Abstract

We develop an optimal tax framework that combines two recent extensions of tax analysis: a tax-systems emphasis on non-rate policy instruments, and taxation with behavioral biases. Although the implications of taxpayers’ biases for optimal tax rates has received considerable attention, a complete analysis of this aspect of optimal tax theory must account for the fact that such biases may be endogenous to the non-rate aspects of a tax system. We extend optimal tax systems analysis to incorporate behavioral biases, illustrating the implications of this perspective using two simple models. We characterize the optimal tax system in this setting under a fairly general formulation of taxpayer bias and (potentially costly) non-rate policies. We also discuss the implications of our framework for applied welfare analysis with behavioral agents, documenting new possibilities and challenges for such work that a tax-systems view reveals.
1 Introduction and Motivation

Recent empirical studies have compellingly established that in many situations some people do not act in accordance with the standard model of rational behavior, under which individuals correctly perceive their budget set and make choices to maximize a stable utility function. Instead, they often exhibit a wide variety of so-called behavioral anomalies. Both positive and normative analyses of taxation have begun to incorporate these developments. A prominent example is Chetty et al. (2009), who argue that many consumers under-react to sales taxes that are not fully salient and develop an applied welfare analysis framework that accommodates both salience effects and other optimization failures. Farhi and Gabaix (2018) build a theory of optimal taxation that encompasses a wider range of behavioral biases including salience, misperceptions, misoptimization, and mental accounting. They show that the impact of behavioral biases on the optimal tax regime can be captured by a set of sufficient statistics.

Another recent development in tax analysis is a widening of focus beyond tax rates and tax bases to encompass the vector of other tax policy instruments available to a tax authority, such as the scope and coverage of audits, evasion penalties, the range of information reporting requirements, the details of the remittance regime, the extent of taxpayer education, and public disclosure. Slemrod and Gillitzer (2013) call this a tax-systems approach, and Keen and Slemrod (2017) characterize the optimal setting of the vector of tax policy enforcement instruments simultaneously with the optimal tax rate structure in the absence of behavioral anomalies.

What has not been studied heretofore are the important connections between the incorporation of behavioral anomalies into tax theory and a tax-systems perspective. To get a sense of the close connection between these subjects, consider the following statement from Farhi and Gabaix (2018, p. 14): “With behavioral biases, estimating these sufficient statistics requires extra care, as they might be highly context dependent, taking different values depending on factors that would be irrelevant in the traditional model, such as: the salience of taxes; the way taxes are collected; the complexity of the tax system; information about the tax system...; the presence of nudges, etc.” Strikingly, each of these aspects of context depend on the choice of certain features of a tax system. A benevolent social planner therefore must consider how non-rate policy choices—such as the remittance system, the enforcement regime, or the accessibility of information about tax rules—interact with taxpayer biases when determining the social welfare maximizing tax system. Indeed, “the way taxes are collected” is a succinct description of what a tax-systems approach adds to traditional tax analysis.

In this paper we develop an optimal tax framework that combines both of these extensions of the traditional
model and draw out several connections between the two. First, as already noted, the extent to which behavioral biases influence responses to taxation may be substantially influenced by the non-rate, non-base aspects of a tax system. For example, the salience of the tax system depends on how much information is dispensed about the system and how successfully it is conveyed to taxpayer. In the extreme, the IRS could make available to any taxpayer who so requested a revenue officer who would explain any aspect of the tax system.¹ This would be the tax equivalent of stationing a police officer on every street corner in the country; while that would certainly reduce street crime, it would almost certainly fail a welfare maximization criterion due to the requisite costs. Just as there is an “optimal” elasticity of taxable income with respect to the tax rate (Slemrod and Kopczuk, 2002) and an optimal tax evasion gap (Keen and Slemrod, 2017), so too is there an optimal amount of behavioral anomaly. Goldin (2015) previously addressed this matter in the context of tax salience but, of course, these issues extend beyond that topic. For example, just as taxpayers may sometimes underestimate their true marginal tax rate, so too might they systematically misperceive the probability of tax evasion being detected and penalized. What is the optimal form of such misperception? If taxpayers rely on decision-making heuristics, should a benevolent social planner seek to influence these heuristics? The framework we develop here can be applied quite generally to the design of optimal tax systems when non-rate policy instruments may alter both taxpayer incentives and taxpayer biases.

Our approach recasts the optimal tax administration framework of Keen and Slemrod (2017) in a model with behavioral biases. In Section 2, we illustrate this using a representative agent model of optimal income taxation without evasion or avoidance. We introduce a new type of policy lever to the social planner’s toolkit that we call bias alteration policies. These policies allow the planner to influence the taxpayer’s behavioral biases at some cost. We use our model to discuss the concept of optimal bias alteration and characterize the optimal tax system in this setting. We show that the optimal tax system is one which adjusts the taxpayer’s biases so to induce behavior that is more consistent with the taxpayer internalizing the fiscal externality associated with her behavioral response to taxation, but only does so to the extent that this is feasible and cost-effective given the bias alteration policy menu available to the planner. We also find that the sufficient statistics needed to evaluate the first-order welfare effects of bias alteration policy are quite similar to those needed to evaluate non-rate policies in the standard model. The key additional informational requirements are measures of taxpayer biases: the same type of statistics that need to be estimated to evaluate optimal tax rate formulae with biased agents. Finally, we explore the possibility of identifying welfare-improving tax

¹Even this kind of radical intervention need not necessarily render taxes fully salient. Limitations on human cognition likely prevent perfect perception of the tax system even with costless access to information. The apparent prevalence of left-digit bias provides an instructive analogy. Left-digit bias describes the tendency of individuals to pay more attention to the leftmost digits of prices (and other numbers) when making decisions. Shlain (2018) provides empirical evidence that consumers exhibit left-digit bias and that retail firms respond to this by disproportionately buying goods with prices that end in 99 cents. This suggests that even when prices are right in front of an individual, they may not properly perceive them.
system changes in the absence of information about taxpayer biases, something which cannot be accomplished by considering the optimal choice of each policy lever in isolation.

In Section 3, we extend the framework to incorporate evasion responses. We consider a model in which the tax authority simultaneously chooses the tax rate, an enforcement policy, and a bias alteration policy. In addition to discussing some analogues to the results presented in section 2, we discuss the implications of this model for applied welfare analysis of such policies, with a particular emphasis on the audit-threat interventions that have received substantial attention in the recent empirical tax compliance literature. In addition, this section highlights some challenges behavioral biases create for the ETI-based approach to normative tax analysis. Section 4 concludes the paper, providing further discussion with a view to the broader implications of our work.

Related Literature

The clearest precedent for our work is Goldin’s (2015) analysis of optimal tax salience. In his model, the social planner simultaneously chooses the tax rate and the salience of the tax. He shows that—when it is costless to do so—the planner will choose a level of salience that enables the attainment of the first-best welfare outcome. In section 2, we extend this result to a much more general formulation of behavioral biases and consider how it changes when altering taxpayer biases is costly.

Elsewhere in the existing literature on optimal taxation with behavioral biases, there has been some limited discussion of policies that alter taxpayer biases. For example, in their work on optimal taxation with heterogeneous tax salience, Taubinsky and Rees-Jones (2017) include a discussion of the welfare effects of changing the distribution of tax salience. Similarly, in their examination of the optimal tax problem when some fraction of taxpayers exhibit ironing\(^2\) behavior, Rees-Jones and Taubinsky (2019) consider the welfare effects of increasing the fraction of taxpayers who use the ironing heuristic. Finally, in their investigation of optimal taxation with biased agents, Farhi and Gabaix (2018) present a model of optimal nudging which includes some analysis of the optimal combination of nudges and commodity taxes in problems with both biased behavioral and externalities.

This paper also relates to the broader literature on optimal policy with behavioral agents. A number of authors consider how to optimally set paternalistic savings policies, such as default or mandatory retirement contribution plans, as in the recent work of Bernheim et al. (2015), Goldin and Reck (2017), and Fadlon and Laibson (2017). In particular, Fadlon and Laibson (2017) emphasizes, as we do, the idea that observed

\(^2\)A taxpayer is said to be using the ironing heuristic when they behave as if their marginal tax rate were equal to their average tax rate.
departures from—or conformity with—rational choice behavior can be the result of deliberate government policy. However, it is important to note that in the context of savings policies, the goals of the social planner and those of private agents are aligned, implying that optimal policies will induce biased agents to behave in a manner more similar to standard rational agents. By contrast, in a tax policy setting, biases can be beneficial to social welfare, and thus the planner may be incentivized to push taxpayer behavior further away from rationality.\(^3\)

## 2 Income Taxation with Optimal Biases

We begin by revisiting the problem of choosing an optimal linear income tax when taxpayer behavior departs from the assumptions of the standard model. We consider a simple two-good setting with a representative agent whose biases are endogenous to government policy.\(^4\)

### 2.1 Model Setup

Consider the following optimal tax problem. The representative agent’s private *experienced* utility is

\[
u(\ell) \equiv (1 - t)w\ell - \psi(\ell)
\]

where \(t \in [0, 1]\) is a linear income tax rate, \(\ell \geq 0\) is the agent’s labor supply, and \(w > 0\) is her wage rate.\(^5,6,7\)

Suppose that \(\psi\) is twice continuously differentiable, \(\psi(0) = 0, \psi_\ell > 0, \psi_{\ell\ell} > 0,\) and \(\lim_{\ell \to 0}\psi_\ell = \infty.\)\(^8\)

Unlike in the usual optimal tax problem, our agent is not an experienced utility maximizer. Instead we...
require that her choice of $\ell$ satisfies the condition

$$\tau^b = (1 - t) w - \psi(\ell) \quad \text{(2)}$$

where $\tau^b \equiv \tau^b(\ell; w, t, \alpha)$ is the agent’s behavioral wedge.\(^9\) We assume $\tau^b$ is twice continuously differentiable in each of its arguments. For the remainder of this section, let $\ell \equiv \ell (w, t, \alpha)$ be the choice of labor supply that satisfies the condition in equation (2) and let $z \equiv w\ell$ be the corresponding taxable income.\(^10\) We further assume that equation (2) admits a unique, positive solution for $\ell$.\(^11\) The value of the behavioral wedge may depend on the taxpayer’s choice of labor supply, as well as on her tax rate and wage rate. Crucially, we also allow the wedge to depend on the government policy parameter $\alpha$. We call the policy indexed by this parameter a bias alteration policy.\(^12\) It is this feature that distinguishes our paper from most prior work in behavioral optimal tax theory, and will receive the bulk of our attention below.

To understand the role of the behavioral wedge, notice that the right-hand side of equation (2) is the first derivative of the agent’s private experienced utility with respect to labor supply: the discrepancy between the true marginal private benefit and cost of increasing labor supply. If $\tau^b = 0$, equation (2) is simply the first-order condition of an experienced utility maximizing agent, as in the standard model. Thus, the behavioral wedge determines the extent to which the taxpayer misoptimizes when choosing her labor supply.

