NBER WORKING PAPER SERIES

TAX EVASION, EFFICIENCY, AND BUNCHING IN THE PRESENCE OF ENFORCEMENT NOTCHES

Daniel M. Hungerman

Working Paper 28826
http://www.nber.org/papers/w28826

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2021

Thanks to the editor and anonymous referees for helpful suggestions. Thanks also to Teja Konduri and especially to Vivek Moorthy for excellent research assistance. The author declares that he has no relevant or material interests that relate to the research described in this paper. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Tax Evasion, Efficiency, and Bunching in the Presence of Enforcement Notches
Daniel M. Hungerman
NBER Working Paper No. 28826
May 2021
JEL No. H21,H26

ABSTRACT

A recent literature has studied bunching at notches in tax systems; but work on the implications of bunching for welfare has been limited. We consider a setting where there are discrete changes in the enforcement of tax compliance at certain levels of reported income, creating notches that can lead to bunching. We find that greater levels of bunching can be associated with greater tax efficiency. A simulation exercise demonstrates that notches with greater bunching can be associated with higher welfare than notches with less bunching, and that a tax system with bunching at a notch can generate higher overall social welfare than a revenue-equivalent no-evasion linear tax.

Daniel M. Hungerman
Department of Economics
University of Notre Dame
3056 Jenkins-Nanovic Halls
Notre Dame, IN 46556-5602
and NBER
dhungerm@nd.edu

1. Introduction

Many tax systems feature notches that discretely change the levels of the choice sets agents face. A recent literature in economics has studied these notches and the propensity of agents to bunch around them. Work in this area has frequently concluded that bunching in reported (as opposed to earned) income is an important driver of bunching behavior. Kleven (2016), in a survey of bunching results, writes that "in contexts in which bunching is likely to require real earnings responses, the observed amount of bunching is very small (or zero) in elasticity terms," but that "in contexts in which evasion and avoidance responses are feasible, observed bunching can be large."

While bunching in reported income has been shown to be empirically important, less work has explored the implications of this behavior for tax efficiency. This paper shows that the implications can be potentially non-standard. We consider a simple model of taxation where the government enforces (or tries to enforce) truthful reporting up to some level of income (e.g., an amount verified by a third party), and tolerates evasion thereafter. This creates a level of reported income where the benefits of evasion discretely change (cf. Carrillo, Pomeranz, and Singhal, 2017). This is denoted an *enforcement notch* or, simply, a notch.

In the simplest version of the model, a government can, by tolerating evasion around the notch, raise tax revenue with an increase in proportional income tax that creates no distortion; this holds even if the tax in question is distortionary in a no-evasion world, if earned income is unobservable and the government's ability to audit is imperfect, and if the act of evasion creates costs that are pure loss. The key to the analysis is the discrete change in the cost of evasion created by the notch, but if the notch coincides with a change in the marginal tax rate (a tax kink), then the findings are potentially stronger since combining the two effects could induce more taxpayers to bunch at the notch.

We then turn to a setting with individuals reporting income across a range of values above and below the notch, derive a simple expression for the welfare effects of a change in the tax rate, and derive a corresponding optimal tax rate for a given enforcement notch.

The welfare effect of a change in the tax rate differs from that in, for example, Chetty (2009), because here the relevant weights used for taxable income and actual earned income differ across different ranges of reported income. The welfare effects depend critically upon the amount of bunching at the notch, where more bunching implies greater efficiency. A simple simulation exercise demonstrates that for a given distribution of productivity, notches that create greater bunching can be associated with higher welfare than notches with less bunching. The simulations also show that a tax system with an enforcement notch that creates bunching can generate higher overall social welfare than a revenue-equivalent no-evasion linear tax.

An implication of this study is that one cannot use observed bunching in an empirical setting to surmise, even in a first-order sense, the efficiency effects of an enforcement notch. In many settings it is intuitive that more bunching means less efficiency. For example, if bunching in response to taxation reflects real behavior, then more bunching may represent greater elasticities of earnings to taxation (as discussed in Saez, 2010), so that more bunching means taxes are more distortionary. Or, if bunching reflects evasion or avoidance, then more bunching means more evasion/avoidance, and that is usually also regarded as inefficient. The conclusions in this setting are different.

Moreover, the intuition here for how notches can be associated with welfare gains differs from, for example, Slemrod (2013) and Sallee & Slemrod (2012), where the focus is on how the inability of notches to mimic the marginal incentives created by (in their model) optimal linear taxes can lead to welfare losses. This paper makes a not-mutually-exclusive but essentially opposite point, which is that removing marginal incentives will in some settings promote efficiency. In this sense, this paper is somewhat similar to Blinder and Rosen (1985), who note that removing the marginal incentives in a proportional tax system by instead using a notch could be socially beneficial.

But the analysis here differs importantly from Blinder and Rosen's work as well. Their focus primarily concerns inducing demand in a certain good, and they note that bunching could

counteract the welfare benefits they consider. This paper obtains a different result by considering a setting with evasion. But simply introducing evasion in the standard notch/kink framework of taxation will not produce the results here. Lockwood (2020) undertakes such a study; he shows that sufficient-statistic taxation results do not extend to cases where notches are present, and that this result is robust to allowing evasion. His analysis concludes that bunching at notches lowers welfare. The difference is that his notches are generated by tax rates, rather than enforcement.

Other papers have suggested that discrete changes in tax enforcement may be beneficial.[1] For example, Bigio and Zilberman (2011) show that a resource-constrained tax authority might in some cases use a threshold as part of an optimal monitoring strategy, and Reinganum and Wilde (1985) argue that audit thresholds can induce truthful reporting at low cost. The present paper attests to the potential benefits of an audit threshold but, unlike these prior papers, our model does so without appealing to the value of truthful reporting or the costs of enforcement. We thus conclude that the efficiency of a tax system can depend crucially upon how tax compliance is enforced, and moreover that the implications of bunching at a notch can depend importantly on the nature of the notch itself.

The next section illustrates the main results using a simple model of income taxation. Section 3 presents results from a simulation, and the final section concludes.

2. The Model

2.1. The Basic Model

The analysis begins with the simplest case of a distortionary tax and will briefly depict outcomes in the model without a notch; these outcomes follow from prior studies. Consider

[1] A large recent literature has discussed the empirical and theoretical importance of tax evasion and tax enforcement, with work covering topics such as measuring tax evasion (e.g., Artavanis, Morse, Tsoutsoura, 2016), the efficacy of efforts to encourage compliance (Meiselman, 2018), and understanding the costs of evasion (Litina and Palivos, 2016); see Alm (2018) for a survey.

an agent with preferences over consumption c and labor l:

$$u(c,l) = c - \Psi(l) \tag{1}$$

where Ψ is the disutility of labor. Denote the *marginal* disutility of labor as $\psi(l) = \frac{\partial \Psi}{\partial l}$, with $\psi > 0$ and $\frac{\partial \psi}{\partial l} > 0$, so that the marginal cost of labor is positive and increasing.[2] Labor effort earns wage rate w and is taxed at rate t so that preferences are maximized subject to:

$$c = y + wl - twl \tag{2}$$

where y is unearned income. Maximizing (1) subject to (2) produces an optimal choice of labor supply:

$$l^* \;=\; \psi^{-1}(w(1-t)) \tag{3}$$

so that taxes are distortionary. Following (e.g.) Chetty (2009), consider a social welfare function that maximizes the sum of this money-metric utility function and tax revenue:

$$W(t) \;=\; \{y + wl^* - twl^* - \Psi(l^*)\} + twl^*. \tag{4}$$

The term in braces represents maximized utility and it is subject to the envelope condition, but this is not true for the term outside the braces. The change in welfare from changing the tax rate is:

$$\frac{dW(t)}{dt} \;=\; -wl^* + wl^* + tw\frac{dl^*}{dt} \;=\; -w^2 t\frac{\partial \psi^{-1}}{\partial l} \;<\; 0. \tag{5}$$

Now suppose that actual labor income wl is unobserved; instead the agent reports income

[2]Aside from matching the models used in some prior studies (so that the results here can be directly compared), the use of quasilinear preferences in (1) has several benefits: it features money-metric utility, it equates compensated and uncompensated elasticities, and it makes the main derivations transparent. But it is not critical for the main point of the paper and, in the Appendix, the main result is derived for general preferences.

wl_r. We assume to start that the act of evasion does not exhaust any resources, but an agent who reports $l_r \neq l$ faces a probability of detection $p(l - l_r)$, with $p' \geq 0$, $p'' \geq 0$. If detected, the agent must pay unreported taxes plus a fine $F(l - l_r)$; we assume $F' \geq 0$, and $F'' \geq 0$. The penalty function for evasion is:

$$z(l,\ l_r,\ t) \ = \ p(l - l_r)[tw(l - l_r) + F(l - l_r)]. \tag{6}$$

Now the agent solves

$$\max_{l,\ l_r} \ \ y + wl - twl_r - z(l,\ l_r,\ t) - \Psi(l) \tag{7}$$

noting from (6) that $z_l + z_{l_r} = 0$, the first order conditions produce: $l^* \ = \ \psi^{-1}(w(1-t))$, which matches (3): the ability to evade does not affect the distortionary impact of taxes on labor supply. It follows that the ability to evade also does not affect the overall welfare effects of taxation.

