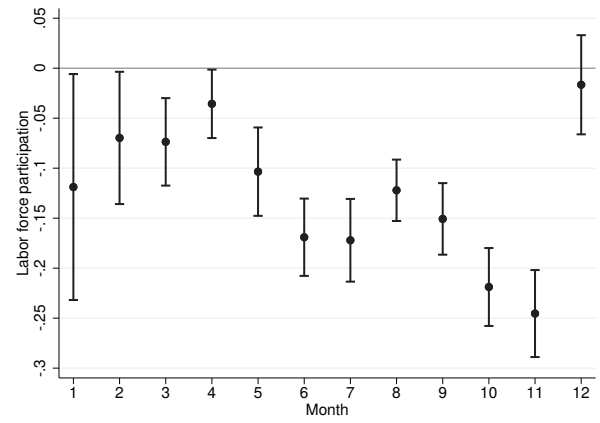
Table OA.1: Missing Mass and the Prevalence of Hours Constraints: Frequencies

	Frequency of missing mass (m_n)		Frequency of year-end bunchers ($b_{12,n}$)		Fraction hours constrained (α)		Obs
Column:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A: Baseline	1,869	(239)	3,151	(80)	.59	(.12)	93,975
B. Work next year	1,583	(154)	2,043	(66)	.77	(.12)	50,791
In same firm	1,513	(145)	1,934	(60)	.78	(.12)	46,162
In different firm	314	(78)	309	(28)	1.02	(.44)	10,573

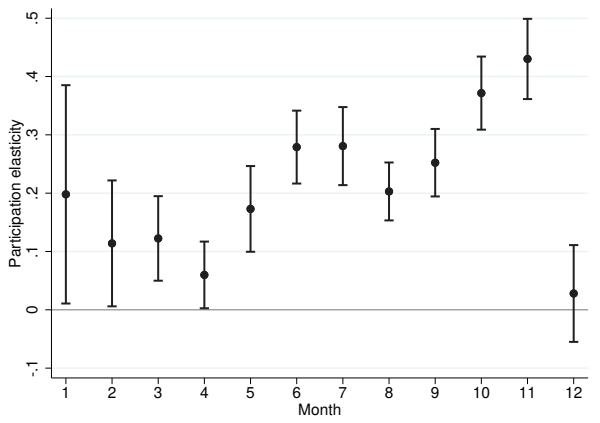
Notes: This table displays the estimated frequency of missing work (m_n), year-end bunching ($b_{12,n}$), and fraction hours constrained ($\alpha = m_n/b_{12,n}$). The sample consists of DI recipients aged 18–66 who receive full (100 percent) benefits, have some labor market earnings during the calendar year, and who were awarded DI before 2015. Standard errors (in parentheses) are calculated by bootstrap using 500 replications.

Figure OA.1: Dynamics in Missing Mass: Benefit Phase-Out Kink in the DI Program

(a) Labor Force Participation (Δp)



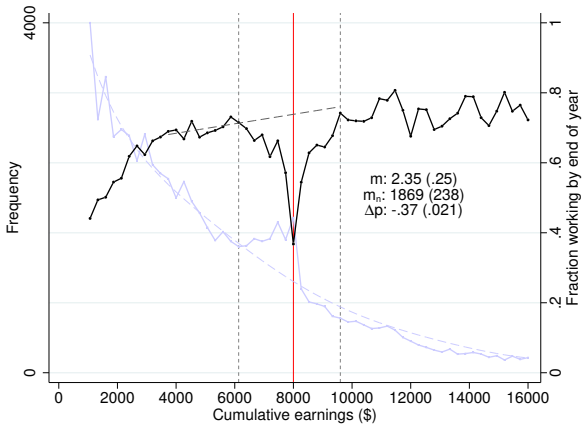
(b) Participation Elasticity (ϵ)



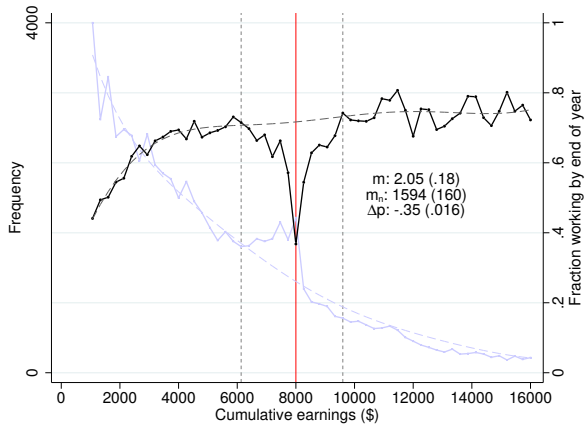
Notes: Figure (a) displays the estimated labor force participation responses ($\triangle p$) for the sample of DI recipients for each calendar month. $\triangle p$ is defined as the difference between the observed probability of working by the end of the calendar year (p) and the estimated counterfactual probability ($\tilde{p}$) at the \$8,000 kink. $\tilde{p}$ is estimated by a linear extrapolation ($i = 1$) using observations to the left of the kink as described in Equation 2. Figure (b) displays the estimated participation elasticity (ϵ) for each calendar month defined as the elasticity of LFP wrt. the (net-of) participation tax rate as defined in Equation 3. The sample consists of DI recipients aged 18–66 who receive full (100 percent) benefits, have some labor market earnings during the calendar year, and who were awarded DI before 2015. Standard errors (indicated by the vertical bars) are calculated by bootstrap using 500 replications.