We define the agent’s internality as the reduction in experienced utility that results from the agent’s biases.\(^13\) Thus, the behavioral wedge can also be described as a measure of the marginal internality associated with the taxpayer’s choice of labor supply, in the sense that it describes the marginal change in the agent’s internality associated with increasing $\ell$. Note that, in some other work, the term internality refers to a specific type of bias, whereas our definition implies that all biases can induce internalities.\(^14\)

\(^9\)Our modeling framework draws on Farhi and Gabaix (2018). This kind of approach to modeling biased taxpayer behavior in normative public finance research was brought to prominence by Mullainathan et al. (2012). A common complimentary approach to welfare analysis with tax rate misperception was introduced by Chetty et al. (2009). In such models the relative responsiveness of demand to taxes and prices (usually denoted as $\theta$) can serve as a sufficient statistic for the welfare costs of taxation given the taxpayer’s misperception. We do not use this approach here, but interested readers can find a discussion that connects it to our modeling framework in appendix A.1.

\(^10\)Note that we define the behavioral wedge as a function that exists independently of the right-hand side of equation (2). That is to say, we treat the behavioral wedge as a primitive feature of the economy which (along with the agent’s experienced utility function) generates a demand function. This differs somewhat from the presentation of wedges in Farhi and Gabaix (2018), who take the biased agent’s demand function as given and define the behavioral wedge as the resulting discrepancy between tax-inclusive prices and the dollar value of marginal utility at the agent’s demand. Our approach permits a more intuitive description of the behavioral responses of biased agents which will prove useful to our discussion.

\(^11\)For example, suppose that for all $\ell \geq 0$, $\tau^b$ is finite and $\partial \tau^b / \partial \ell > -\psi(\ell)$. These are sufficient conditions for the existence of a unique, positive solution given the assumptions we made about $\psi$ above.

\(^12\)Notice that the type of policy we consider here would have no impact on unbiased taxpayers. However, many policies of interest might impact taxpayer biases while simultaneously changing the incentives of unbiased agents. We provide some discussion of such policies in section 3.

\(^13\)That is to say, the difference between her experienced utility at her actual choice and her experienced utility at the counterfactual choice she would have made if unbiased (i.e. if $\tau^b = 0$ for all $\ell$).

\(^14\)For example, Farhi and Gabaix (2018) often use this term to describe biases resulting from an agent that acts so as to maximize some decision utility function rather than their experienced utility.
Planner’s Problem

In our model, the planner chooses the tax rate and a set of costly policy actions designed to influence taxpayer biases. Specifically, the social planner’s problem is to choose the tax system \((t, \alpha)\) that maximizes the welfare function

\[
W(t, \alpha) \equiv u(\ell) + v(g),
\]

where \(g \equiv tz - a(\alpha)\) is public expenditure and \(a(\alpha)\) is the administrative cost associated with some choice of the bias alteration policy \(\alpha \geq 0\). We assume that \(v_g > 0\), and \(v_{gg} < 0\) for all \(g\). Furthermore, let \(a_\alpha > 0\) for all \(\alpha\) and \(a(0) = 0\).

Example: Tax Rate Misperception

Because the properties of the behavioral wedge depend on the nature of the behavioral bias under consideration, it will be useful for us to have a specific example to rely on for building intuition about our results. Suppose that the agent misperceives the income tax rate as \(t^s(t, \alpha) \in [0, 1]\) when it is in fact \(t\), but is otherwise rational. Then she will behave as if she is maximizing the decision utility function

\[
u^s(\ell) \equiv (1 - t^s(t; \alpha)) w\ell - \psi(\ell).
\]

In this case, the behavioral wedge takes the form \(\tau^b = [t^s(t, \alpha) - t] w\). Here, \(\alpha\) might represent, for example, investment in a taxpayer education program designed to improve the accuracy of the taxpayer’s perceptions. Alternatively, it could represent investment in some kind of disinformation program, designed to make agents believe taxes are lower than they really are.

2.2 Optimal Tax Rate

At the optimal tax rate \(t\) (when there is an interior solution)\(^{17}\)

\[
W_t = -z + v_g \cdot (z + tz_t) + [(1 - t) w - \psi]\ell_t = 0.
\]

\(^{15}\)Models of tax rate misperception have been discussed extensively in the behavioral public finance literature. See, for example, Chetty et al. (2009); Goldin (2015); Taubinsky and Rees-Jones (2017); Farhi and Gabaix (2018). Chetty et al. (2009), Goldin (2015), Taubinsky and Rees-Jones (2017), and Farhi and Gabaix (2018).

\(^{16}\)This follows from the first-order condition of this misperceiving taxpayer: \([1 - t^s(t, \alpha)] w = \psi\ell\). The behavioral wedge of this taxpayer must be such that equation (2) is equivalent to this first-order condition. Notice that in this example treating decision utility as the model primitive that produces observed behavior is equivalent to treating the behavioral wedge as the model primitive, as long as the taxpayer’s decision utility maximization problem has a unique interior solution.

\(^{17}\)A sufficient condition for a unique interior optimal tax rate is that \(W\) is strictly concave in \(t\): \(W_{tt} < 0\) for all \(t\). The second derivatives of the welfare function are presented in appendix A.2.
The first two terms in this expression are present in the non-behavioral formulation of the optimal tax problem (i.e., when $\tau^b = 0$). The third component is attributable to taxpayer misoptimization. Recall, the behavioral wedge $\tau^b$ describes how a marginal increase in $\ell$ changes the taxpayer’s private experienced utility. In the standard model without behavioral biases, the envelope condition implies that this term will be zero. If it is positive, the taxpayer is under-working relative to her private experienced utility maximizing choice; if negative, she is over-working.

Let $\tilde{\tau}^b \equiv \frac{\tau^b}{wv_g}$. This normalized behavioral wedge measures the marginal internality of the agent’s choice of taxable income—rather than labor supply—in terms of the marginal social value of public goods expenditure ($v_g$). Rearranging equation (5), we obtain

$$\frac{t + \tilde{\tau}^b}{1 - t} = \left(\frac{v_g - 1}{v_g}\right) \frac{1}{E^{1-t}_z},$$

where $E^{1-t}_z \equiv -\frac{1-t}{zt}$ is the ETI.\footnote{The result is derived from equation (5) via the following intermediate steps:}

$$(v_g - 1) z + \left(\frac{t + \tilde{\tau}^b}{1 - t}\right) z t = 0$$

$$\left(\frac{v_g - 1}{v_g}\right) z - \left(\frac{t + \tilde{\tau}^b}{1 - t}\right) \left(-\frac{1 - t}{zt}\right) z t = 0$$

when $\tilde{\tau}^b = 0$, this equation reduces to the standard optimal tax formula.

The inclusion of the normalized behavioral wedge in equation (6) ensures that the impact of the tax rate on the size of the internality is accounted for when determining the optimal tax rate. Again returning to our example, suppose the taxpayer consistently underperceives the tax rate so that $t^s (t, \alpha) < t$ for all $t$. This means the agent will work more than is privately optimal because she believes her private after-tax return to labor supply is higher than it really is. The resulting negative marginal internality ($\tilde{\tau}^b = \frac{t^s - t}{v_g} < 0$) puts upward pressure on the optimal tax rate because increasing the rate induces a decrease in labor supply that partially corrects the tendency to overwork.

Differences from a Model with Externalities

Variants of equation (6) appear frequently in the existing behavioral public finance literature, including Farhi and Gabaix (2018). As they note, such optimal tax rate formulae may give the impression that there is no difference between optimal taxation with internalities and with externalities, at least in the case of a representative agent. However, this superficial similarity masks some important differences. For one thing, note that although $\tilde{\tau}^b$ is analogous to a classic Pigouvian wedge in that it represents marginal social
costs/benefits the taxpayer does not internalize, a behavioral wedge differs in that its value can be directly impacted by tax system policy parameters. For example, in the case of tax rate misperception, $\tilde{\tau}^b$ depends directly on $t$. By contrast, in standard tax analyses with externalities the marginal social cost/benefit of some externality-generating good only indirectly depends on policy choices, via their impact on consumption. That distinction is a primary focus of this paper and is explored further in section 2.3, where we discuss the optimal choice of $\alpha$.

The second key difference between an optimal tax analysis with behavioral biases and one with externalities is that behavioral biases (by definition) directly influence taxpayer behavioral responses. This is not a feature of most analyses of externalities. To see this, consider the comparative statics of consumer behavior in our model. Applying the implicit function theorem to equation (2), we can see that the responsiveness of labor supply to changes in the tax rate is described by

$$\ell_t = -\left(\frac{w + \partial\tau^b/\partial t}{\psi_{\ell t} + \partial\tau^b/\partial \ell}\right).$$

(7)

Absent behavioral biases, this expression reduces to $\ell_t = -\frac{w}{\psi_{\ell t}}$. Notice, behavioral biases enter the right-hand side of the expression in two ways. First, the inclusion of $\partial\tau^b/\partial t$ in the numerator reflects the fact that, in response to a small tax change, a biased agent will not necessarily act as if their marginal return to labor supply has decreased in proportion to their wage rate. Second, the inclusion of $\partial\tau^b/\partial \ell$ in the denominator, reflects the fact that, when considering making a small change to their labor supply, a biased agent will not necessarily act as if their marginal cost of labor supply adjusts according to the true curvature of their labor supply function ($\psi_{\ell t}$).\(^{19}\)

To be more concrete about this issue, consider our tax rate misperception example once again. Given the behavioral wedge in that case, we have $\frac{\partial\tau^b}{\partial \ell} = (\frac{\partial\tau^c}{\partial \ell} - 1) w$ and $\frac{\partial\tau^b}{\partial t} = 0$, so that $\ell_t = -\frac{w}{\psi_{\ell t}} (\frac{\partial\tau^c}{\partial \ell})$. Thus, unless the taxpayer correctly perceives the magnitude (and sign) of tax rate changes (i.e. $\frac{\partial\tau^c}{\partial \ell} = 1$), she will under- or over-estimate the change in the marginal return to her labor supply. To see how this can influence optimal tax rates, suppose the taxpayer underperceives the magnitude of tax rate changes so that $\frac{\partial\tau^c}{\partial \ell} < 1$ for all $t$. If she also underperceives the level of tax rates so that $t^* < t$ for all $t$ misperception of this kind will, in general, have the effect of attenuating the ETI at any given choice of $t$.\(^{20}\) Equation (2.2) shows that this would push up the optimal tax rate.\(^{21}\)

\(^{19}\)Note, the comparative statics results we present in equation (7) and elsewhere in this paper are somewhat novel. Similar results for the case of general behavioral biases are not present in Farhi and Gabaix (2018). These results are enabled by the different approach we have taken to modeling general behavioral biases, as discussed in footnote 10.

\(^{20}\)The agent’s ETI is $E_{1-t}^{1-t} = \frac{\partial\tau^c}{\partial \ell}$. Note that if $t^* < t$, the agent’s labor supply will be higher than that of an unbiased agent at any given tax rate $t$. Thus, as long as $\psi_{\ell t}$ is not too negative over the relevant range of values, the ETI of a biased agent will be attenuated relative to that of an unbiased agent at any given tax rate $t$.

\(^{21}\)The role of $\partial\tau^c/\partial \ell$ in equation (7) can be understood by considering the example of an agent who misperceives their cost of
The Role of Bias Alteration Policy

The discussion above helps to motivate our analysis in the subsections that follow. Note that equations like (6) are familiar to the literature on optimal taxation with behavioral biases. As argued by Farhi and Gabaix (2018) and others, such optimal tax formulae suggest two types of “sufficient statistics” for welfare analysis with behavioral agents: elasticities that measure the behavioral response to taxation; and behavioral wedges that measure marginal internalities.