Comparing this to the presumably unobtainable outcome from simple lump-sum taxation: suppose that instead of an income tax, the government could impose a lump-sum tax τ that could not be evaded, so the budget constraint (2) became $c = y + wl - \tau$. In this case the optimal choice of labor (denoted $l^{\tau *}$), the welfare function ($W(\tau)$) and its derivative with respect to τ are:

$$l^{\tau *} \ = \ \psi^{-1}(w) \tag{8}$$

$$W(\tau) \ = \ \{y - \tau + wl^* - \Psi(l^*)\} + \tau \tag{9}$$

$$\frac{dW(\tau)}{d\tau} \ = \ -1 + 1 \ = \ 0. \tag{10}$$

2.2. Evasion with Notches

Now we will suppose that the government's efforts to deter evasion change discretely for a certain exogenously given level of income, wl^n. To begin with, and to make the basic intuition as transparent as possible, we assume that if individuals report income such that

$wl_r < wl^n$, that is, $l_r < l^n$, then any evasion is detected with certainty and subsequently they must pay all taxes owed and a fine. If they instead report $l_r \geq l^n$, they pay taxes on reported income and any evasion will go undetected. The cost of evasion thus discretely changes at the notch wl^n.

As this is the central novel feature of the analysis, it merits a few comments. First, the extreme simplicity of the above notch, and in particular the unrealistic assumption that for a certain range of reported income the government can perfectly identify evasion, eases the presentation but will be relaxed momentarily. Second, while the main focus of this analysis is conceptual, it is worth noting that this type of notch resembles assumptions made in some prior work as well as components of real-world tax policy. Regarding the former, Saez (2010) presents his own model of agent behavior to explain his bunching results; similar to here, his model assumes that there is an audit trigger based on whether reported income falls below a certain level and that this trigger acts as a constraint on reported income. Saez's analysis is primarily concerned with matching his empirical results rather than implications for tax efficiency. However, in the Appendix, we derive a similar result with Saez's model and compare it to the result derived here.[3] Keen and Slemrod (2017), who present a model of tax enforcement, also note that many components of enforcement may be non-continuous and give discrete changes in audit probability as an example.

Further, discretely changing the costs of evasion at a certain level of income is an intuitively simple and readily applicable idea. Some real-world tax systems feature such changes; for example the IRS employs the Automated Underreporter (AUR) program which can notify taxpayers if the IRS detects underreporting, for instance by reporting earnings below those reported by a third party (National Taxpayer Advocate, 2017). The results below also involve the idea of agents bunching in reported income at discrete changes in the tax system, and as noted several papers have produced empirical evidence of this behavior—e.g., Car-

[3]See section 2C of Saez's paper for his model. Seaz's model, as well as the analysis motivated by his model in the Appendix, takes third-party reported income as given, but assumes one can vary the reporting of other (called "informal" by Saez) earnings.

rillo, Pomeranz, and Singhal (2017) looking at bunching in reported income at agent-specific values based on third party-reporting, and Almunia and Lopez-Rodriqguez (2018) looking at an economy-wide income-based threshold faced by firms.[4] Here, the amount wl^n could be thought of as applying idiosyncratically to each individual (e.g., the amount reported on a W2 or 1099 form for US income taxation) or a change that coincides with a bracket; in our model of one agent they are similar and the exploration of factors that could create meaningful differences between agent-specific and system-wide notches is a topic left for future work. Since evasion here is not enforced for some ranges of income, the notion of evasion becomes similar to *avoidance*, that is, the legal reduction of taxable income given a certain level of earned income. But the term *evasion* seems more appropriate since the behavior in question crucially will invoke a discrete penalty beyond a certain threshold.

The enforcement notch as described above will matter only for an individual who chooses an earnings level $l^* \geq l^n$. If $l^* < l^n$, the individual will always report truthfully (to avoid the fine) and the analysis in that case follows from section 2.1. Further, the notch here lowers the price of labor l as labor increases, so that the choice set is non-convex and it is possible that an individual would be indifferent to choosing a certain spot above the notch versus below. However, there will never be more than one optimal choice of l^* that is above l^n. For the moment we will assume that the agent optimizes by picking a labor outcome $l^* > l^n$; we consider a case with multiple agents and various choices of l later on.

The agent now maximizes:

$$\max_{l,\ l_r} \quad y + wl - twl_r - z(l,\ l_r,\ t) - \Psi(l) \tag{11}$$

where $z(l,\ l_r,\ t) = tw(l - l_r) + F(l - l_r)$ for $l_r < l^n$, and zero otherwise. Supposing that this individual chooses an optimum $l^* > l^n$, what can we say about l_r? First, the individual will never choose $l_r < l^n$; it would be better to set $l_r = l^*$ and avoid the fine. But the individual

[4]Also related are changes in enforcement based on changes in the use of presumptive taxes or the use of particular tax regimes at certain levels of reported income; Agostini (2016) gives some examples.

will similarly never choose $l_r > l^n$; by lowering reported income the individual can lower their tax payment at no cost. The individual will consequently choose $l_r^* = l^n$.[5]

Given this, we consider the first order condition for actual labor l:

$$w - z_l - \psi(l) \;=\; 0. \tag{12}$$

But for $l^* > l^n$, $z_l = z = 0$, so that (12) becomes:

$$l^* \;=\; \psi^{-1}(w). \tag{13}$$

There will be no deadweight loss from taxation. This does not depend upon the linearity of the tax system; a nonlinear tax function $t(wl_r)$ producing a corner solution in l_r would yield the same equations (12) and (13).

Moreover, for individuals reporting at the notch, this holds if evasion is costly to the taxpayer and if the government's enforcement of tax compliance is imperfect, although both costly evasion and imperfect enforcement could induce an individual to no longer report at the notch and then the effects of taxes could vary. To see this, we will first allow the act of evasion to be costly, and then further allow enforcement to be imperfect.

Suppose first that the act of evasion consumes taxpayer resources $g(l - l_r)$, where g is a positive function increasing in the amount of evasion $l - l_r$. The consumer solves

$$\max_{l,\; l_r} \quad y + wl - twl_r - z(l,\; l_r,\; t) - g(l - l_r) - \Psi(l) \tag{14}$$

with first order conditions

$$l: \quad w - z_l - g_l - \psi(l) \;=\; 0 \tag{15}$$

[5]Similar intuition comes from inspecting first order condition of (11) with respect to l_r, which is $-tw \;-\; z_{l_r} \;=\; 0$. For $l_r < l^n$ the function $z_{l_r} = -tw + F_{l_r}$ so that the left-hand side of this first order condition is always positive. For $l_r > l^n$ the condition becomes $-tw = 0$, so that the left hand side is always negative. The optimal choice of l_r is thus at the notch.

$$l_r: \quad -tw - z_{l_r} - g_{l_r} \;=\; 0. \tag{16}$$

Again the individual will choose to report $l_r \geq l^n$, so that $z = 0$. There are two possible cases. For the first case, the costs of evasion are sufficiently low so that the individual selects $l_r = l^n$. Then (16) does not hold and (15) becomes $\psi(l) + g_l(l - l^n) = w$, which implicitly defines the optimal choice l^* so that taxes are still efficient. The second case is that $l_r^* > l^n$. In this case, lowering l_r incurs costs g_{l_r} that are high enough to discourage so much evasion that the individual reaches the notch. Then the first order condition in (16) becomes $-tw - g_{l_r} = 0$. Noting this and the fact that $-g_l = g_{l_r}$, the first order condition for l in (15) becomes $l^* = \psi^{-1}(w(1-t))$, which matches (3) in the simple distortionary case.

Suppose next that the penalty function is $p(l - l_r)[tw(l - l_r) + F(l - l_r)]$ for $l_r < l^n$ and zero otherwise. The optimization problem is still given by (14), and the first order conditions can still be expressed by (15) and (16), but the there are now three cases to consider. First, if the probability of detection is sufficiently low the agent may choose to evade more income than the amount necessary to reach the notch. Second, the individual may choose to report income at the notch. Third, if the marginal cost of evasion g_l is high enough, then the agent may choose a reported income level above the notch. One can summarize the three cases, and the discussion to this point, as follows:

Proposition 1. *Consider a taxpayer choosing income wl and reported income wl_r, with wage rate w, and facing an enforcement notch wl^n. The possible solutions are:*

A.) Reported income is below wl^n, taxes are distortionary and the optimal choices are given by:

$$wl^* = w\psi^{-1}(w(1-t)), \qquad wl_r^* = wl^* - wg_l^{-1}(tw + z_{l_r}). \tag{17}$$

B.) Reported income is above wl^n, taxes are distortionary and the optimal choices are given by:

$$wl^* = w\psi^{-1}(w(1-t)), \qquad wl_r^* = wl^* - wg_l^{-1}(tw) \tag{18}$$

C.) Reported income equals wl^n, taxes are not distortionary and the optimal choices are

implicitly given by:

$$\psi(l^*) + g_l(l^* - l^n) = w, \qquad wl_r^* = wl^n \tag{19}$$

The solutions in Proposition 1 are derived in the Appendix but follow simply from applying the earlier discussions of (6) & (7) and (12) & (13) to the first order conditions in (15) & (16). Intuitively, the result builds upon an insight from Saez (2001), who notes that in a nonlinear system with discrete changes in the the tax treatment of income, inframarginal components of the tax system can be modeled as wealth effects. The combination of evasion with an enforcement notch produces such a setting.

We make three other observations. First, with an interior solution of l_r in cases A and B in the proposition, the taxpayer equates the gain in tax evasion with the marginal cost of evasion, and an interior solution for l equates the gain in greater l with the gain in greater l_r. But in case C, the notch eliminates the relation between the marginal gain and marginal cost of evasion, so that the tax rate no longer matters at the margin for either choice variable.

Next, the analysis highlights a potentially unexpected feature of a bracket-based tax system, which is that by offering a change in tax rates that coincides with a change in the enforcement notch, one could affect the overall efficiency of the tax system by inducing corner solutions at a notch.[6]

Finally, the proposition shows that the effects of taxation vary by reported income, which is observable to the tax authority. The next subsection applies this result to a setting of heterogeneous agents reporting a range of income amounts, and considers welfare effects of taxation.