Figure OA.2: Missing Mass Estimation Methods

(a) Linear Extrapolation



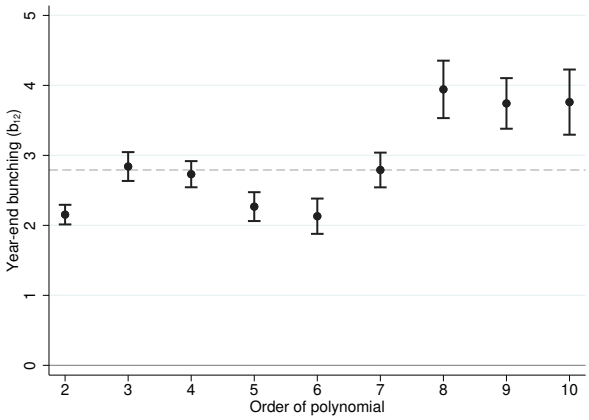
(b) Polynomial Method



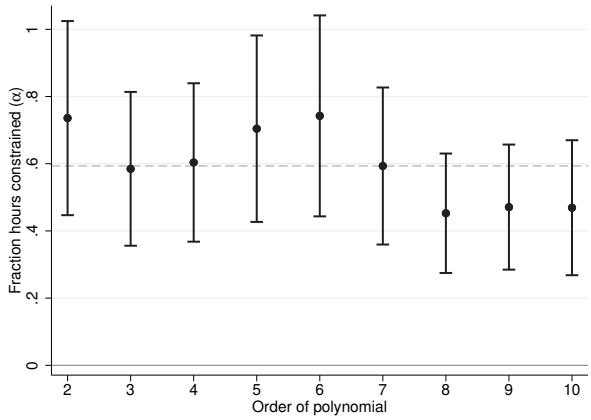
Notes: Figure (a) displays the observed fraction of recipients who work by the end of the calendar year (p) for each earnings bin (using a binwidth of \$267 = 2,000 NOK) and the counterfactual probability ($\tilde{p}$) estimated by a linear extrapolation ($i = 1$) using observations to the left of the kink as described in Equation 2 (right axis). Similarly, Figure (b) displays the observed p and estimated counterfactual $\tilde{p}$ which is estimated by a 7th degree polynomial omitting bins in the excluded region as described in Equation 2. Both figures display estimated parameters using each respective estimation method and the observed cumulative earnings distribution for the sample of DI recipients and a fitted counterfactual using a 7th degree polynomial as described in Equation 2 (left axis). For full details of the estimated parameters, see Section 2.2. The sample consists of DI recipients aged 18–66 who receive full (100 percent) benefits, have some labor market earnings during the calendar year, and who were awarded DI before 2015. Standard errors (in parentheses) are calculated by bootstrap using 500 replications.

Figure OA.3: Sensitivity in Order of Polynomial in Bunching Estimation

(a) Year-End Bunching (b_{12})



(b) Fraction Hours Constrained (α)