Consistent with previous contributions to the tax systems literature, our objective is to highlight the fact that neither of these types of “sufficient statistics” should be thought of as structural features of the economy. Because taxpayer biases may be endogenous to non-rate government policies, there may be scope for a social planner to beneficially influence both the behavioral responses and marginal internalities of taxpayers through the use of such policy instruments. In principle, the social planner should be interested not in the optimal tax rate given the current behavioral wedge and ETI, but the optimal tax rate when bias alteration policy is itself set to its optimal value.

To be more explicit, let \( ˆt(\alpha) \) be the optimal tax rate described by equation (6) for some choice of bias alteration policy \( \alpha \). How does this optimal rate change in response to a marginal increase in \( \alpha \)? Applying the implicit function theorem to equation (5) we obtain

\[
\frac{d ˆt(\alpha)}{d \alpha} = \frac{W_{t\alpha}}{W_{tt}}.
\]

For simplicity, suppose that \( v_g \) is constant. Then, because \( W_{tt} < 0 \) at any interior optimum tax rate, the sign of \( \frac{d ˆt(\alpha)}{d \alpha} \) is determined by the sign of

\[
W_{t\alpha} = (v_g - 1) z + [v_g tw + \tau^b] \ell, \quad \frac{\partial \tau^b}{\partial \alpha} + \frac{\partial \tau^b}{\partial \ell} \ell.
\]

This equation highlights the three ways \( \alpha \) influences the optimal tax rate. The \( M \) term accounts for the impact \( \alpha \) has on the mechanical effect of taxation, \( B \) accounts for the impact \( \alpha \) has on the agent’s behavioral response to taxation, and \( I \) accounts for the impact \( \alpha \) has on the marginal internality associated with labor supply. It is the \( I \) term that is unique to a setting with behavioral agents. To understand its role, suppose \( \tau^b > 0 \), \( \frac{\partial \tau^b}{\partial \alpha} < 0 \), and \( \frac{\partial \tau^b}{\partial \ell} = 0 \). Then increasing \( \alpha \) reduces the positive marginal internality associated with \( \ell \). This in turn implies that the marginal net benefit of raising the tax rate (and thus decreasing \( \ell \)) is increasing in \( \alpha \) (i.e. that \( I > 0 \)).

\[22\] In particular, it is a simple case of a result presented by Farhi and Gabaix (2018).

\[23\] If \( v_g \) is not constant, an additional term must be added to equation (8) which accounts for the impact of a marginal change in \( \alpha \) on the marginal value of public expenditure. See appendix A.2.
2.3 Optimal Bias Alteration

We now turn our attention to the optimal choice of bias alteration policy $\alpha$ (given $t$). If the optimal choice of $\alpha$ is an interior solution, it will satisfy the first-order condition

$$W_\alpha = v_y \cdot (tz_\alpha - a_\alpha) + [(1 - t)w - \psi] \ell_\alpha = 0. \tag{9}$$

This condition can be re-written as

$$(t + \tilde{\tau}^b) z_\alpha = a_\alpha. \tag{10}$$

The right-hand side of this equation is the marginal social cost of increasing $\alpha$—marginal administrative costs ($a_\alpha$)—while the left-hand side represents the marginal benefit of doing so—the marginal social benefit of taxable income ($t + \tilde{\tau}^b$) multiplied by the taxpayer’s behavioral response to a marginal increase in bias alteration ($z_\alpha$). The marginal social benefit of taxable income consists of marginal tax revenue ($t$) and the taxpayer’s marginal internality of taxable income ($\tilde{\tau}^b$).

Thus, the condition states that the social planner will choose a value of $\alpha$ to equalize the marginal social cost and social benefit of bias alteration policy.

We can also derive an elasticity-based formula for optimal bias alteration:

$$E^\alpha_z = \frac{\alpha a_\alpha}{(t + \tilde{\tau}^b) z}, \tag{11}$$

where $E^\alpha_z = \frac{\partial z_\alpha}{\partial z}$ is the bias alteration elasticity (BAE) of taxable income.

Equations (10) and (11) are perhaps surprisingly simple. In spite of the complex role bias alteration policy plays—simultaneously influencing the taxpayer’s behavioral response and her marginal internality—these equations suggest sufficient statistics for bias alteration policy that are very similar to those suggested by Keen and Slemrod (2017) for other non-rate policy instruments. Relative to similar expressions in their work, the only new term that appears here is the behavioral wedge $\tilde{\tau}^b$. While such wedges might prove difficult to estimate, this problem must also be overcome for empirically evaluating the optimal tax rate formula in equation (6).

---

24 Again, if $W$ is strictly concave in $\alpha$, there exists a unique interior optimum for $\alpha$. This holds if $W_{\alpha \alpha} < 0$ for all $\alpha$.

25 Recall, $\tilde{\tau}^b$ measures the marginal internality of taxable income relative to the marginal social benefit of public expenditure, so all costs and benefits in equation (10) are measured in commensurate units.

26 In their most general formulation of nudging, Farhi and Gabaix (2018) model nudges in a similar manner to our bias alteration policies, assuming that a nudge alters taxpayer behavior in some way but does not enter the taxpayer’s budget constraint. The main difference is that, whereas we assume bias alteration policies have accompanying administrative costs, Farhi and Gabaix (2018) assume that nudges can impose private costs on taxpayers, perhaps due to some innate distaste for being nudged. We note, however, that their analysis of this general case does not extend beyond presenting an analogue of our equation 10. For most of their paper, they model nudges as a kind of psychic tax that the planner can impose on commodities. By contrast, we retain our focus on a general model of biases and bias alteration throughout this paper, presenting numerous results that have no analogue in Farhi and Gabaix (2018).
The Role of the Bias Alteration Elasticity

The simplicity of these optimal bias alteration formulae results from the fact that the BAE is a sufficient statistic for the impact of \( \alpha \) on the marginal internality. To see how this works, notice that the BAE is

\[
E^\alpha_z = -\frac{\alpha}{z} \left( \frac{\psi t^b}{\psi t^* + \frac{\partial z}{\partial t}} \right).
\]

Thus, the taxpayer’s behavioral response to \( \alpha \) is proportional to the marginal effect of bias alteration on the taxpayer’s marginal internality (\( \frac{\partial z}{\partial \alpha} \)).

In addition to its value as a sufficient statistic, the BAE can provide insight into the nature of the taxpayer’s biases. For example, consider two circumstances under which taxable income does not respond to \( \alpha \) (i.e. when \( E^\alpha_z = 0 \)). This implies \( \frac{\partial z}{\partial \alpha} = 0 \), which indicates either that the taxpayer is unbiased or the bias alteration program is ineffective at the margin. Therefore, if we have \( E^\alpha_z \neq 0 \), we know that the taxpayer is indeed biased and, furthermore, that the bias alteration policy in question influences her biases.

The sign of a non-zero bias policy elasticity can be informative about the type of bias the taxpayer faces. For instance, in our tax rate misperception example, if \( E^\alpha_z < 0 \) at some \( t \) and \( \alpha \), this implies that \( \frac{\partial z}{\partial \alpha} > 0 \).

Suppose the policy indexed by \( \alpha \) is a taxpayer education program that (weakly) improves the accuracy of the taxpayer’s perception of the tax rate: \( \frac{\partial z}{\partial \alpha} \leq 0 \). From this assumption, we can conclude that the taxpayer must be underperceiving the tax rate (\( t^* < t \)).

When Is Bias Alteration Beneficial?

As equation (9) highlights, \((t + \tilde{t}^b) z_\alpha > 0\) is a necessary condition for additional investment in bias alteration policy to be welfare-enhancing. There are three scenarios in which this will obtain. First, suppose the taxpayer has a positive marginal internality (\( \tau^b > 0 \)). In this case, increasing \( \alpha \) will only be beneficial if it increases the taxable income so as to produce benefits in the form of both tax revenue and a reduced internality for the taxpayer. As noted above \( z_\alpha > 0 \iff \frac{\partial z}{\partial \alpha} < 0 \), so a beneficial bias alteration policy in this context would operate by reducing the magnitude of the agent’s behavioral wedge. That is to say, when a taxpayer exhibits a positive marginal internality, bias alteration policies are beneficial only if they serve to correct the taxpayer’s biases.

Alternatively, suppose the taxpayer has a negative marginal internality. In this case, whether or not increasing taxable income is socially beneficial depends on the magnitude of the marginal internality. If \( t > -\tilde{t}^b \) then

\[ \frac{\partial z}{\partial \alpha} < 0 \implies \frac{z_\alpha}{z} > 0. \]

This implies that a beneficial bias alteration policy in this context would operate by reducing the magnitude of the agent’s behavioral wedge. That is to say, when a taxpayer exhibits a negative marginal internality, bias alteration policies are beneficial only if they serve to correct the taxpayer’s biases.

\[ ^{27} \text{Applying the implicit function theorem to equation (2) yields } \ell_\alpha = -\frac{a_{\psi b}^t}{\psi t^* + \frac{\partial z}{\partial t}}. \text{ The result follows.} \]
the marginal internality is not sufficiently negative to outweigh the social benefit of additional tax revenue so, as in the case above, increased bias alteration will be beneficial only if it increases taxable income. Once again, this implies that \( \frac{\partial \tau}{\partial b} < 0 \), but in the context of a negative marginal internality this means that bias alteration policy is beneficial here only insofar as it increases the magnitude of the agent’s behavioral wedge.

Thus, when the taxpayer exhibits a negative marginal externality there are circumstances where it is actually beneficial for the planner to exacerbate the taxpayer’s biases. On the other hand, if \( t < -\tilde{\tau}_b \) the taxpayer’s negative marginal internality is so severe as to render a marginal increase in taxable income welfare-decreasing. In this case, as in the case of positive marginal internalities, bias alteration policy is beneficial only insofar as it exhibits corrective properties because \( z_\alpha < 0 \iff \frac{\partial \tau}{\partial b} > 0 \).

Notice, that there is a critical threshold where the desired direction of bias alteration policy changes: \( \tilde{\tau}_b = -t \). Whenever \( \tilde{\tau}_b \neq -t \), there is scope for the planner to pursue bias alteration policies which move \( \tilde{\tau}_b \) closer to \(-t\), though they should do so only when cost-effective.\(^{28}\)

### Biases at the Optimum

To better understand what bias alteration policy can potentially accomplish, consider the case where it is costless. That is to say, suppose \( a(\alpha) = 0 \) for all \( \alpha \) and, consequently, \( a_\alpha = 0 \). Then at the optimal bias alteration policy described by equation (10) it must be the case that either \( \tilde{\tau}_b \equiv \frac{\tau}{wv_g} = -t \) or \( z_\alpha = 0 \).

First, suppose \( \tilde{\tau}_b = -t \) at the optimum. Combining this with equation (2) implies that \( \ell \) satisfies

\[
(1 - t) w + v_g tw = \psi_\ell.
\]

Because \( v_g tw \) is the social value of the tax revenue generated by a marginal increase in taxpayer’s labor supply, this equation implies that when \( \tilde{\tau}_b = -t \) the taxpayer’s biases induce them to behave as if she were a rational agent who directly incorporated the social value of her tax payments into her labor supply decisions.

As we discuss in the next section, this scenario has the interesting implication that the optimal tax system can attain the first-best welfare outcome.