[6]More precisely, suppose that the government introduced a tax kink so that for $wl_r \geq wl^n$ the tax rate became $t^n > t$. Based on equation (16), an interior solution for l_r means that, in the pre-tax-kink setting, $-tw - g_{l_r} = -tw + g_l > 0$ when $l_r = l^n$. For a sufficiently high post-kink tax rate—specifically, for $t^n \geq g_l(l - l^n)/w$—it will be possible to (all else equal) induce this person to report income at the notch.

2.3. Heterogeneity and Welfare Effects

A tax authority with an enforcement notch would likely observe individuals reporting different levels of income, so that all three of the above cases could apply. Suppose then that a tax authority, with an enforcement notch at wl^n, collects tax revenue from a population of individuals whose size is normalized to unity, and suppose that individuals' disutility of labor can now be given by $\Psi(l, \alpha)$, where α is an exogenous parameter that affects the disutility of labor. The marginal disutility of labor is $\psi(l, \alpha)$. Based on the α that individuals privately observe, individuals choose a level of labor l and report income wl_r to the tax authority. This is essentially a special instance of the more general set of solutions given in (17) (18) (19) in the prior section, so the solutions there continue to apply for a given individual.

We will make no assumptions about how exactly α maps into optimal solutions of l^* and l_r^*, as it turns out that such assumptions are not necessary for deriving the welfare effects of taxation.[7] We will however assume (a) for a given α all individuals will choose to report income under the notch, on the notch, or above the notch (ie those with a given α do not "split" across these solutions) and (b) that any change in taxes is sufficiently small so that it does not induce people to switch from (e.g.) reporting income under the notch to reporting income above the notch.[8]

The distribution of α across individuals is given by $\mathcal{H}$, so that $\text{Prob}(\alpha \leq x) = \mathcal{H}(x)$. The probability mass function (PMF) of α is given by $h(\alpha)$, representing the mass of individuals associated with a given disutility level α. We will denote the distribution of αs associated with individuals reporting income under the notch as $h^u(\alpha)$. That is, $h^u(\alpha) = h(\alpha)$ for all αs associated with reported income under the notch, and zero for all other αs. Similarly, let the distribution of α for individuals at the notch be $h^n(\alpha)$, and the distribution of α for

[7] Moreover, even if we specified a particular functional form for $\Psi(l, \alpha)$, the relationship between α and the optimal choices of l and l_r would still depend the (unspecified) cost-of-evasion function g.

[8] In Appendix Section A5 we consider an example where individuals initially reporting income above the notch respond to an increase in the tax rate by bunching. Individual utility falls by more when a notch necessitates bunching after a tax increase, but in this case the presence of the notch yields greater social welfare. The intuition here can thus extend to the case of such switchers. Moreover, the numerical simulations in Section 3 also allow for switching in reported income above, below, or onto the notch.

reported incomes above the notch be $h^a(\alpha)$.

The PMF distributions $h(\alpha)$, $h^u(\alpha)$, $h^n(\alpha)$, and $h^a(\alpha)$ are all unobservable to the tax authority. The observable distribution of reported incomes is given by $\mathcal{G}(x)$, so that $\text{Prob}(wl_r < x) = \mathcal{G}(x)$. Let the fraction of individuals reporting taxable income at the notch be given by β. Then $\mathcal{G}(wl^n)$ is the fraction of individuals who report income at or below the notch, and removing β from this group $\mathcal{G}(wl^n) - \beta$ leaves the fraction reporting strictly under the notch. Lastly $1 - \mathcal{G}(wl^n)$ is the fraction reporting strictly above the notch. Thus the distribution of α and reported income are related, as $\sum h^n(\alpha) = \beta$, and $\sum h^u(\alpha) = \mathcal{G}(wl^n) - \beta$, and $\sum h^a(\alpha) = 1 - \mathcal{G}(wl^n)$.

Let $W(t)$ denote the social welfare from taxing a given individual:

$$W(t) = \{y + wl^* - twl_r^* - z(l^*,\ l_r^*,\ t) - \Psi(l^*, \alpha) - g(l^* - l_r^*)\} + z(l^*,\ l_r^*,\ t) + twl_r^* \quad (20)$$

where again braces indicate that the envelope theorem applies. The social welfare function for all individuals can then be given as:

$$\mathbb{W}(t) = \sum_{l_r < l^n} W(t)h^u(\alpha) + \sum_{l_r = l^n} W(t)h^n(\alpha) + \sum_{l_r > l^n} W(t)h^a(\alpha). \quad (21)$$

Denote taxable income wl_r^* as TI for each individual, and wl^* as LI (labor income), and denote the weighting term $\mu = (tw - g_l)/tw$. Then the following proposition shows the marginal effect of a change in the tax rate on social welfare:

Proposition 2. *Let the fraction of individuals below, above, and bunching at the notch l^n be given by $\Big(\mathcal{G}(wl^n) - \beta\Big)$, $\Big(1 - \mathcal{G}(wl^n)\Big)$, and β, respectively. Then the change in social welfare from a change in the tax rate is:*

$$\frac{d\mathbb{W}(t)}{dt} = t\Big(\mathcal{G}(wl^n) - \beta\Big)E\Big[\ \mu\frac{dLI}{dt} + (1-\mu)\frac{dTI}{dt}\ \Big|\ l_r < l^n\Big] + t\Big(1 - \mathcal{G}(wl^n)\Big)E\Big[\ \frac{dTI}{dt}\ \Big|\ l_r > l^n\Big] \quad (22)$$

The derivation is given in the Appendix. Expression (22) is the sum of two terms, one for the

group of individuals reporting income under the notch and one for the group of individuals reporting above. Each term depends upon the overall size of the group, and the group's expected value of the welfare effect from a change in the tax rate. Further, the welfare effects depend upon how taxable income TI and, for one group, labor income LI respond to the change in the tax rate.

The expression (22) is shrinking (getting closer to zero) in the fraction β of individuals who report income at the notch. All else equal, the welfare costs of increasing taxes are smaller when bunching is greater. An important takeaway from (22) is thus that one cannot use observed bunching in this empirical setting to surmise the efficiency effects of an enforcement notch. An increase in bunching holding all else equal is similar to what Lockwood (2020) calls the aggregate-bunching response but (as noted before) his analysis considers a different type of notch and suggests that efficiency falls as bunching increases.

Intuitively, the proposition is like an example of the theory of second best: given a distortionary tax, adding another distortion (from the enforcement notch) may be efficient (I thank a referee for noting this).[9] Figure 1 illustrates the intuition of the welfare effects for an individual who chooses earnings above the notch and facing a tax increase from t to $t^{'}$. Increases in after-tax income above l_r are nonlinear because of $g()$.[10] The initial optimal choice of this consumer, for tax rate t, is given at bundle A, with associated reported taxable income l_r^A.

Suppose that taxes increased to $t^{'}$, but that the individual continued to report the original amount of taxable income. Then the individual's budget line would be given by the new lower line with slope $1-t^{'}$ below reported income and the dashed line that has a kink at the original l_r^A. Above reported income, this dashed line is parallel to the original curved line–it is akin to a lump sum tax. However, an increase in the tax will change taxable income to $l_r^{A'}$. (Taxable income will fall; this follows from equation (16)). For a "large" change in t

[9]Similar intuition can also apply to situations with tax evasion and *expenditure* decisions; Hungerman (2014) considers such a case.

[10]Both l and l_r are chosen together; in this figure different values of real income $l > l_r$ are depicted for a given value of reported income l_r.

the new choice of labor income will also be distorted at A', although for small changes in t this behavioral response does not impact welfare from the envelope theorem. The change in taxable income from the move in l_r however is not negated by any envelope-style argument and this is the source of distortion for these individuals in (22).

One might be interested instead in an increase in bunching that reflected an increase in income elasticities. Greater elasticities would likely increase both β and the $\frac{d\text{LI}}{dt}$ $\frac{d\text{TI}}{dt}$ terms, making the overall relation between bunching and efficiency ambiguous. Moreover, policy-induced changes in the amount of bunching might be generated from (a) moving the location of a notch or (b) eliminating a notch, and in both of these cases it is again unclear whether the intuition of (22) could continue to hold. We undertake an empirical exercise that explores these extensions below. The simulation also allows changes in both reported and actual labor supply, reflecting both possible sources of distortion shown in Figure 1.

Next, the expression (22) resembles the derivation in Chetty (2009) (see his equation 23), where the welfare effects of taxation are a weighted average of taxable and earned income and the weights depend upon the ratio of the marginal private cost of evasion g_l and the tax rate t. As in that paper, the weights for the group below the notch occur because these individuals do not equate the marginal cost of hiding, g_l, with the tax rate. But here, and unlike in Chetty's derivation, the role of these weights depends critically upon whether reported income is above or below the notch. The difference between the two models comes from the asymmetric treatment of the transfer function z. Notably, however, if the tax authority were able to estimate the change in labor income and weights μ, it would be straightforward to apply the results in Proposition 2 across individuals since the expression differs in taxable income, which is observed. The expression in Proposition 2 also obtains, for high income earners above the notch, the result in Feldstein (1999) that welfare is a function of taxable income alone, and this holds even in the presence of transfers and costly evasion and no matter what the marginal cost of evasion is. However, the result here is different in the sense that in this setting evasion can be associated with an increase in taxes that creates

no deadweight loss, even though for the *marginal* dollar earned the evasion rate is 100%.