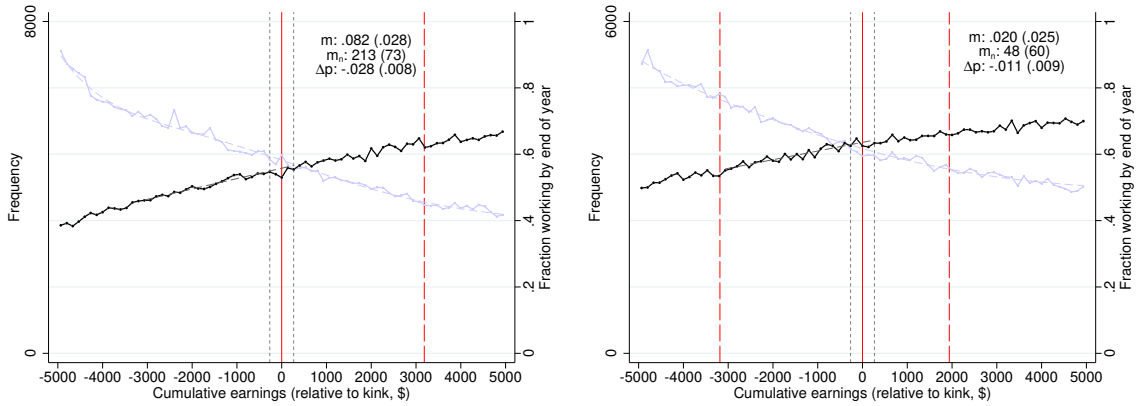


Notes: Figure (a) displays estimated year-end bunching (b_{12}) for the sample of DI recipients where the counterfactual earnings distribution is estimated by a n-th degree polynomial (indicated on the horizontal axis). Similarly, Figure (b) displays the estimated fraction hours constrained (α) for different orders of polynomial to estimate the number of year-end bunchers. For the full details of the estimated parameters, see Section 2.2. In both figures, the dashed horizontal lines indicate our main specification using a 7th degree polynomial to estimate the counterfactual density of year-end cumulative earnings. The sample consists of DI recipients aged 18–66 who receive full (100 percent) benefits, have some labor market earnings during the calendar year, and who were awarded DI before 2015. Standard errors (indicated by the vertical bars) are calculated by bootstrap using 500 replications.

Figure OA.4: Missing Mass at Kinks in the Income Tax Schedule

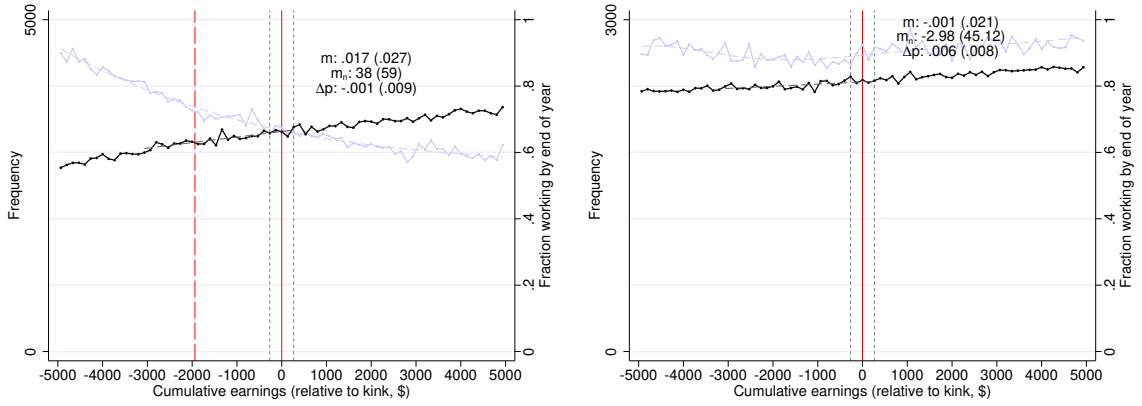
(a) $6,667 Kink

(b) $9,851 Kink



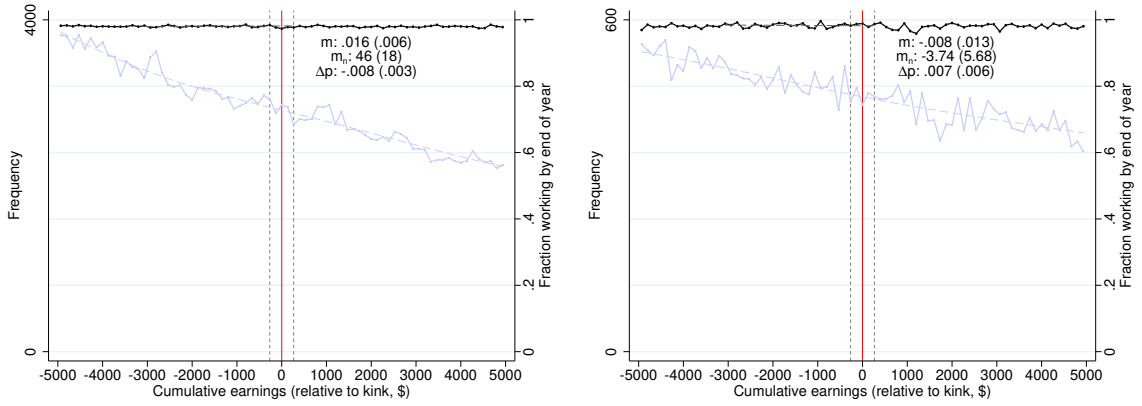
(c) $11,789 Kink

(d) $27,612 Kink



(e) $73,407 Kink

(f) $118,080 Kink



Notes: These figures display the observed cumulative earnings distributions around each kink in the income tax schedule for the workforce sample (non-DI) along with fitted counterfactuals using a 7th degree polynomial as described in Equation 2 (left axes). The figures further display the observed fraction of individuals who work by the end of the calendar year (p) for each earnings bin (using a binwidth of $133 = 1,000 NOK) and the estimated counterfactual probabilities ($\tilde{p}$) estimated by a linear extrapolation ($i = 1$) using observations to the left of the kink. See Figure 4 for more details. The sample consists of the overall workforce (non-DI) aged 18–66 years with some labor market earnings in 2015. Earnings at round numbers in multiples of 1,000 NOK are excluded. Standard errors (in parentheses) are calculated by bootstrap using 500 replications.