Alternatively, suppose that \( z_\alpha = 0 \) at the optimal bias alteration policy. This case reflects the possibility that Fadlon and Laibson (2017) note that, in the context of savings policy, apparently rational behavior by taxpayers might in fact result from deliberate government policy choices intended to induce such behavior. Our analysis above makes it clear when a similar conclusion might hold in a tax policy setting. When the taxpayer exhibits a positive marginal internality or a sufficiently negative one, then both the taxpayer’s private welfare and social welfare are increased when taxpayer bias is alleviated. Consider the case of a positive marginal internality. Suppose it is feasible for the planner to eliminate the taxpayer’s bias but not to induce the desired negative marginal internality \( \tilde{\tau}_b = -t \). Then optimal bias alteration—absent costs—would imply \( \tilde{\tau}_b = 0 \) and, as in Fadlon and Laibson (2017), our model implies that apparently rational behavior by taxpayers could be the product of optimal policy intervention. But note, if it is feasible and cost-effective, our model implies that the planner should attempt to induce a specific form of biased behavior by the taxpayer: a negative marginal internality such that \( \tilde{\tau}_b = -t \).
any given bias alteration policy need not necessarily be capable of altering taxpayer biases so as to ensure that \( \tilde{\tau}^b = -t \). In such cases, absent administrative costs, investment in any bias alteration policy that moves \( \tilde{\tau}^b \) closer to \(-t\) should be increased until it is no longer effective at doing so.\(^{29}\) Exactly how close \( \tilde{\tau}^b \) will be to \(-t\) at the optimum in this circumstance depends on the functional form of \( \tau^b \).

Reintroducing administrative costs provides another reason to expect that \( \tilde{\tau}^b \neq -t \) at the optimum. In general, it may not be cost-effective for the planner to pursue bias alteration that succeeds in manipulating the taxpayer into behaving as if she were fully taking into account the fiscal externality associated with her labor supply choice. Rather, equation (10) shows that at the optimum choice of \( \alpha \), \( \tilde{\tau}^b = \frac{a_\alpha}{\alpha} - t \). In this case, equation (2) implies that, at the optimal \( \alpha \), the taxpayer’s choice of \( \ell \) satisfies

\[
(1 - t) w + v_g w \left[ t - \frac{a_\alpha}{\alpha} \right] = \psi_\ell. \tag{14}
\]

Thus, in the general case with costly bias alteration policy, optimal bias alteration policy involves inducing taxpayer biases which only partially induce the taxpayer to behave as if she had internalized the fiscal externality of her behavior. The degree of internalization which occurs at the optimum is decreasing in \( a_\alpha \).\(^{30}\)

### 2.4 Characterizing the Optimal Tax System

The optimal tax system consists of jointly optimal choices of the tax rate and bias alteration policy. Suppose \((t^*, \alpha^*)\) is an interior solution to the planner’s problem. Then both \( W_\ell (t^*, \alpha^*) = 0 \) and \( W_\alpha (t^*, \alpha^*) = 0 \) must be satisfied. Let \( \ell^* \equiv \ell (w, t^*, \alpha^*) \) and \( z^* \equiv w \ell^* \). We consider two key cases.

**The Optimal Tax System without Administrative Costs**

As we did in the preceding subsection, it is instructive to consider the case without administrative costs.

As we noted above, this implies that whenever feasible, the planner will choose \( \alpha \) so as to ensure \( \tilde{\tau}^b = -t \).

\(^{29}\)This is consistent with our discussion above regarding when bias alteration is desirable.

\(^{30}\)Note that although in our example ideal optimal bias alteration involves inducing a negative marginal internality \((\tilde{\tau}^b \neq -t)\), other cases might arise. Consider the question of optimal commodity taxation with some externality-generating good. Suppose the externality is atmospheric in nature, so that in the standard model optimal tax treatment involves levying a Pigouvian tax on the externality-generating good equivalent to the marginal social cost of the externality. As has been noted by Farhi and Gabaix (2018) among others, this prescription may not hold when agents are biased. For example, misperception of commodity taxes might imply that optimal taxation involves levying corrective tax rates that exceed the marginal social cost of the externality or include subsidizing substitutes or taxing complements of the externality-generating good. In this case, ideal bias alteration could actually mean ensuring that the agent behaves as if she were a rational agent who accurately perceives any corrective tax levies on the externality-generating good. That is to say, in contrast to our example, the planner in this case would like to use such policies to eliminate the taxpayer’s internality.
Suppose this is indeed possible. Then the first-order welfare effect of increasing the tax rate simplifies to

\[ W_t = (v_g - 1) z^*. \]  

(15)

Because \( W_t = 0 \) at the full optimum, this implies that it is optimal for the planner to increase the tax rate until the marginal value of public expenditure is equalized with the marginal utility of consumption \( (v_g = 1) \).

This result is quite intuitive: because optimal bias alteration involves manipulating the taxpayer into behaving as if she were accounting for the fiscal externality of her actions, the social planner does not need to worry about the agent’s behavioral response to taxation when evaluating the first-order welfare effects of taxation.

This is not because there is no behavioral response by the taxpayer, but rather reflects the fact that no matter what the taxpayer’s behavioral response to taxation is, if \( \tilde{\tau}^b = -t \) then the marginal social cost of forgone tax revenue due to the agent’s response is exactly offset by the negative marginal internality of taxable income.

That is to say, the behavioral response caused by an increase in the tax rate not only reduces tax revenue but also (beneficially) reduces the size of the taxpayer’s internality. At the full optimum, these effects have a net zero first-order influence on welfare: they cancel each other out.

That said, it will in fact be true that in this full optimum the taxpayer will choose the same value of labor supply that a rational agent would in response to a lump-sum tax. To see this, notice that evaluating equation (13) when \( v_g = 1 \) implies that the agent’s choice of \( \ell \) at the optimum will satisfy \( w = \psi \ell \). It must be noted, however, that this does not in general mean that the taxpayer’s ETI at the optimum is zero. Referring to equation (7), we can see that this would only occur if \( \frac{\partial \tau^b}{\partial \ell} = -w \), but this is not necessary to ensure \( \tilde{\tau}^b = -t \).

A biased agent can, in principle, be induced to choose labor supply at a given tax rate as if she were a rational agent responding to lump-sum taxes, even if she is responsive to changes in that the tax rate.

This implies that in the optimal tax system without administrative costs, if \( \tilde{\tau}^b = -t \) is feasible, the taxpayer’s biases will be adjusted so as to render the first-best welfare outcome attainable, because \( v_g = 1 \) is exactly the outcome that would be optimal if the taxpayer were unbiased and the social planner had access to a lump-sum tax. This generalizes a result from Goldin (2015), who shows—in a model without administrative costs—that the optimal choice of tax salience similarly may permit the social planner to obtain the first-best welfare outcome.\footnote{Goldin’s model involves choosing tax salience by combining two tax instruments of differing salience that apply to the same good. In his model the first-best-enabling level of tax salience is thus only attainable if it can be expressed as a convex combination of the salience of the two component tax instruments. This issue is conceptually similar to one we highlighted above in our model: it need not be the case that a given bias alteration policy is capable of inducing \( \tilde{\tau}^b = -t \).}

As we discussed in the previous section, it is important to note that it may not be feasible to use bias alteration to achieve the desired welfare outcome. To see this, note that \( \tilde{\tau}^b \equiv \frac{\tau^b}{w v_g} \) and therefore \( \tilde{\tau}^b = -t \) implies that \( v_g t w + \tau^b = 0 \). Inserting this into equation (5) gives the result.

\footnote{To see this, note that \( \tilde{\tau}^b \equiv \frac{\tau^b}{w v_g} \) and therefore \( \tilde{\tau}^b = -t \) implies that \( v_g t w + \tau^b = 0 \). Inserting this into equation (5) gives the result.}
alteration policy to induce $\tilde{\tau}^b = -t$ for any given tax rate $t$. This situation is even more important to keep in mind when considering the optimal tax system. In a rational inattention-style model, where agent biases are rationalized by some optimal allocation of cognitive processing, it might be the case that bias alteration policy can induce $\tilde{\tau}^b = -t$ for low values of $t$ but not for higher values of $t$, when the private costs of holding such biases would be larger for the agent.

To conclude, the optimal tax system without administrative costs satisfies

$$\tilde{\tau}^b (\ell^*; w, t^*, \alpha^*) = -w t^*$$

and

$$t^* z^* = [v_g]^{-1}$$

where $[v_g]^{-1} (\cdot)$ is the inverse function of $v_g (\cdot)$.

The Optimal Tax System with Administrative Costs

In the previous subsection, we noted that optimal bias alteration at any given tax rate implies setting $\alpha$ so that $\tilde{\tau}^b = \frac{\alpha z}{\alpha} - t$ or, alternatively, $v_g w t + \tau^b = \frac{v_g a}{\alpha}$. Therefore, the first-order welfare effect of increasing the tax rate when bias alteration policy is optimally set is

$$W_t = (v_g - 1) z + v_g a \left( \frac{z_t}{z^*_\alpha} \right).$$

Thus, in general, the optimal tax system will not involve tax revenues sufficiently high to drive $v_g$ below 1, as $\frac{\alpha z}{\alpha}$ is generally non-zero. If $\frac{\alpha z}{\alpha} < 0$ in the optimal tax system, then marginal social value of public expenditure will be higher that the marginal utility of consumption ($v_g > 1$) at the optimum. The most natural circumstance satisfying this condition is $z_t < 0$ and $z_\alpha > 0$. This implies that $\frac{\partial \tau^b}{\partial t} < 0$ because $\ell_\alpha = -\frac{\partial \tau^b}{\partial \ell} > 0$. That is to say, if the taxpayer exhibits a positive marginal internality or a sufficiently low magnitude negative marginal internality, bias alteration policy will be beneficial if it reduces $\tau^b$. In the optimal tax system, $\tilde{\tau}^b = \frac{\alpha z}{\alpha} - t > -t$ and $v_g = \frac{z}{z^*_\alpha(\frac{1}{\alpha})} > 1$.

Alternatively, suppose that $z_t < 0$ and $z_\alpha < 0$. This implies that, at the optimal tax system, $v_g < 1$. This is a somewhat strange result which suggests the government is actually collecting more tax revenue than they would in the first-best case. This occurs because in this case $\tilde{\tau}^b = \frac{\alpha z}{\alpha} - t < -t$, so the taxpayer exhibits a very large marginal negative internality. Indeed, the taxpayer is over-working to such a degree that she ends

---

33Unless there exists a cost-effective bias alteration policy which induces $z_t = 0$ at tax rates high enough to deliver $v_g = 1$. 
up giving the government an inefficiently large amount of tax revenue. Although it is worthwhile to spend some funds on bias alteration to alleviate this problem, the administrative costs of such a program imply that complete correction might be undesirable. Thus, $v_g$ can be less than one in the optimal tax system.

To summarize, assuming it is feasible to use bias alteration policy to induce $\tilde{\tau}_b = -t$ for any given tax rate $t$, the optimal tax system with administrative costs can be described as one which gets as close as possible to the first-best optimum without administrative costs where $\tilde{\tau}_b = -t$ and $v_g = 1$, but is constrained from doing so fully by administrative costs.\footnote{In appendix B, we show that heterogeneity provides an additional independent force that constrains the planner from attaining the first-best.}

The resulting optimal system will satisfy the following equations

$$\tau^b (t^*; w, t^*, \alpha^*) = \left[ \frac{a_\alpha (\alpha^*)}{z^*} - t^* \right] w v_g (g^*), \quad (19)$$

$$t^* z^* = g^* + a (\alpha^*), \quad (20)$$

and

$$g^* = [v_g]^{-1} \left( \frac{z^*}{z^* + a_\alpha (\alpha^*) \left( \frac{z^*}{z^*} \right)} \right). \quad (21)$$

Comparing these equation to (16) and (17) shows how administrative costs alter the optimal tax system.