2.4. The Optimal Tax Rate

This normative result extends to deriving an optimal tax rate. Suppose that a dollar of government spending is worth $1+\phi$ dollars of private income. Assume $\phi > 0$, so that social welfare matches (21) except each individual's tax revenue is weighted by ϕ.

$$\tilde{W}(t,\phi) = \{y + wl^* - twl_r^* - z(l^*,\ l_r^*,\ t) - \Psi(l^*,\alpha) - g(l^* - l_r^*)\} + z(l^*,\ l_r^*,\ t) + twl_r^*(1+\phi) \quad (23)$$

Using (23) in the social welfare expression (21),[11] consider the choice of t that sets the first order condition to zero.[12] Let $\varepsilon_t^i = -\mathrm{E}[d\mathrm{TI}/dt][(1-t)/\bar{\mathrm{TI}}^i]$ represent the elasticity of average taxable income with respect to the tax rate, evaluated for group $i \in \{u, n, a\}$; $\bar{\mathrm{TI}}^i$ is average taxable income in i. The elasticity term ε_ℓ^i is the same but uses labor income instead of taxable income. The taxable income elasticity for all individuals is given by $\varepsilon_t = -\mathrm{E}[d\mathrm{TI}/dt](1-t)/\bar{\mathrm{TI}}$, where $\bar{\mathrm{TI}} = \mathrm{E}[\mathrm{TI}]$. Lastly, define θ_t^i as the ratio of taxable income for group i over taxable income for all taxpayers and θ_ℓ^i as the ratio of labor income for group i over taxable income for all taxpayers. The following proposition then gives the optimal tax rate:

Proposition 3. *The optimal tax rate t^* that maximizes social welfare is given by:*

$$\frac{t^*}{1-t^*} = \frac{\phi}{\left(\mu\ \varepsilon_\ell^u\ \theta_\ell^u + (1-\mu)\ \varepsilon_t^u\ \theta_t^u + \varepsilon_t^a\ \theta_t^a + \phi\ \varepsilon_t\right)} \quad (24)$$

The proof is in the Appendix. This optimal tax equates the marginal benefit of a dollar of revenue with its marginal cost. The marginal cost depends upon how distortionary taxes

[11] Weighting the penalty function by ϕ complicates the analysis but does not change the intuition.

[12] We thus allow endogenous responses to changes in t (outside of the braces) and allow individuals to reach corner solutions in taxable income (ie bunch at the notch), but maintain the assumption that small tax changes do not induce individuals to start or stop bunching in taxable income. Below we consider a simulation that allows agents to change their labor supply across the notch as tax rates change.

are across different groups of income, weighted by the income shares of these groups. The final term in the denominator, weighted by ϕ, which does not have a corresponding term in the welfare expression in (22), reflects the fact that changes in taxation not only affect distortion but also directly affect tax revenue.

The greater the fraction of income reported by those bunching at the notch, all else equal, the smaller the share of income in the non-notch groups, so that the θ terms in the denominator will shrink and the optimal tax rate will grow. Suppose $\mu = 0$ and that for those above and below the notch $\varepsilon_t^i = e$; these hold if the marginal cost of evasion equals its private benefit for those under the notch and if all taxpayers away from the notch have the same tax-price elasticity in taxable income. Then the Appendix shows that (24) simplifies further to:

$$\frac{t^*}{1-t^*} = \frac{\phi}{e(1-\theta_t^n)(1+\phi)} \tag{25}$$

where θ_t^n is the ratio of taxable income for those at the notch over taxable income for all taxpayers. The larger is the share of income from those bunching, the larger is θ_t^n and the higher is the optimal tax rate as the marginal cost of increasing the tax rate declines. With no bunching, $\theta_t^n = 0$ and the right-hand-side expression simplifies to the familiar $\phi/[e(1+\phi)]$.[13]

The results of this section thus show that higher levels of bunching can be associated with smaller-in-absolute-value welfare costs of taxation and higher optimal levels of taxes. The following section considers an empirical exercise that (a) illustrates this intuition in a simple setting (b) does so while allowing individuals to relocate around a notch in response to a change in the tax rate (c) compares the use of a notch to a no-notch no-evasion world and (d) for a given group of taxpayers, compares different notches that create different amounts of bunching.

[13]Real-world estimates of θ_t^n could generate meaningful changes in the optimal tax condition in (25). Best et. al (2015) show that low-rate firms in their data display double the density they should near a kink (apparently through evasion), with roughly 25% of firms locating near the kink. See panel B of their Figure 3, and row 2 of Table 2. If $\theta_t^n = .25$, (25) would consequently be 1/(1-.25) =1.33 times larger in size than in a no-evasion case where $\theta_t^n = 0$. They find even larger bunching behavior for some other types of firms.

3. A Simulation

We illustrate and expand on the above results with a simulation. Preferences are given by iso-elastic utility $u(c,l) = c - \frac{\alpha}{1+1/e}\left(\frac{l}{\alpha}\right)^{1+1/e}$. The parameter e is the elasticity of labor with respect to $(1-t)$ in the no-notch, no-evasion world.

We proceed in several steps. First, for each of 1,000 individuals we draw a parameter of the disutility of work α from a log-normal distribution where the underlying normal distribution has mean $\mu = 0$; and standard deviation $\sigma = .5$. Second, we set the wage rate to unity and for a given notch we calculate the optimal decisions for each individual using the simple case of perfect detection of evasion below the notch and no detection of evasion above, as in (11). Since the choice set is non-convex, we calculate possible solutions under each piece-wise linear component of the choice set and then select the feasible solution providing the greatest utility. Third, we calculate total tax revenue given each individual's solution, $\Sigma t w l_r^* = R$ noting that for some individuals (those earning below the notch) reported income will be actual labor supply. Social welfare is $W(t) = \Sigma\Big(V(\alpha, w, t, e) + t w l_r^*\Big)$ where V is indirect utility. In the no-notch no-evasion case, each person pays taxes of $t w l^* = t w \alpha w^e (1-t)^e$. To make the two settings revenue equivalent, while accounting for endogenous taxable income, we use the revenue R calculated above and solve $\Sigma t w \alpha w^e (1-t)^e = R$ for t; call the smallest non-negative t that solves this equation t^*. The final step is to solve for each individual's optimal choice in a no-notch, no-evasion case with tax rate t^*, and calculate social welfare in this no-notch no-evasion setting, $W^0(t) = \Sigma\Big(V(\alpha, w, t^*, e) + t^* w l^*\Big)$.

Table 1 reports simulation outcomes for several different sets of parameter values. Panel A reports for each simulation the ratio of social welfare with the notch over no-notch social welfare, W/W^0. Panel B reports the corresponding revenue-equivalent no-evasion tax t^* values, Panel C reports for each simulation the fraction of taxpayers earning income above the notch (when the notch is in effect) and thus bunching in reported income, and Panel D reports the social welfare values in the notch case alone, W. Each row in each panel uses a different notch-based tax rate, ranging from 0.1 to 0.7.

In the first column of each panel we choose the elasticity $e = .25$ (as in Saez, 2001) and set the notch value at $l^n = .5$. Looking at Panel A, for each row in the first column the ratio is greater than 1, indicating greater social welfare in a setting with an enforcement notch relative to a no-notch, no-evasion, revenue-equivalent setting.

The first column of Panel B shows the corresponding t^* values. All values of t^* are lower than the notch-based tax rates, as one would expect—the t^* values are applied to all earnings, while the notch-based tax rates will only be applied to earnings at or below the notch, so that with inelastic labor supply, a lower tax rate can be applied to a larger tax base while generating the same revenue. Often the t^* values are about half as large as the tax rate with the notch. For example, the last entry in column 1 of Panel B shows that, compared to a notch system with a tax rates of 0.7, a system with no evasion could raise the same revenue with a tax rate of 0.331. The last entry in Panel A shows that in this case the notch system provides 1.42% higher social welfare. Panel C reports the fraction earning income over the notch. The fraction reporting income over the notch is high but, as shown momentarily, efficiency gains from a notch are still possible even when the fraction earning over the notch is lower. Column 1 of Panel D (and all other columns of the panel) show that as t increases, all else equal, welfare in the notch setting unsurprisingly falls.

In the second column of each panel, we increase the elasticity of labor supply to .5, which Gruber and Seaz (2002) report as the elasticity of taxable income (while arguing that the elasticity of real income is lower, so that .5 represents a strong upper bound for e). The relative efficiency of the notch increases further compared to the ratios in column 1. The values in column 2 of Panels B and C are reasonably close to those in column 1. The results in column 2 indicate that, given similar values of t^* as in column 1, the greater elasticity of labor supply leads to greater distortion from taxation and greater efficiency gains from using the notch.

Column 3 uses the parameter values from column 2, except the value of the notch has been increased to 0.8. As the notch increases, the number of taxpayers below the notch

grows (Panel C). Not only do these individuals prefer the no-notch system with its lower t^*, but deadweight loss from these individuals will be smaller in the no-notch system as well. Nevertheless, column 3 of Panel A makes clear that the social gains from using a notch are actually *larger* in this case. Here, for a given tax rate under the notch t^* will increase, and the notch-based cost of increasing distortion for low earners shrinks and the benefit of reducing distortion for high earners grows, so that the social gains from using a notch in A are larger in column 3 than in the other columns despite the fact that the number of individuals actually above the notch is smaller.

Columns 2 and 3 utilize the same preferences so that one can compare notch-based welfare between them in Panel D. There are three takeaways from comparing columns 2 and 3. First, in both settings notch-based welfare is higher than no-notch welfare (Panel A). Second, column 2 both has higher levels of bunching (as shown in Panel C) and higher levels of notch-based welfare (Panel D); this shows that greater bunching can be associated with greater welfare, holding preferences and the underlying distribution of productivity constant, expounding on the intuition from (22) earlier. Lastly, even though column 2 consistently has higher notch-based welfare than the results in column 3, as noted above its welfare gains relative to a no-notch setting are smaller. Increases in bunching can be associated with both greater welfare and with greater welfare relative to no-notch taxation, but the two concepts of "greater welfare" are distinct. Moreover, for these two columns, the differences in the amount of bunching seen in the last few rows of Panel C are similar, but the differences in utility in the last few rows of Panel D vary greatly. One cannot examine the amount of bunching, or compare the amount of bunching in two settings—even for settings where preferences and the underlying distribution of productivity are constant—and draw conclusions about welfare.