A Necessary Condition Without Behavioral Wedges

The planner’s first-order conditions (equations 6 and 11) imply that the optimal tax system $(t^*, \alpha^*)$ satisfies the condition

$$\left[ t + \tilde{\tau}_b \right] \frac{z E^w}{a_\alpha} = 1 = \left( \frac{[t + \tilde{\tau}_b] E^w_{z}^{1-t}}{1-t} \right) \left( \frac{v_g - 1}{v_g} \right). \quad (22)$$

To build some intuition about equation (22), consider the alternative presentation

$$\left[ t + \tilde{\tau}_b \right] \frac{z_{\alpha}}{a_\alpha (i)} = 1 = -\left[ t + \tilde{\tau}_b \right] \frac{z_t}{a_\alpha (ii)}. \quad (23)$$

Starting from an optimal tax system satisfying this condition, there are two ways to change the taxpayer’s labor supply: by using $\alpha$ or by using $t$. Part $(i)$ of this equation is the ratio of the marginal social benefit of increasing $z$ using $\alpha$ to the marginal social cost of doing so. Part $(ii)$ is similarly a ratio of the marginal social benefit to the marginal social cost of increasing $z$ by lowering the tax rate.\footnote{The benefit is the social value of increased tax revenue and of the change in internality due to the taxpayer’s behavioral response to the tax change. The cost is the loss in social welfare due to the transfer of taxable income from the government to} Thus, equation (22)
reflects the fact that—at an interior optimum—the social planner will choose policy parameters to equalize the marginal social benefit and marginal social cost of each policy lever and that, therefore, the marginal benefit-cost ratios associated with each policy parameter must be equal to one another at such an optimum.

The former feature of the optimal tax system is particularly interesting because it can be examined without any reference to behavioral wedges. This obtains because the numerators of both \((i)\) and \((ii)\) are proportional to \([t + \tilde{\tau}_b]\), so the condition that the marginal benefit-cost ratios are equal to one another can be written as

\[
\frac{zE_z^{\alpha}}{\alpha a_\alpha} = \frac{1}{1-t} \frac{E_z^{1-t}}{v_g - 1}.
\]  

This equation constitutes a necessary condition for an interior optimal tax system that, strikingly, does not contain any types of sufficient statistics beyond those discussed in Keen and Slemrod (2017) despite the introduction of behavioral biases into the model.\(^{36}\) As noted above, the reason it is possible to determine whether the marginal social benefit-cost ratios of \(t\) and \(\alpha\) are equal without knowing the behavioral wedge is that both policy instruments produce benefits in the same way: by altering the taxpayer’s choice of taxable income. The value of a marginal unit of tax revenue \((t + \tilde{\tau}_b)\) is the same for both instruments, and thus we do not need to know its precise value when checking for equality of the marginal benefit-cost ratios.

**Welfare Improvements Without Behavioral Wedges**

The same logic that underlies equation (24) can be used to identify tax system reforms that are welfare-improving when the system is away from the optimum. Consider some marginal change to the tax system which increases \(\alpha\) and simultaneously alters the tax rate in such a way as to leave taxable income unchanged. The welfare impact of such a change can be expressed as

\[
dW = \left. W_t \frac{dt}{d\alpha} \right|_z + W_\alpha \\
= - (v_g - 1) z \left( \frac{z_\alpha}{z_t} \right) - v_g a_\alpha
\]

because \(\frac{dt}{d\alpha} \big|_z = - \frac{z_\alpha}{z_t}\). Therefore, if

\[
-(v_g - 1) z \left( \frac{z_\alpha}{z_t} \right) - v_g a_\alpha > 0,
\]

the taxpayer (the mechanical effect of a tax decrease). All marginal costs and benefits are measured in terms of the marginal value of public expenditure.

\(^{36}\)Indeed, an identical condition could be obtained by combining equations (7) and (10) in their paper and setting \(c_\alpha = 0\).
then the joint change of policy parameters \((dt, d\alpha)\) such that \(d\alpha > 0\) and \(\frac{dt}{d\alpha} = -\frac{z_t}{z_\alpha}\) is welfare-improving. Our preceding discussion highlights the rationale for this result. When the condition above is satisfied, it implies that the marginal benefit-cost ratios of the policy parameters are not equalized. In particular, it implies that the marginal benefit-cost ratio of \(\alpha\) is higher than that for \(t\). Thus, a marginal decrease in \(t\) accompanied by a compensating increase in \(\alpha\) which ensures that realized taxable income remains unchanged will improve social welfare by producing the same \(z\) using a relatively less costly combination of policy tools.

As with equation (24), condition (26) can be evaluated without knowledge of the taxpayer’s perceived tax rate. The intuition behind this result is that, when \(z\) is held constant, marginal changes in \(t\) and \(\alpha\) only have second-order effects on the size of the taxpayer’s internality. To understand this, recall that equations (5) and (9) show that the first-order effects of policy changes on the internality occur entirely via the changes in labor supply they induce. Thus, although in general changing \(t\) and \(\alpha\) certainly alters the size of the taxpayer’s internality, when \(z\) is held constant it does not do so for small policy changes. When considering a large policy change, this approach would need to be adapted to consider second-order effects. For additional discussion, see appendix A.3.

### 3 Tax Enforcement with Optimal Biases

In this section, we introduce the possibility of tax evasion and consider the problem of optimally choosing policy instruments that address it. Tax evasion is a first-order concern in developing countries and a non-trivial one in developed countries. As in the previous section, we depart from Keen and Slemrod (2017) by assuming that the representative taxpayer may exhibit behavioral biases and the planner has access to non-rate tax system instruments that may influence these biases.

#### 3.1 Model Setup

The representative taxpayer’s private experienced utility is

\[
u(\ell, e) \equiv [1 - t] (w\ell - e) + e - \psi(\ell) - c(e, w\ell; p),
\]

where \(e\) is the taxpayer’s choice of evaded income and \(c(\cdot, \cdot; p)\) is a function determining the private cost of evasion for a given enforcement policy \(p \in \mathbb{R}\). The parameter \(p\) might represent features of the tax enforcement regime such as the audit probability, the intensity of audits, or the penalties associated with detected evasion. Note that the costs of evasion depend on both the taxpayer’s choice of evasion and her choice of true labor
income $y \equiv w\ell$.\footnote{Keen and Slemrod (2017) consider this model in an extension of their main results. See Slemrod (2001) for a thorough discussion of this model with rational agents. As he notes, it has some desirable properties, including its relative simplicity as well as its ability to capture interdependence of labor supply and avoidance decisions.} Let $z \equiv w\ell - e$ be the taxpayer’s reported taxable income. As in section 2, we assume that $\psi$ is twice continuously differentiable, $\psi(0) = 0$, $\psi_\ell > 0$, $\psi_{\ell\ell} > 0$, and $\lim_{\ell \to 0} \psi_\ell = \infty$. Further, suppose that $c$ is twice continuously differentiable in all three of its arguments.

As in the previous section, we consider a very general formulation of behavioral biases. In particular, suppose the taxpayer’s labor supply and evasion choices satisfy the following equations:

\begin{equation}
\tau^{b,\ell} (\ell, e; w, t, p, \alpha) = (1 - t) w - \psi_\ell - wc_y, \tag{28}
\end{equation}

and

\begin{equation}
\tau^{b,e} (\ell, e; w, t, p, \alpha) = t - ce. \tag{29}
\end{equation}

Here, $\tau^{b,\ell}$ and $\tau^{b,e}$ are the taxpayer’s behavioral wedges for labor supply and evasion respectively, each defined similarly to the wedge from the last section. They are allowed to depend endogenously on choices of labor supply and evasion, as well as on a number of exogenous parameters. This includes both the parameter $\alpha \in \mathbb{R}$, which represents a bias alteration policy, and the parameter $p$. We assume these wedges are twice continuously differentiable in each of their arguments, that $\frac{\partial \tau^{b,\ell}}{\partial e} = \frac{\partial \tau^{b,e}}{\partial \ell}$, and furthermore, that there exists some unique solution $(\ell, e)$ to the nonlinear system of equations defined by (28) and (29).

**Planner’s Problem**

We analyze the decision of a social planner who chooses a tax system $(t, p, \alpha)$ to solve

\begin{equation}
\max_{t, p, \alpha} W \equiv u(\ell, e) + v(g), \tag{30}
\end{equation}

where $g \equiv t (w\ell - e) - a(p, \alpha)$ is public goods spending and $a(p, \alpha)$ is a continuously differentiable function representing the administrative costs of the non-rate policies $p$ and $\alpha$.

**Example: Dual Misperception**

Suppose the taxpayer exhibits a very particular form of bias: she misperceives both the enforcement parameter $p$ and the tax rate $t$. As she is otherwise rational, her labor supply and evasion choices solve the decision
utility maximization problem

\[
\max_{\ell,e} u^* (\ell, e) \equiv [1 - t^*] (w\ell - e) + e - \psi (\ell) - c (e, w\ell; p^*),
\]

where \( t^* \equiv t^* (t, p, \alpha) \) is her perceived tax rate and \( p^* \equiv p^* (t, p, \alpha) \) is her perceived enforcement parameter.

This setup can, for example, be interpreted as modeling an agent who is an expected utility maximizer but misperceives the probability of being audited (and thus, of evasion detection) as well as the tax rate.\(^{38}\) The parameter \( \alpha \in \mathbb{R} \) represents a bias alteration policy that alters the taxpayer’s perception of the enforcement parameter. For example, it might represent the intensity of an audit threat program, as implemented in Bérgolo et al. (2017).\(^{39}\)

Let \( c^* \equiv c (e, w\ell; p^*) \). The agent’s behavioral wedges will be defined as follows:\(^{40}\)

\[
\tau_{b,\ell} \equiv \left[ (t^* - t) + (c^*_y - c_y) \right] w
\]

\[
\tau_{b,e} \equiv - (t^* - t) + (c^*_e - c_e)
\]

This follows from combining the agent’s first-order conditions with equations (28) and (29).

### 3.2 Optimal Tax Rate

Setting \( W_t = 0 \), we can show that the optimal tax rate satisfies

\[
\frac{t + \left( \frac{\nu_t}{\nu_{t^*}} \right) \tilde{\tau}_{b,y} + \left( \frac{\nu_e}{\nu_{t^*}} \right) \tilde{\tau}_{b,e}}{1 - t} = \left( \frac{v_{t^*} - 1}{v_{t^*}} \right) \frac{1}{E_{t^* - t}},
\]

where \( \tilde{\tau}_{b,y} \equiv \frac{\tilde{\tau}_{b,y}}{w_{t^*}} \) and \( \tilde{\tau}_{b,e} \equiv \frac{\tilde{\tau}_{b,e}}{v_{t^*}} \). The derivation parallels that in section 2.\(^{41}\) This equation differs from the usual optimal tax formula via the inclusion of two behavioral wedge terms in the numerator on the left-hand side: \( \tilde{\tau}_{b,y} \) and \( \tilde{\tau}_{b,e} \) are the marginal internalities labor income and evasion, respectively, each measured in terms of the marginal value of public expenditure. As in section 2, these wedge terms serve to incorporate into the social planner’s calculus the internalities that result from the taxpayer’s biased behavior.

\(^{38}\)Note that this model does not readily capture behavioral biases such as loss aversion or ambiguity aversion, which could be more easily modeled by assuming the taxpayer misperceives the cost of evasion function \( c \) as \( c^* \). Although it cannot capture loss aversion, our model can capture the other key feature of prospect theory: incorrect weighting of probabilities.

\(^{39}\)They conduct an RCT in which firms in different treatment arms received different types of letters from the tax authority. Only some firms received letters containing information about the likelihood of an audit and, among those, some firms received information indicating a higher likelihood than others. Other audit threat programs might also be said to have varied the “intensity” of treatment. For example, Boning et al. (2018) conduct an audit threat RCT in which firms in some treatment arms received letters from the tax authority, while others received in-person visits from representatives of the tax authority.

\(^{40}\)The agent’s first-order conditions are \( (1 - t^*) w - \psi_t - wc^*_y = 0 \) and \( t^* - c^*_e = 0 \).

\(^{41}\)Refer to appendix (C.2) to see the derivatives of the welfare function.
To be more concrete about these internalities, consider our example. The taxpayer’s private experienced utility is lower than it would otherwise be because she misperceives both the tax rate and the costs of evasion. This has two implications. On the one hand, she may be evading too much (too little) at the margin because she thinks the marginal cost of evasion is lower (higher) than it really is or because she thinks the marginal benefit of evasion—in the form of reduced tax liability—is higher (lower) than it really is. On the other hand, she may be supplying too much or too little labor at the margin not only because she misperceives the tax rate, but also the manner in which labor income affects the costs of evasion.\footnote{Note, this implies that the inclusion of behavioral wedges in equation (32) reflects a decision to assign normative significance to misoptimization by the taxpayer with respect to her evasion decisions. This is a natural result of the consequentialist normative framework typical of optimal tax theory. Nonetheless, it might be difficult to convince policymakers that policies which induce sub-optimally low rates of evasion impose a welfare-relevant cost on the taxpayer. Note, however, that our model could also be interpreted as pertaining to some legal forms of tax avoidance.}

**Insufficiency of the Elasticity of Taxable Income**

The optimal tax formula presented in equation (32) highlights a challenge for empirical application of this framework. The terms $\frac{\eta}{\ell^t}$ and $\frac{\xi}{\ell^t}$ measure the proportion of the responsiveness of taxable income to taxation that is accounted for by labor income and evasion responses, respectively. As evasion is generally not directly observable, this poses a challenge for applied welfare analysis. Recall that, in section 2, the ETI was a sufficient statistic for the behavioral response to taxation in the optimal tax formula (equation 6). Because the model in this section involves two choices that each exhibit differing marginal internalities, and because the planner cannot alter the private net benefit of evasion and labor income separately, that sufficiency property breaks down.\footnote{The latter point explains why the result of Kopczuk (2003) regarding the additivity principle of optimal taxation in the presence of externalities does not apply here. The planner cannot simply levy separate Pigouvian taxes on each of the internality-generating choices equivalent to their marginal social cost, and then conduct standard analysis ignoring the internalities.} Unless $\hat{\tau}^{b,y} = -\hat{\tau}^{b,e}$, empirically evaluating equation (32) requires decomposing the behavioral response to taxation.

This is just one example of a more general issue in optimal tax theory with behavioral biases that has received little attention. When the planner cannot directly alter the after-tax prices of all choices agents make, there are many circumstances in which the ETI is a sufficient statistic for the behavioral responses of agents under the standard model but where this is not true in a model with biases. In other words, not only do behavioral biases necessitate the estimation of behavioral wedges for solving optimal tax problems, they can also increase the number of behavioral response parameters that must be estimated to conduct applied welfare analysis.\footnote{It should be noted that this insufficiency result would also arise in a model in which labor income and evasion were associated with different externalities.}

Although this result is implicit in the optimal tax formulae presented in Farhi and Gabaix (2018), we think it is important to highlight the issue explicitly, given its implications for applied welfare analysis. In section 3.4,
we show that adopting a tax systems perspective may offer a partial resolution of the problem.

The Role of Non-Rate Policies

As with the model in section 2, it is important to emphasize the fact that both the behavioral wedges and the ETI appearing in equation 32 are endogenous to non-rate policy choices. The social planner can use both enforcement policies and bias alteration policies to change taxpayer biases. Thus, the planner can use these policy tools to directly change the taxpayer's marginal internalities as well as her behavioral response to taxation.

To be more explicit, let \( \hat{t}(p, \alpha) \) be the optimal tax rate for a given choice of \( p \) and \( \alpha \), as defined by equation (32). Apply the implicit function theorem, we have

\[
\frac{d\hat{t}}{dp} = -\frac{W_{tp}}{W_{tt}} \quad \text{and} \quad \frac{d\hat{t}}{d\alpha} = -\frac{W_{t\alpha}}{W_{tt}}.
\]

Suppose, for simplicity, that \( v_g \) is constant.\(^{45}\) As with the model in section 2, \( W_{tt} < 0 \) at a unique interior solution so the sign of \( \frac{d\hat{t}}{dp} \) is determined by the sign of

\[
W_{tp} = (v_g - 1) z_p + v_g t z_{tp} + \tau_{t}\ell_{tp} + \tau_{b,e} e_{tp} + \frac{d\tau_{b,e}}{dp} e_{tp} + \frac{d\tau_{b,\ell}}{dp} \ell_{tp} + \frac{d\tau_{b,\ell}}{dp} \ell_{tp}.
\]

As in section 2 there are three important factors to consider: the impact \( p \) has on the mechanical of effect of taxation (\( M \)); the way \( p \) changes the efficiency costs of taxation by changing the behavioral response to taxation (\( B \)); and, the way \( p \) changes the efficiency costs of taxation by directly changing the agent’s marginal internality (\( I \)). It is this final consideration that is unique to our setting because, although the presence of wedges in the behavioral response term \( B \) is similar to what would occur in a case with externalities, we would not generally expect tax system policies to directly alter the size of externality wedges.

What types of effects are being captured by \( I \)? Consider our example with dual misperception. Of course we might expect a change in \( p \) to change \( p^* \) unless our taxpayer is completely inattentive. But, notice we have allowed both \( t^* \) and \( p^* \) to depend on all three parameters of the tax system \( (t, p, \alpha) \). This allows for interesting rational inattention-style behavior on the part of our example taxpayer. For example, suppose that \( p \) is the evasion detection rate. As it increases and the returns to evasion decline, the cost of misperceiving the true tax rate also falls. If improved perception is cognitively costly, this could imply that \( \frac{\partial t^*}{\partial p} < 0 \).

The possibility of such behavior highlights an aspect of optimal tax systems analysis with biased agents that we could not highlight in the simple model of section 2. Full welfare maximization in the presence of

\(^{45}\)Appendix C.2 presents all the second derivatives of the social welfare function allowing for a non-constant marginal social value of public expenditure \((v_g)\).
behavioral biases requires not only taking into account the endogeneity of biases to bias alteration policies, which have no impact in the standard model. It also requires that the planner account for the impact that more traditional non-rate policies like \( p \) can have on taxpayer biases. Indeed, the analysis of \( p \) and \( \alpha \) in our model is quite similar. For both policies, their impact on optimal taxation can be decomposed into a mechanical component, a behavioral response component, and their impact on the marginal internality.

3.3 Optimal Enforcement and Bias Alteration

The planner’s first-order conditions with respect to \( p \) and \( \alpha \) imply that the optimal enforcement policy satisfies

\[
\left[ t + \left( \frac{y_p}{z_p} \right) \tilde{\tau}_{b,y} + \left( \frac{e_p}{z_p} \right) \tilde{\tau}_{b,e} \right] z_p = a_p + \frac{c_p}{v_p},
\]

and, similarly, the optimal bias alteration policy satisfies

\[
\left[ t + \left( \frac{y_{\alpha}}{z_{\alpha}} \right) \tilde{\tau}_{b,y} + \left( \frac{e_{\alpha}}{z_{\alpha}} \right) \tilde{\tau}_{b,e} \right] z_{\alpha} = a_{\alpha}.
\]

These optimality conditions are quite similar. In each case, the left-hand side represents the marginal benefit of increasing the policy parameter, which consists of the resulting change in tax revenue and in the taxpayer’s internality. As with the tax rate, a given change in taxable income produced by changing one of these parameters has different welfare consequences depending on the composition of the change. That is to say, for example, that $1000 in additional tax revenue produced by more stringent enforcement policy will not have the same internality costs as $1000 produced via some audit threat program unless the decomposition of the additional revenue into real and evasion components is the same in both cases.

The right-hand sides of equations (34) and (35) are the marginal cost of \( p \) and \( \alpha \), respectively. In both cases this includes the marginal administrative costs of the policy. In addition, the marginal cost of enforcement policy \( p \) includes some private marginal cost \( c_p \). Such costs might include additional time the taxpayer devotes to correctly completing tax forms or the change in the expected cost of being audited.\(^{46}\)

Once again, we can of course replace equations (34) and (35) with equivalent formulations in terms of elasticities:

\[
E_p^y = \frac{\alpha a_p + \alpha \left( \frac{c_p}{v_p} \right)}{t + \left( \frac{y_p}{z_p} \right) \tilde{\tau}_{b,y} + \left( \frac{e_p}{z_p} \right) \tilde{\tau}_{b,e}}
\]

\(^{46}\)See Keen and Slemrod (2017) for a more detailed discussion.
and

\[ E^2_\alpha = \frac{\alpha a_\alpha}{t + \left( \frac{\alpha a_\alpha}{t} \right) \tilde{\tau}^{b.g} + \left( \frac{\alpha a_\alpha}{t} \right) \tilde{\tau}^{b,e}}. \]  

(37)

In the equations above, \( E^2_\alpha \equiv \frac{\alpha}{z} \) is the enforcement elasticity of taxable income (Keen and Slemrod, 2017) and \( E^2_\alpha \) is the bias alteration elasticity (BAE) of taxable income, which we introduced in section 2.3. To reiterate, in contrast to the model of Keen and Slemrod (2017), the enforcement elasticity of taxable income is not a sufficient statistic for the welfare effects of increased enforcement policy in our model.

As in section 2, direct knowledge of the manner in which \( p \) and \( \alpha \) alter the taxpayer’s biases is not required to estimate the first-order welfare effects of changing these parameters. The response of labor supply and evasion to changes in \( p \) and \( \alpha \) provides all the required information about that matter. As in section 2, this is due to the fact that such behavioral responses are partially caused by such changes in biases. In appendix C.1, we present comparative statics results for our representative taxpayer which document this more precisely.