In summary, Table 1 illustrates that notches can not only lead to greater efficiency under marginal changes in tax rates, but that a system allowing evasion with an enforcement-induced notch can generate higher welfare than a revenue equivalent system with no evasion

and no notch. We leave further empirical work on the implications of notches to the future, and discuss other issues for future research in the conclusions.

Conclusions

This paper considers bunching in reported income at enforcement notches that create discrete changes in the cost of tax evasion, and shows that the welfare effects of taxation can be increasing with the amount of such bunching. We find that efficiency costs of taxation are closer to zero the greater is bunching, all else equal, and that optimal taxes can be higher the greater is bunching. A general formula for the welfare effects of taxation in this case depends not only upon taxable and labor income, but also on whether reported income is above or below the notch. A simulation shows that a change in the location of a notch that leads to higher bunching can lead to higher social welfare, and social welfare can be greater in a system with bunching and evasion than in a revenue equivalent no-notch, no-evasion world. This paper builds on and provides novel intuition behind an observation in Section 4 of Slemrod (2013) that "general statements about the welfare effects of notches cannot be made."

This paper also has implications for several other topics related to evasion and enforcement in the presence of notches. For example, evasion may involve issues of "horizontal equity" whereby efforts to distort matter more in some professions or for some individuals than others. The result here is obtained by allowing evasion for relatively high earners—those earning above the notch. In that sense, the intuition here, which is driven by the single-agent case, is an example of an equity-efficiency trade off. But if notches are taxpayer-specific, then it might be possible for the intuition here to apply in a way that does not sacrifice equity for efficiency. Considering a case of individual-specific notches (such as from third-party reporting, if such amounts are taken as given), an implication of the analysis here would be that taxes are less distortionary as more taxpayers report income matching third-party-reported

income.

However, some people may always report income truthfully, in which case the results here would suggest that (as in the standard evasion story) the presence of evasion could shift additional tax burden onto truthful reporters. The model here, moreover, assumes that the enforcement system is common knowledge and that evasion is essentially sanctioned for high levels of income; making such a setting explicit could change individuals' views of the tax system itself: a system built upon tacit approval of evasion could lead to lower tax morale or regard for the tax authority in general. Similar ideas are raised in Frey (1997), see also Luttmer and Singhal (2014). Further, over time a tax authority that has promised to allow full evasion beyond a certain level of reported income may be tempted to surprise taxpayers with an unannounced audit. A tax authority using an enforcement notch would have to consider its ability to commit to the notch.

In an important sense the notches in the analysis here are the mirror image of those considered in some prior work, where bunching responds to an increase (rather than a decrease) in costs from passing some threshold: a higher marginal tax rate (Saez, 2010); higher average tax rates (Kleven & Waseem, 2013), or the cap on a tax credit (Hungerman & Wilhelm, 2021). The marginal-cost-decrease from crossing the threshold here leads to a non-convex budget set. One might wonder in this case why there would be any bunching at all; the analysis here highlights that such notches may in fact generate bunching in reported income (similar in spirit to the observations in Klevin, 2016) and consequentially both lead to bunching and lead to important effects on welfare. Interpreting bunching depends upon the nature of the notch.

It has been pointed out to me that the evasion story here resembles the famous result of Mirrlees' model (1971) that in an optimal income tax system the top earner should face a marginal tax of zero. The analysis here does rely on individuals above the notch facing a zero marginal tax rate. It is known that Mirrlees' result is local, but that is not the case here. This paper provides a different, and non-local, rationale for allowing *de facto* zero

marginal tax rates for high earners in a nonlinear tax system.

An implication of this paper is that evasion is not always the same as distortionary behavior, and that evasion may combine in complex and unexpected ways with the tax system. Large recent literatures have explored the potential importance of bunching and the importance of evasion. The results here suggest that future work should carefully consider how these two phenomenon may interact.

References

[1] Agostini, Claudio. 2016. "Small Firms and Presumptive Tax Regimes in Chile: Tax Avoidance and Equity." Working paper.

[2] Alm, James. 2018. "What Motivates Tax Compliance?" *Economic Surveys* June: 1-36.

[3] Almunia, Miguel, and David Lopez-Rodriguez. 2018. "Under the Radar: the Effects of Monitoring Firms on Tax Compliance." *American Economic Journal: Economic Policy* 10: 1-38.

[4] Artavanis, Nikolaos, Adair Morse, and Margarita Tsoutsoura. 2016. "Measuring Income Tax Evasion Using Bank Credit: Evidence From Greece." *The Quarterly Journal of Economics*: 739-798.

[5] Best, Michael, Anne Brockmeyer, Henrik Kleven, Johannes Spinnewijn, and Mazhar Waseem. 2015. "Production versus Revenue Efficiency with Limited Tax Capacity: Theory and Evidence from Pakistan." *Journal of Political Economy* 123: 1311-1355.

[6] Bigio, Saki, and Eduardo Zilberman. 2011. "Optimal Self-employment Income Tax Enforcement." *Journal of Public Economics* 95: 1021-1035.

[7] Blinder Alan and Harvey Rosen. 1985. "Notches." *The American Economic Review* 75: 736-747.

[8] Carrillo, Paul, Dina Pomeranz, and Monica Singhal. 2017. "Dodging the Taxman: Firm Misreporting and Limits to Tax Enforcement." *American Economic Journal: Applied Economics* 9: 144-64.

[9] Chetty, Raj. 2009. "Is the Taxable Income Elasticity Sufficient to Calculate Deadweight Loss? The Implications of Evasion and Avoidance." *American Economic Journal: Economic Policy* 1: 31-52.

[10] Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review* 99: 1145-77.

[11] Feldstein, Martin. 1999. "Tax Avoidance and the Deadweight Loss of the Income Tax." *Review of Economics and Statistics* 81: 674-80.

[12] Frey, Bruno. 1997. "A Constitution for Knaves Crowds out Civic Virtues." *The Economic Journal* 107: 1043-1053.

[13] Gruber, Jonathan, and Saez, Emmanuel. 2002. "The Elasticity of Taxable Income: Evidence and Implications." *Journal of Public Economics* 84, 1-32.

[14] Hungerman, Daniel. 2014. "Public Goods, Hidden Income, and Tax Evasion: Some Nonstandard Results from the Warm-glow Model." *Journal of Development Economics* 109: 188-202.

[15] Hungerman, Daniel, and Mark Wilhelm. 2021. "Impure Impact Giving: Theory and Evidence." *Journal of Political Economy*, 129: 1553-1614.

[16] Keen, Michael, and Joel Slemrod. 2017. "Optimal Tax Administration." *Journal of Public Economics* 152: 133-42.

[17] Kleven, Henrik. 2016. "Bunching." *Annual Review of Economics* 8: 435-64.

[18] Kleven, Henrik, Martin Knudsen, Claus Kreiner, Søren Pedersen and Emmanuel Saez. 2011. "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark." *Econometrica* 79: 651-692.

[19] Kleven, Henrik, and Mazhar Waseem. "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan." *Quarterly Journal of Economics* 128, 669-723.

[20] Litina, Anastasia, and Theodore Palivos. 2016. "Corruption, Tax Evasion, and Social Values." *Journal of Economic Behavior & Organization* 124: 164-177.

[21] Lockwood, Ben. 2020. "Malas Notches." *International Tax and Public Finance* 27: 779-804.

[22] Luttmer, Erzo F. P., and Monica Singhal. 2014. "Tax Morale." *Journal of Economic Perspectives* 28: 149-68.

[23] Meiselman, Ben. 2018. "Ghostbusting in Detroit: Evidence on Nonfilers from a Controlled Field Experiment." *Journal of Public Economics* 158: 180-193.

[24] Mirrlees, J.A. 1971. "An Exploration in the Theory of Optimum Income Taxation." *The Review of Economic Studies* 38: 175-208.

[25] National Taxpayer Advocate. 2017. *Annual Report to Congress, Volume 1.*

[26] Reinganum, Jennifer, and Louis Wilde. 1985. "Income Tax Compliance in a Principal-Agent Framework." *Journal of Public Economics* 26: 1-18.

[27] Saez, Emmanuel. 2010. "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy* 2: 180-212.

[28] Saez, Emmanuel. 2001. "Using Elasticities to Derive Optimal Income Tax Rates." *Review of Economic Studies* 68: 205-29.

[29] Sallee, James, and Joel Slemrod. 2012. "Car Notches: Strategic Automaker Responses to Fuel Economy Policy," *Journal of Public Economics* 96: 981-999.

[30] Slemrod, Joel. 2013. "Buenas Notches: Lines and Notches in Tax System Design." *eJournal of Tax Research* 11: 259-283.