**Interaction Between Non-Rate Policies**

Just as changing \( p \) and \( \alpha \) alters the optimal tax rate, changing one non-rate policy instrument influences the optimal value of the other. Changes to the enforcement policy \( p \) could change the taxpayer’s marginal internalities directly, while also changing the taxpayer’s behavioral response to the bias alteration policy \( \alpha \). Consider the phenomenon of audit rate misperception. A rational inattention framework might lead us to expect that the taxpayer will be more attentive to the audit rate when it is larger, so we might expect increasing \( p \) to decrease the magnitude of the taxpayer’s marginal internalities as well as her behavioral response to bias alteration policy. On the other hand, suppose \( \alpha \) has the effect of increasing the taxpayer’s perceived audit rate \( \left( \frac{\partial \alpha}{\partial \alpha} \right) \). If the true audit rate is very low, the taxpayer may never have been audited or observed anyone who was audited, and it may prove difficult to convince them that the rate is not approximately zero. In this case, increasing \( p \) might actually increase the behavioral response to \( \alpha \). In either scenario, changes to one non-rate policy parameter will alter the optimal choice of the other.\(^{47}\)

\(^{47}\)As in our discussion of the impact of non-rate policies on the optimal tax rate in sections 2.2 and 3.2, we can formalize this discussion by applying the implicit function theorem to the planner’s first-order conditions. If \( \hat{\alpha} (t, p) \) is the optimal choice of bias alteration policy for a given tax rate and enforcement policy, then \( \frac{d\hat{\alpha}}{dp} = -\frac{W_{\alpha p}}{W_{\alpha \alpha}} \), so the sign of \( W_{\alpha p} \) determines the direction of the effect of enforcement policy on optimal bias alteration policy. As with the discussion of tax rates, enforcement policy has an impact here by changing both the taxpayer’s marginal internalities and her behavioral response. It also could change the marginal administrative costs of bias alteration. Analyzing the impact of \( \alpha \) on the optimal choice of \( p \) is similar, but \( \alpha \) might also influence the marginal cost of evasion \( (c_p) \) for the taxpayer. See appendix C.2 for the second derivatives of the social welfare function.
3.4 Applied Welfare Analysis

The results of this section have implications for a large body of existing empirical work on interventions designed to alter taxpayer beliefs about the enforcement environment: the audit threat literature. This body of work examines interventions in which a tax authority communicates information to taxpayers designed to alter their perception of the audit probability. This literature began with Slemrod et al. (2001), which presents an RCT examining such an intervention among a group of taxpayers in Minnesota. Many subsequent studies have followed suit, examining similar interventions in several countries, addressing both firms and individual taxpayers. Slemrod (ming) provides a recent review of the modern empirical tax compliance literature that includes an extensive discussion of this literature.

It is illustrative to consider our example in the special case where $t^s = t$, so that the only form of bias the taxpayer faces is a misperception of $p$ as $p^s$. Consider an audit threat intervention through the lens of our framework. Suppose that $p$ is the audit probability and that the mechanism through which the audit threat program might function is by changing the perceived audit rate. We might say that $\alpha$ represents the intensity of an audit threat program, perhaps capturing factors such as how taxpayers are contacted (through letters, by phone, or via in-person visits), how many attempts are made to contact the taxpayer, and similar program features that are costly to implement. For simplicity, we maintain all the assumptions that have accompanied our model throughout this section, so that $\alpha$ is a continuous measure of this intensity. Further, suppose for simplicity that the tax rate does not influence the perceived audit probability so that $p^s \equiv p^s(p, \alpha)$. The taxpayer’s behavioral wedges are thus $\tau^{b,t} \equiv (c^y_s - c_y) w$ and $\tau^{b,e} \equiv (c_e^e - c_e)$.

Holding $t$ and $p$ constant, what information would the tax authority need to determine whether they have currently set $\alpha$ to its optimal value? Equation (37) provides the answer to this question. On the left-hand side is the bias alteration elasticity ($E_\alpha^\gamma$), which can in principle be estimated by the tax authority via an appropriate RCT that examines how different choices of $\alpha$ impact reported taxable income. Audit threat studies typically provide estimates of this kind of information (i.e., on how taxpayers change their reported taxable income in response to audit threats). On the right-hand side, the numerator does not necessarily present any difficulties: presumably the tax authority has information about marginal administrative costs ($a_\alpha$) and has a good idea of what policy they’ve currently chosen ($\alpha$). The denominator has some familiar terms such as the tax rate ($t$) and the marginal social value of public expenditure ($v_g$). The former should be known to the tax authority and while the latter is clearly much more difficult to estimate, that is not a problem unique to our setting but one which must be addressed in many attempts to apply optimal tax theory to policy decisions. More importantly, the denominator also includes some new terms that are

48 Although not those in which public goods expenditure is held constant.
unique to models with behavioral biases: the behavioral wedge terms; the responsiveness of labor income to audit threats \((y_a)\); and separately the response of evasion to audit threats \((e_a)\).

Regarding the behavioral wedges, estimating these would require making some assumption about the functional form of the cost of evasion function \(c(e, wℓ; p)\) as well as knowledge of the current values of \(p\) and \(p^\star(p, α)\). Even setting aside the cost function, almost none of the existing work in the audit threat literature attempts to estimate taxpayer perceptions of the audit probability. Bérgolo et al. (2017) provides a noteworthy exception to this: in their audit threat experiment, they survey taxpayers in all treatment arms about their perceptions of the audit probability following the intervention. Note, however, that such survey responses do not necessarily map directly to \(p^\star\) in our model. For example, if \(p^\star\) partially reflects biased weighting of probabilities by agents when making decisions, even when probabilities are fully known, such survey results might be misleading.\(^{49}\)

Even if these behavioral wedges proved estimatable, because the tax authority will generally only observe taxable income, the decomposition of the response to taxation into a labor income response and an evasion response may be infeasible. Indeed, we are not aware of any existing attempts to do so in the audit threat literature.\(^{50}\)

**An Alternative Optimal Tax Rate Formula**

Given the concerns raised above, it is of practical interest to develop methods which render applied optimal tax analysis more tractable in a model with evasion. Here we discuss one possible approach. Suppose we can assume that \(\frac{c^e - c_y}{p^\star - p} \approx \frac{\partial c^e}{\partial p^\star}\) and \(\frac{c^e - c_y}{p^\star - p} \approx \frac{\partial c^e}{\partial p^\star}\). Note that these approximations are exact in the limit of small \(|p^\star - p|\) (i.e. as misperception grows small). If they are exact, we have the following alternative optimal tax formula

\[
t + \left(\frac{p^\star(p; α) - p}{v_g}\right) \frac{z_p^\star}{z_t} = \left(\frac{v_g - 1}{v_g}\right) \frac{1}{E\frac{1}{z_t}}.
\]

This presentation suggests another possible advantage of adopting a tax systems perspective in problems with behavioral biases: information about how taxable income responds to non-rate policy parameters can potentially render the optimal tax rate problem more manageable by eliminating the need to estimate each behavioral wedge. If \(z_p^\star\) and \(p^\star\) can be estimated, the remaining components of the formula are part of the conventional optimal linear income tax formula.

\(^{49}\)Such biased weighting is a key component of prospect theory.

\(^{50}\)More recently, a number of papers outside the audit threat literature have attempted to decompose the response of taxable income to tax policy parameters into real and evasion components. For example, see Bachas and Soto (2018) and Velayudhan (2018).
Although estimating $p^*$ may be difficult, this is just one example of the broader difficulty associated with estimating behavioral wedges, and is not unique to our setting. Furthermore, knowledge of $p^*$ is also required to estimate the behavioral wedges that enter the usual optimal tax formula in this example (equation 32), along with first derivatives of the taxpayer’s cost of evasion function $c_y$ and $c_e$. Estimating $z_{p^*}$, on the other hand, presents a distinct challenge: the plausibly estimatable quantities $z_p$ and $z_\alpha$ can provide only indirect evidence about $z_{p^*}$. Knowledge of $\frac{\partial p^*}{\partial p}$ or $\frac{\partial p^*}{\partial \alpha}$ would be required to estimate $z_{p^*}$ using an estimate of $z_p$ or $z_\alpha$.

As we noted above, Bérgolo et al. (2017) collected information on taxpayer perceptions of audit rates. Assuming taxpayer survey responses reflect their true perceived audit rates, this study collects some of the information one would need to evaluate equation (38). The experiment allows for the estimation of both the response of taxable income to the audit threat program ($z_\alpha$) and the average response of taxpayer perceptions of the audit rate to the program ($\frac{\partial p^*}{\partial \alpha}$). As noted above, these could be combined to estimate $z_{p^*}$. Together with their estimates of $p^*$ and some external information about the ETI and the marginal value of public expenditure ($v_g$), it is possible to evaluate the optimal tax formula above. Note, however, that it would be inappropriate to apply our framework directly to their setting.\textsuperscript{51}

Welfare Improvements Without Wedges

One might wonder if there is an analogue in this model to the intriguing result we presented in section 2.4, which showed that it was possible to identify welfare-improving tax system changes without knowledge of the taxpayer’s behavioral wedge. In appendix C.3, we show that a marginal tax system change $(dt, dp, d\alpha)$ where $d\alpha > 0$,

$$\frac{dt}{d\alpha} = \frac{\ell_p e_\alpha - \ell_\alpha e_p}{\ell_t e_p - \ell_p e_t},$$

(39)

and

$$\frac{dp}{d\alpha} = \frac{\ell_\alpha e_t - \ell_t e_\alpha}{\ell_t e_p - \ell_p e_t}$$

(40)

is welfare-improving if and only if:

$$(v_g - 1) z \left[ \frac{\ell_p e_\alpha - \ell_\alpha e_p}{\ell_t e_p - \ell_p e_t} \right] - (c_p + v_g a_p) \left[ \frac{\ell_\alpha e_t - \ell_t e_\alpha}{\ell_t e_p - \ell_p e_t} \right] - v_g a_\alpha > 0.$$ 

(41)

This makes it clear that the result we discuss in section 2.4 can, in principle, be extended to the case where agents make multiple choices.\textsuperscript{52} However, as equations (39) and (40) show, identifying a tax system change

\textsuperscript{51}There are several reasons for this. Most importantly, Bérgolo et al. (2017) study firms whereas our framework is designed with individual taxpayers in mind.

\textsuperscript{52}Extending this result to the case of agents making $n$ choices with $n$ policy parameters is straightforward.
that can be evaluated this way might be quite demanding empirically: it requires decomposing the response of taxable income to each policy parameter into a labor supply and evasion component.

Finally, we note that at the optimal tax system, the left-hand side of equation (72) would be equal to zero. Thus, as with the model in section 2.4, we can identify a necessary condition for the optimal tax system that does not include behavioral wedges.

4 Conclusion

Existing work on optimal tax theory in the presence of behavioral biases has generally treated biases as fixed features of the economy. However, as Farhi and Gabaix (2018) note, such biases are surely endogenous to broader features of the tax policy environment. Given the large body of existing empirical literature that evaluates interventions designed to tweak taxpayer perceptions of the tax system, as well as the increasing ubiquity of “nudge units” and related programs among tax authorities, it is important to develop a formal, normative framework for evaluating these policy choices. We propose extending the burgeoning literature on optimal tax systems to the consideration of policies that alter taxpayer biases as a natural way to address these issues.

We highlight several implications of this approach for optimal taxation with behavioral biases. First, we emphasize the fact that the behavioral wedges of Farhi and Gabaix (2018) should not be treated as exogenous. Policymakers can and should undertake actions to affect these biases where doing so may be beneficial and cost-effective. This point is analogous to the argument made by Slemrod and Kopczuk (2002): behavioral responses such as the ETI do not generally reflect deep structural parameters of the economy. Rather, they may be influenced by non-rate features of the policy environment including enforcement policies and the tax base definition. Policymakers should take this into account when determining how best to structure the tax system as a whole.