Appendix

A1. General Preferences

Here we derive the main result from (2.2) for general preferences. The agent maximizes

$$\max_{l,\ l_r} \quad U(y - \tau + wl - twl_r - z(l,\ l_r,\ t),\ l) \tag{26}$$

where τ is a lump sum tax that will be used to derive comparative statics. This yields the first order conditions:

$$l: \quad U_c(\cdot)(w - z_l) + U_l(\cdot) \ = \ 0 \tag{27}$$

$$l_r: \quad U_c(\cdot)(-tw - z_{l_r}) = 0 \tag{28}$$

Consider a change in the lump sum tax rate τ in a world with no evasion. In this case the first order condition in (27) still holds, with the function z set to zero. Taking the derivative of (28), we have:

$$U_{cc}(\cdot)(-1+wl_\tau)w + U_{cl}(\cdot)l_\tau w + U_{lc}(\cdot)(-1+wl_\tau) + U_{ll}(\cdot)l_\tau = 0 \tag{29}$$

Gathering terms,

$$l_\tau = \frac{wU_{cc}(\cdot) + U_{lc}(\cdot)}{w^2U_{cc}(\cdot) + 2wU_{lc}(\cdot) + U_{ll}(\cdot)} \tag{30}$$

Now consider a case with $l_r^* = l^n$, where again by assumption the individual is at a notch so that (28) does not hold and $z = 0$. Differentiating (27) with respect to t, we have:

$$U_{cc}(\cdot)(wl_t - wl^n)w + U_{cl}(\cdot)l_t w + U_{lc}(wl_t - wl^n) + U_{ll}(\cdot)l_t = 0. \tag{31}$$

And solving for l_t yields:

$$l_t = \frac{w^2l^nU_{cc}(\cdot) + wl^nU_{lc}(\cdot)}{w^2U_{cc}(\cdot) + 2wU_{lc}(\cdot) + U_{ll}(\cdot)} = (wl^n)l_\tau \tag{32}$$

and so for a change in the tax rate $d\tau = wl^n dt$ the two effects will be the same. Thus, we once again have that a change in the proportional tax rate t produces no change in l_r^*, and elicits the same effect on l^* as does an appropriately sized lump sum tax τ.

A2. Derivation for the Saez Model

Here we will show that it is possible to have efficient taxation in a model akin to the one used in Sectio 2.C in Saez (2010). That model differs from the model used in this study in four ways. First, he assumes linear preferences. Second, he introduces fixed costs of both reporting earned income, denoted q_A, and of reporting earned income different from actual earned income, denoted q_M. Third, he assumes earnings do not respond to taxes. Fourth, he assumes that the tax function is single peaked.

In Saez's notation, total earnings (which are perfectly inelastic) are $w + y$, where w is formal earnings that cannot be evaded, and y is informal earnings that can be evaded. The individual chooses $\hat{y}$ to report; w is assumed fixed. The individual faces a tax function $-T(w+\hat{y})$ which is single peaked. We will suppose that this tax function is differentiable in a parameter t, so that the function is $T(w+\hat{y},\ t)$. Otherwise, we would not have a parameter to optimize over for social welfare. The individual maximizes:

$$\max_{\hat{y}} \quad y + w - T(w+\hat{y}) - q_A 1(\hat{y} > 0) - q_M 1(\hat{y} \neq y) \tag{33}$$

(cf. equation 7 in Saez). The only optimal choices would be to report no informal income, to truthfully report, or to report so that one is at the peak. (Saez proves this.) Suppose we are at the peak so that the first order condition holds:

$$-\frac{\partial T}{\partial \hat{y}} = 0. \tag{34}$$

The welfare function is:

$$W(t) = \{w + y - T(w + \hat{y},\ t) - q_A 1(\hat{y} > 0) - q_M 1(\hat{y} \neq y)\} + T(w + \hat{y},\ t) \tag{35}$$

As in the text, assume that a change in the parameter t does not induce any "switching," for example between choosing the be at the peak versus choosing to report all (or no) informal income. Then the derivative of (35) is:

$$\frac{dW(t)}{dt} = -\frac{\partial T}{\partial t} + \frac{\partial T}{\partial t} + \frac{\partial T}{\partial \hat{y}}\frac{d\hat{y}}{dt} = 0. \tag{36}$$

where the first two terms cancel and the last term is zero by (34). Thus, the model again shows that under evasion, the marginal welfare cost of a change in the tax system must be zero. Note here that the result is obtained not just for a linear tax rate, but for a more general change in parameters in the face of a nonlinear (but single peaked) tax system.

The result is worth a second thought, however, as one might observe that in the above solution there is a discontinuity in the cost of evasion, but this discontinuity occurs away from the optimal choice of evasion. Here, and unlike in the main analysis in the paper, evasion is determined by a first order condition rather than by an enforcement notch. The difference comes from Saez's strong assumption that actual labor supply is perfectly inelastic: in Saez's model it is *actual* labor supply that is at a "corner" solution (in that no interior first order condition characterizes it), while *reported* labor supply is determined by an interior solution that, due to single peaked taxes, sets the marginal tax rate to zero. Saez's assumption of a nonlinear tax system with linear preferences is also the reverse of what is done in the paper, although any combination of the two that admits a convex solution characterized by a first order condition independent of a tax rate could suffice. Aside from highlighting the overlooked importance of enforcement notches, and enforcement more generally in discussions of taxation and welfare, a benefit of the model used in the main text of this paper is that it allows taxation to be distortionary in labor supply, whereas by assumption that channel is shut down in Saez's model. In the Saez model, if reported income equals true income, then the tax system has no effect (since true income is by assumption fixed) and if reported

income is zero, then the tax system creates no deadweight loss but also creates no revenue. Indeed, in Saez's model, which was not intended for serious consideration of efficiency issues, if $q_A = 0$ then there will *never* be deadweight loss from taxation. But both models highlight how a corner solution in one choice variable and an interior solution in the other can facilitate an efficient tax system that nonetheless manages to raise revenue.

A3. Proof of Proposition 1

The first order conditions are

$$l: \quad w - z_l - g_l - \psi(l) \;=\; 0 \tag{37}$$

$$l_r: \quad -tw - z_{l_r} - g_{l_r} \;=\; 0. \tag{38}$$

For part A, $g_l = -g_{l_r}$ and $z_l = -z_{l_r}$. This last equality becomes $z_l = tw - g_l$ using (38). Plugging this into (37) produces the solution for l^* in (17). Replacing $-g_{l_r}$ with g_l and taking its inverse leads to the solution for l_r^*. For part B, the solution for l_r^* is the same as in A noting $z_{l_r} = 0$. Plugging (38) into (37) produces the solution for l^*. For part C, l_r^* is given and l^* follows directly from (37). Multiply both sides of each solution by w for the expressions in the proposition.

A4. Proof of Proposition 2

Differentiating (21) with respect to t yields:

$$\frac{d\mathbb{W}(t)}{dt} = \sum_{l_r<l^n} \frac{dW(t)}{dt} h^u(\alpha) + \sum_{l_r=l^n} \frac{dW(t)}{dt} h^n(\alpha) + \sum_{l_r>l^n} \frac{dW(t)}{dt} h^a(\alpha). \tag{39}$$

The above expression consists of three summations, each taken over different values of α. For each of the three summations, we can write the general expression:

$$\sum \frac{dW(t)}{dt} h^i(\alpha) = \sum \Big(\{-wl_r^* - \frac{\partial z}{\partial t}\} + \frac{\partial z}{\partial t} + \frac{\partial z}{\partial l}\frac{dl^*}{dt} + \frac{\partial z}{\partial l_r}\frac{dl_r^*}{dt} + wl_r^* + tw\frac{dl_r^*}{dt} \Big) h^i(\alpha) \tag{40}$$

where the superscript $i \in \{u, n, a\}$ determines whether the summation in question is for individuals under, at, or above the notch. The expression in (40) simplifies to

$$\sum \frac{dW(t)}{dt} h^i(\alpha) = \sum \Big[\frac{\partial z}{\partial l}\frac{dl^*}{dt} + \frac{dl_r^*}{dt}\Big(\frac{\partial z}{\partial l_r} + tw \Big) \Big] h^i(\alpha). \tag{41}$$

The evaluation of equation (41) depends upon whether the sum in question is for individuals reporting income under, at, or above the notch. For those under the notch, where $l_r^* < l^n$, the optimal solutions are given by (17) (where one could write $\psi^{-1}(w(1-t))$ as $\psi^{-1}(w(1-t), \alpha)$). By the first order condition for (14), for this group $z_{l_r} = -tw - g_{l_r} = -tw + g_l$, and $z_l = -z_{l_r}$ and finally $g_l = -g_{l_r}$. Then (41) becomes:

$$\sum_{l_r<l^n} \frac{dW(t)}{dt} h^u(\alpha) = \sum_{1_r<l^n} \Big[\frac{dl^*}{dt}(tw - g_l) + \frac{dl_r^*}{dt} g_l\Big] h^u(\alpha). \tag{42}$$

For those at the notch, $z = 0$ and by (19) we have $\frac{dl^*}{dt} = \frac{dl_r^*}{dt} = 0$ so that the summation equals zero. For those reporting income above the notch, $z = 0$ and the first order condition becomes $tw = g_{l_r}$. Plugging these results in for all individuals yields:

$$\frac{d\mathbb{W}(t)}{dt} = \sum_{l_r<l^n} \Big[\frac{dl^*}{dt}(tw - g_l) + \frac{dl_r^*}{dt} g_l\Big] h^u(\alpha) + 0 + \sum_{l_r>l^n} \frac{dl_r^*}{dt}\ tw\ h^a(\alpha). \tag{43}$$

Lastly, while each of the above summations resembles an expected value, as noted above the PMFs h^u and h^a do not sum to unity. Re-scaling them to sum to unity (ie, multiplying by $\mathcal{G}(l^n) - \beta$ over $\mathcal{G}(l^n) - \beta$, for the first summation) allows us to express (43) as:

$$\frac{d\mathbb{W}(t)}{dt} = t\Big(\mathcal{G}(wl^n) - \beta\Big)\mathrm{E}\Big[\ \mu\frac{d\mathrm{LI}}{dt} + (1-\mu)\frac{d\mathrm{TI}}{dt}\ \Big|\ l_r < l^n\Big] + t\Big(1 - \mathcal{G}(wl^n)\Big)\mathrm{E}\Big[\ \frac{d\mathrm{TI}}{dt}\ \Big|\ l_r > l^n\Big] \tag{44}$$

which matches the proposition.

A5.Moving onto the Notch: An Example

Consider a taxpayer strictly above the notch, so that $z(l^*,\ l_r^*,\ t) = 0$. The initial tax rate is t and let $w = 1$. Let the cost of evasion be quadratic, $g(l - l_r) = \frac{1}{2}(l - l_r)^2$, and suppose labor effort is isoelastic with unit elasticity of supply e so that $\Psi = \frac{1}{1+e} l^{(1+1/e)} = \frac{1}{2} l^2$. The individual solves:

$$\max_{l,\ l_r}\ \ y + l - tl_r - \frac{1}{2}(l - l_r)^2 - \frac{1}{2}l^2. \tag{45}$$

The first order conditions are $1 - l + l_r - l = 0$ and $-t + l - l_r = 0$. Combining yields $l^* = 1 - t$ and $l_r^* = 1 - 2t$. Assume an interior optimal choice for l_r^*, so that $t < 1/2$. Of course, if the tax rate increases, labor supply and reported income both fall, but reported income falls by

more. Using the first order conditions (noting $l - l_r = t$), the value function is

$$V(l^*, l_r^*, t) = y + 1 - t - t(1-2t) - \frac{1}{2}t^2 - \frac{1}{2}(1-t)^2 \quad (46)$$

$$= y + 1/2 - t + t^2. \quad (47)$$

Now introduce a notch l^n. For an individual who bunches, $l_r^* = l^n$ and the first order condition for l in (45) becomes $l^* = \frac{1}{2}(1 + l^n)$. A bunching individual desires labor income above the notch and reported income below the notch; the first will hold if $l^n < 1$ and the second if $1 - 2t < l^n$. The value function now is:

$$V(l^*, l^n, t) = y + \frac{1}{2} + \frac{l^n}{2} - tl^n - \frac{1}{2}\Big(\frac{1}{2} - \frac{l^n}{2}\Big)^2 - \frac{1}{2}\Big(\frac{1}{2} + \frac{l^n}{2}\Big)^2 \quad (48)$$

$$= y + \frac{1}{4} + \frac{l^n}{2} - tl^n - \frac{(l^n)^2}{4}. \quad (49)$$

It can be shown that (47) > (49); this is omitted for brevity but logically it must be true, else the individual optimized wrong.

Consider then an individual initially at an interior solution who faces an increase in taxes from t to $\bar{t} > t$. We will compare a purely-interior case to a case where a notch is present and the change results in "switching," i.e. bunching at the notch. From the above it follows that the individual facing the notch who must bunch will be worse off than if they were not constrained at the notch. But how will social welfare compare between the two cases?

Begin with the purely-interior case. Social welfare is: $V(l^*, l_r^*, t) + tl_r^* = y + 1/2 - t^2$. With the higher tax rate, this is simply $y + 1/2 - \bar{t}^{\,2}$. Next, if the notch is present and the tax increase induces a person to bunch, initial welfare is the same as before, and after the tax rate change welfare is $V(l^*, l^n, \bar{t}) + tl^n = y + 1/4 + l^n/2 - (l^n)^2/4$.

A tax rate that induces bunching onto notch l^n will thus have higher social welfare than in the no-notch case if (and only if):

$$y + 1/4 + l^n/2 - (l^n)^2/4 \;>\; y + 1/2 - \bar{t}^{\,2}. \quad (50)$$

The left-hand side is increasing in l^n for $l^n < 1$. Since $l^n > 1 - 2\bar{t}$, the left hand side must then be greater than what is obtained by substituting $1 - 2\bar{t}$ for l^n:

$$y + 1/4 + l^n/2 - (l^n)^2/4 \;>\; y + 1/4 + (1-2\bar{t})/2 - (1-2\bar{t})^2/4 \;=\; y + 1/2 - \bar{t}^{\,2}. \quad (51)$$

The equality follows from routine algebra but again logically must be true—if the notch

were set to the initial level of reported income, social welfare will be the same as before. Thus for "switchers", individual utility falls by more if a notch necessitates bunching after a tax increase, but the presence of the notch yields greater social welfare.

A6. Optimal Taxation and Proof of Proposition 3

Consider first the proposition. Social welfare is given by

$$\tilde{\mathbb{W}}(t,\phi) \;=\; \sum_{i\in\{u,n,a\}} \sum_{\alpha} \tilde{W}(t,\phi) h^i(\alpha) \tag{52}$$

where

$$\tilde{W}(t,\phi) \;=\; \{y+wl^*-twl_r^*-z(l^*,\; l_r^*,\; t)-\Psi(l^*,\alpha)-g(l^*-l_r^*)\}+z(l^*,\; l_r^*,\; t)+twl_r^*(1+\phi) \tag{53}$$

This matches the social welfare function in section 2.3 used to derive equation (22), with an additional term ϕtwl_r^* for each individual. Using equation (22), then, the first derivative of (52) can be written:

$$\begin{aligned}\frac{d\tilde{\mathbb{W}}(t)}{dt} \;&= t\Big(\mathcal{G}(wl^n)-\beta\Big)\mathrm{E}\Big[\mu\tfrac{d\mathrm{LI}}{dt}+(1-\mu)\tfrac{d\mathrm{TI}}{dt}\Big|l_r<l^n\Big]+t\Big(1-\mathcal{G}(wl^n)\Big)\mathrm{E}\Big[\tfrac{d\mathrm{TI}}{dt}\Big|l_r>l^n\Big] \\ &\qquad +\phi\sum_{i\in\{u,n,a\}}\sum_{\alpha}(tw\tfrac{dl_r^*}{dt}+wl_r^*)h^i(\alpha)\end{aligned} \tag{54}$$

where the first row is from (22). The second row can be written $\phi\big(\bar{\mathrm{TI}}+t\mathrm{E}[\frac{dTI}{dt}]\big)$, where $\bar{\mathrm{TI}}=\sum_{i\in\{u,n,a\}}\sum_{\alpha}(wl_r^*)h^i(\alpha)$ represents the sum and mean of taxable income for all individuals since the total population is of size unity. Using this, setting (54) to zero, and denoting $\mathcal{G}(wl^n)$ as $\mathcal{G}$, we have:

$$t\Big(\mathcal{G}-\beta\Big)\mathrm{E}\Big[\mu\frac{d\mathrm{LI}}{dt}+(1-\mu)\frac{d\mathrm{TI}}{dt}\Big|l_r<l^n\Big]+t\Big(1-\mathcal{G}\Big)\mathrm{E}\Big[\frac{d\mathrm{TI}}{dt}\Big|l_r>l^n\Big]+\phi\Big(\bar{\mathrm{TI}}+t\mathrm{E}\Big[\frac{dTI}{dt}\Big]\Big)=0 \tag{55}$$

Define mean elasticities thusly:

$$\begin{aligned}
\varepsilon_\ell^u &= -\mathrm{E}\Big[\frac{d\mathrm{LI}}{dt}\Big| l_r < l^n\Big]\frac{1-t}{\bar{\mathrm{LI}}^u},\ \text{where } \bar{\mathrm{LI}}^u = \mathrm{E}[\mathrm{LI}|l_r < l^n] \\
\varepsilon_t^u &= -\mathrm{E}\Big[\frac{d\mathrm{TI}}{dt}\Big| l_r < l^n\Big]\frac{1-t}{\bar{\mathrm{TI}}^u},\ \text{where } \bar{\mathrm{TI}}^u = \mathrm{E}[\mathrm{TI}|l_r < l^n] \\
\varepsilon_t^a &= -\mathrm{E}\Big[\frac{d\mathrm{TI}}{dt}\Big| l_r > l^n\Big]\frac{1-t}{\bar{\mathrm{TI}}^a},\ \text{where } \bar{\mathrm{TI}}^a = \mathrm{E}[\mathrm{TI}|l_r > l^n] \\
\varepsilon_t &= -\mathrm{E}\Big[\frac{d\mathrm{TI}}{dt}\Big]\frac{1-t}{\bar{\mathrm{TI}}},\ \text{where } \bar{\mathrm{TI}} = \mathrm{E}[\mathrm{TI}]
\end{aligned} \tag{56}$$

And define income shares as:

$$\begin{aligned}
\theta_\ell^u &= \frac{(\mathcal{G}-\beta)\bar{\mathrm{LI}}^u}{\bar{\mathrm{TI}}} \\
\theta_t^u &= \frac{(\mathcal{G}-\beta)\bar{\mathrm{TI}}^u}{\bar{\mathrm{TI}}} \\
\theta_t^a &= \frac{(1-\mathcal{G})\bar{\mathrm{TI}}^a}{\bar{\mathrm{TI}}}
\end{aligned} \tag{57}$$

Then the first order condition in (55) can be written as

$$\frac{-t}{1-t}\ \bar{\mathrm{TI}}\ \Big(\mu\ \varepsilon_\ell^u\ \theta_\ell^u + (1-\mu)\ \varepsilon_t^u\ \theta_t^u + \varepsilon_t^a\ \theta_t^a\Big) + \phi\ \bar{\mathrm{TI}} - \phi\frac{t\ \bar{\mathrm{TI}}}{1-t}\ \varepsilon_t = 0 \tag{58}$$

yielding:

$$\frac{t^*}{1-t^*} = \frac{\phi}{\Big(\mu\ \varepsilon_\ell^u\ \theta_\ell^u + (1-\mu)\ \varepsilon_t^u\ \theta_t^u + \varepsilon_t^a\ \theta_t^a + \phi\ \varepsilon_t\Big)} \tag{59}$$

which matches equation (24) in the proposition.

Equation (25), that $\frac{t^*}{1-t^*} = \frac{\phi}{e(1-\theta_t^n)(1+\phi)}$, follows. First, by construction

$$\theta_t^u+\theta_t^n+\theta_t^a = \frac{(\mathcal{G}-\beta)\sum\limits_{l_r<l^n}\frac{\mathrm{TI}h(\alpha)}{\mathcal{G}-\beta} + \beta\sum\limits_{l_r=l^n}\frac{\mathrm{TI}h(\alpha)}{\beta} + (1-\mathcal{G})\sum\limits_{l_r>l^n}\frac{\mathrm{TI}h(\alpha)}{(1-\mathcal{G})}}{\sum \mathrm{TI}h(\alpha)} = \frac{\sum \mathrm{TI}h(\alpha)}{\sum \mathrm{TI}h(\alpha)} = 1 \tag{60}$$

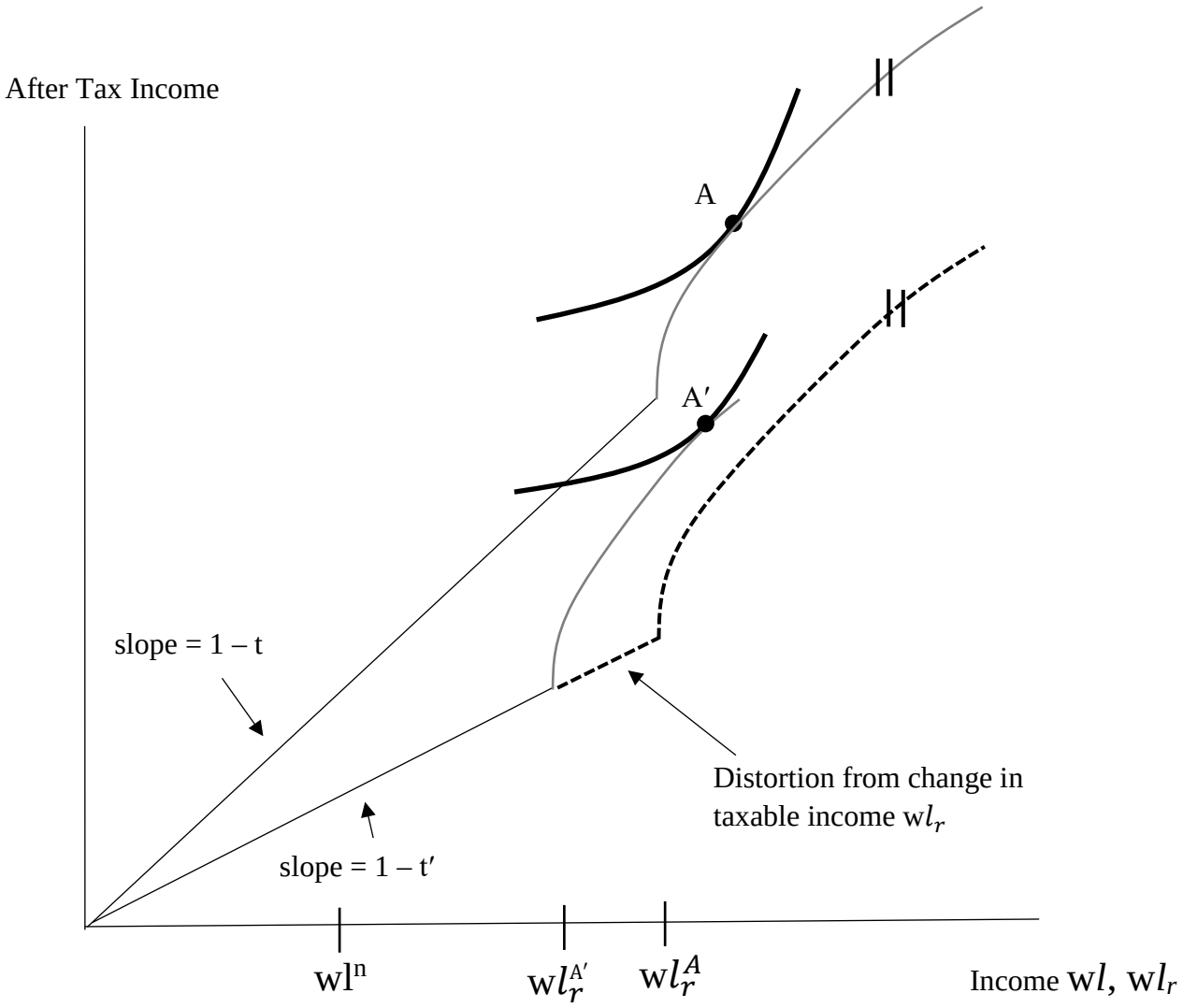
so that setting $\mu = 0$ in equation (24) leads to $\frac{t^*}{1-t^*} = \frac{\phi}{e\theta_t^u + e\theta_t^a + \phi\varepsilon_t} = \frac{\phi}{e(1-\theta_t^n)+\phi\varepsilon_t}$.

Next, we show that $\varepsilon_t = e(1-\theta_n^n)$, and the result follows. This can be seen given:

$$
\begin{aligned}
\varepsilon_t &= -\mathrm{E}\Big[\frac{d\mathrm{TI}}{dt}\Big]\frac{1-t}{\bar{\mathrm{TI}}} \\
&= \frac{1-t}{\bar{\mathrm{TI}}}\sum\frac{d\mathrm{TI}}{dt}h(\alpha) \\
&= \frac{1-t}{\bar{\mathrm{TI}}}\Big[\sum_{l_r<l^n}\frac{d\mathrm{TI}}{dt}h^u(\alpha)+\sum_{l_r=l^n}\frac{d\mathrm{TI}}{dt}h^n(\alpha)+\sum_{l_r>l^n}\frac{d\mathrm{TI}}{dt}h^a(\alpha)\Big] \\
&= \Big[\sum_{l_r<l^n}\frac{d\mathrm{TI}}{dt}\frac{1-t}{\bar{\mathrm{TI}}^u}h^u(\alpha)\frac{\bar{\mathrm{TI}}^u}{\bar{\mathrm{TI}}}+0+\sum_{l_r>l^n}\frac{d\mathrm{TI}}{dt}\frac{1-t}{\bar{\mathrm{TI}}^a}h^a(\alpha)\frac{\bar{\mathrm{TI}}^a}{\bar{\mathrm{TI}}}\Big] \\
&= \frac{\bar{\mathrm{TI}}^u}{\bar{\mathrm{TI}}}(\mathcal{G}-\beta)e+0+\frac{\bar{\mathrm{TI}}^a}{\bar{\mathrm{TI}}}(1-\mathcal{G})e \\
&= e\theta_t^u+e\theta_t^a=e(1-\theta^n). \qquad (61)
\end{aligned}
$$

Plugging this in for ε_t yields the result.

Figure 1: Tax Efficiency above an Enforcement Notch



The figure shows the choices made by an individual who earns and reports income above notch wl^n. The initial tax rate is t and initial reported taxable income is wl_r^A; income above this amount is evaded. The nonlinear cost of evasion results in a kinked budget constraint above reported income. When the tax rate increases to t′ the budget line rotates downward; the dashed segment shows the budget line if the individual were to continue reporting wl_r^A . Above reported income, this dashed line is parallel to the original budget line. The budget line kinked at w$l_r^{A'}$ shows the optimal response by the taxpayer in reporting income after taxes increase. The movement of the kink is the source of distortion.

Table 1: An Evasion- and Notch-Based Tax

Panel A: The Ratio of Notch Welfare over No-Notch Welfare

		e = .25, l_n = .5	e = .5, l_n = .5	e = .5, l_n = .8
	0.1	1.0003	1.0006	1.0010
	0.2	1.0011	1.0026	1.0041
	0.3	1.0025	1.0062	1.0094
Tax Rate	0.4	1.0046	1.0112	1.0171
with Notch	0.5	1.0073	1.0178	1.0272
	0.6	1.0107	1.0254	1.0434
	0.7	1.0142	1.0341	1.0657

Panel B: Revenue-Equivalent No-Evasion Tax Rate t*

		e = .25, l_n = .5	e = .5, l_n = .5	e = .5, l_n = .8
	0.1	0.044	0.045	0.067
	0.2	0.090	0.092	0.136
	0.3	0.136	0.140	0.209
Tax Rate	0.4	0.183	0.191	0.285
with Notch	0.5	0.232	0.244	0.364
	0.6	0.281	0.299	0.450
	0.7	0.331	0.356	0.542

Panel C: Proportion of Taxpayers Earning Income above the Notch

		e = .25, l_n = .5	e = .5, l_n = .5	e = .5, l_n = .8
	0.1	0.911	0.908	0.647
	0.2	0.907	0.898	0.633
	0.3	0.901	0.889	0.604
Tax Rate	0.4	0.898	0.872	0.575
with Notch	0.5	0.891	0.852	0.541
	0.6	0.882	0.828	0.518
	0.7	0.869	0.807	0.487

Panel D: Social Welfare with the Notch

		e = .25, l_n = .5	e = .5, l_n = .5	e = .5, l_n = .8
	0.1	894	745	744
	0.2	894	745	743
	0.3	893	744	739
Tax Rate	0.4	893	742	732
with Notch	0.5	892	740	721
	0.6	890	735	706
	0.7	888	728	683

Results are from simulations of 1,000 inviduals with iso-elastic quasilinear preferences with labor elasticity e and with a lognormal disutility of work parameter with $\mu = 0$ and $\sigma = .5$. Panel A shows for each set of parameters the ratio of social welfare with the notch over revenue-equivalent no-notch social welfare. Panel B shows the no-notch tax rate. Panel C shows those earning above the enforcement notch (and reporting income at the notch) when the notch is in place, and panel D shows aggregate social welfare under the notch alone.