Our work in section 3 also highlights a barrier to empirical evaluation of the welfare costs of taxation with biased agents that has not appeared in the existing literature. When the planner cannot directly adjust the private net benefits of all choices an agent makes, as occurs in the presence of evasion opportunities, the number of behavioral responses that must be estimated can expand considerably relative to what is required under the standard model. Given that tax evasion is fairly pervasive, this result might pose a substantial barrier to applied welfare analysis in a setting with behavioral biases. In equation (38), we put forward a preliminary solution to this problem, based on the connection between the behavioral responses to tax rates and to non-rate tax policies. Future work should interrogate the robustness of this approach.
Finally, some of our results present the intriguing possibility that an optimal tax systems perspective may permit the identification of welfare-improving policy changes in a world with behavioral agents without requiring estimation of behavioral wedges. If this result extends to a broader set of modeling assumptions, it would be of substantial value. The challenge of estimating behavioral biases is the subject of much discussion in the existing behavioral public economics literature.\footnote{For example, it features heavily in Bernheim and Taubinsky’s (2018) recent review of behavioral public finance.} No proposed approach to this issue seems entirely satisfactory, and it is not at all clear that any approach ever could be given the ambiguity that necessarily accompanies welfare analysis in the absence of revealed preference methodology. Given this, it would be quite useful if some questions about optimal tax policy in the presence of such biases can be addressed while side-stepping such concerns.

This paper also develops many results that demonstrate how the consideration of behavioral biases can alter optimal tax systems analysis. The most obvious change is the inclusion of behavioral wedge terms in the equations characterizing the optimal tax system. These terms serve to incorporate the internality benefits/costs of changing tax system parameters into the determination of optimal policy. This builds on existing work that has incorporated internalities into welfare analyses of taxation, as in Farhi and Gabaix (2018), and optimal tax systems, as in Goldin (2015).

A final issue we highlight is one which has been noted elsewhere: behavioral biases themselves alter agents’ behavioral responses to changes in the tax rate and other policy parameters. In addition to considering how changes in non-rate features of the tax system impact the behavioral responses of taxpayers to tax rate changes, policymakers should consider how they might directly alter taxpayer biases. Not only does this suggest that a reconsideration of well-known tax systems problems—such as the joint choice of bases and rates—may be in order, it also highlights the importance of considering policy tools that are irrelevant in the standard model.

Future work should consider how well these results generalize to more complex and realistic models. Developing such extensions is important: interpreting existing and future empirical work on policies that plausibly alter taxpayer biases requires a richer description of taxpayer behavior than we have provided here. Important features include incorporating heterogeneous agents and examining the robustness of results to alternative normative assumptions, such as the welfare relevance of cognitive costs in endogenous bias models.\footnote{Appendix B presents a basic heterogeneous agents extension of the model from section 2.} Additional motivating examples should be investigated to help establish how prominent qualitative results from the literature on optimal taxation with biased agents and optimal tax systems might be adjusted in a framework that combines the two perspectives.
References


A Additional Results For Income Taxation Model

A.1 Chetty, Looney, and Kroft’s Theta

Readers familiar with the behavioral public finance literature might be interested in understanding what the \( \theta \) parameter introduced by Chetty et al. (2009) looks like in our model. We can define a version of CLK’s \( \theta \) for our labor supply model as

\[
\theta = \frac{E_1 - t}{E_1 - t} = -\frac{1}{1-t} \ell_t \quad \text{and} \quad E_\ell = \frac{w}{t} \ell_w.
\]

This parameter measures the relative responsiveness of the representative agent’s labor supply choices to changes in the net-of-tax rate and to changes in her wage rate. In the standard model, these are identical and, consequently, \( \theta = 1.55 \). Thus, if \( \theta \neq 1 \), this may provide some indication of taxpayer bias.

Once again applying the implicit function theorem to equation (2), we obtain

\[
\ell_w = \frac{1 - t - \frac{\partial z^h}{\partial \tau}}{\psi' + \frac{\partial z^h}{\partial \tau}}
\]

(42)

which, in combination with equation (7) gives us the following expression for CLK’s \( \theta \):

\[
\theta = \frac{1 + \left[ \frac{\partial z^h}{\partial \tau} / w \right]}{1 - \left[ \frac{\partial z^h}{\partial \tau} / (1-t) \right]}.
\]

(43)

Notice that in our model it is possible to have \( \theta = 1 \) even if agents are biased. This will occur whenever \((1-t) \frac{\partial z^h}{\partial \tau} = -w \frac{\partial z^h}{\partial \tau} \). The reason for this difference is that CLK assume that agents respond rationally to changes in pre-tax prices, whereas we allow for the possibility that the agent does not respond to changes in her wage rate as a rational agent would.

---

\(^{55}\) Though, as noted by Slemrod (2001), even in the standard model this need not be the case when tax evasion is introduced to the model.
A.2 Second Derivatives of Social Welfare

For reference, here we present the second derivatives of the social welfare function with respect to \((t, \alpha)\).

\[
W_{tt} = 2 (v_g - 1) w t_t + \left( w + \frac{\partial \tau^b}{\partial t} + \frac{\partial \tau^b}{\partial \ell} t_t \right) \ell_t + (v_g t w + \tau^b) \ell_{tt} + v_g \cdot (w t + tw t_t)^2 \tag{44}
\]

\[
W_{t\alpha} = (v_g - 1) w t_{\alpha} + \left( \frac{\partial \tau^b}{\partial \alpha} + \frac{\partial \tau^b}{\partial \ell} \ell_{t} \right) \ell_t + (v_g t w + \tau^b) \ell_{t\alpha} + v_g \cdot (w t + tw t_t) (tw t_{\alpha} - a_{\alpha}) \tag{45}
\]

\[
= W_{\alpha t}
\]

\[
W_{\alpha\alpha} = -v_g a_{\alpha\alpha} + \left( \frac{\partial \tau^b}{\partial \alpha} + \frac{\partial \tau^b}{\partial \ell} \ell_{\alpha} \right) \ell_{\alpha} + (v_g t w + \tau^b) \ell_{\alpha\alpha} + v_g \cdot (tw t_{\alpha} - a_{\alpha})^2 \tag{46}
\]

Interestingly, these derivatives can also be presented in a manner that excludes any direct reference to partial derivatives of \(\tau^b\) because

\[
w + \frac{\partial \tau^b}{\partial t} + \frac{\partial \tau^b}{\partial \ell} t_t = -\psi_{tt} t_t
\]

and

\[
\frac{\partial \tau^b}{\partial \alpha} + \frac{\partial \tau^b}{\partial \ell} \ell_{\alpha} = -\psi_{t\ell} t_{\alpha}.
\]

In principle, these equalities might provide a way to reduce the sufficient statistics required to estimate the second-order welfare effects of changes to the tax system, because they seem to offer a way to bypass the need to estimate \(\frac{\partial \tau^b}{\partial t}, \frac{\partial \tau^b}{\partial \alpha},\) or \(\frac{\partial \tau^b}{\partial \ell}\). However, knowledge of \(\psi_{tt}\) would be required to use this approach in practice, which presents its own difficulties.

A.3 Further Explanation of Welfare Improvements Without Behavioral Wedges

To better understand why equation (26) can be used to identify welfare-improving changes to the tax system without knowledge of behavioral wedges, consider the following presentation of the social welfare function:

\[
W \equiv u(\ell^R) + \underbrace{\left[ u(\ell) - u(\ell^R) \right]}_{\equiv \mathcal{I}(\ell, \ell^R)} + v (t z - a)
\]

where \(\ell \equiv \ell(t, \alpha)\) is the representative taxpayer’s choice of labor supply, \(\ell^R \equiv \ell^R(t)\) is the choice of labor supply the taxpayer would make if she were unbiased, and \(\mathcal{I}(\ell, \ell^R)\) is the taxpayer’s internality: the reduction in her private experienced utility due to her biases. For any marginal change to the tax system \((dt, d\alpha)\) we
have
\[
dW = dW_t dt + dW_\alpha d\alpha
\]
\[
= u' (\ell^R) [\ell_t^R dt + \ell_\alpha^R d\alpha] + u' (\ell) [\ell_t^R dt + \ell_\alpha^R d\alpha] - u' (\ell^R) [\ell_t^R dt + \ell_\alpha^R d\alpha] + v_g \cdot ([z + tz] dt + [tz_\alpha - a_\alpha] d\alpha).
\]

The term (i) is the change in the private experienced utility of an unbiased agent due to this change, (ii) is the change in the taxpayer’s internality, and (iii) is the change in social welfare due to the resulting change in public goods expenditure. Note that the envelope condition (which holds for unbiased agents) implies that \( u' (\ell^R) [\ell_t^R dt + \ell_\alpha^R d\alpha] = 0 \). Therefore (i) = 0 and (ii) = \( u' (\ell) [\ell_t^R dt + \ell_\alpha^R d\alpha] \), which reflects the marginal change in the taxpayer’s internality as a result of a marginal change to the tax system. This makes sense because \( u' (\ell) = (1 - t) w - \psi = \tau^b \) (see equation (2)).

This makes it clear why a marginal tax system change that leaves \( z \)—and therefore \( \ell \)—unchanged does not alter the taxpayer’s internality. The taxpayer’s internality has two components: \( u (\ell) \) and \( u (\ell^R) \). Because \( \ell \) does not change, \( u (\ell) \) does not change. Furthermore, even though this tax system change alters \( \ell^R \)—the labor supply of an unbiased taxpayer—the envelope theorem means that this has no first-order impact on the utility of the unbiased taxpayer. Together, these two facts imply that changes to the tax system that leave \( z \) constant have no first-order impact on the taxpayer’s internality. This is why equation (26) does not include behavioral wedges: knowledge of \( \tau^b \) is not required to determine the first-order welfare effect of the tax system change under consideration.

\section{B Heterogeneous Agents}

Suppose have a continuum of agents indexed \( h \in [0, 1] \). Agent \( h \)'s private experienced utility is
\[
u^h (\ell) \equiv (1 - t) w^h \ell - \psi^h (\ell)
\]
where \( t \in [0, 1] \) is a linear income tax rate, \( \ell \geq 0 \) is the agent’s labor supply, and \( w^h > 0 \) is her wage rate.

Each agent’s choice of labor supply \( \ell^h \) satisfies the condition
\[
\tau^{h,h} = (1 - t) w^h - \psi^h (\ell^h)
\]
where \( \tau^{h,h} \equiv \tau^{h,h} (\ell; w^h, t, \alpha) \) is agent \( i \)'s behavioral wedge. Let \( z^h \equiv w^h \ell^h \) be the corresponding taxable
income.

Let each agent’s $\tau^{b,h}$ and $\psi^h$ functions satisfy the same conditions we imposed for the case of the representative agent in section 2.1.

**Comparative Statics**

The responsiveness of each agent’s labor supply with respect to the policy parameters is summarized by

$$
\ell^h = -\frac{w^h + \frac{\partial \tau^{b,h}}{\partial t}}{\psi^h_{\ell\ell} + \frac{\partial \psi^h}{\partial t}} \quad \text{and} \quad \ell^h = -\frac{\frac{\partial \tau^{b,h}}{\partial \alpha}}{\psi^h_{\ell\ell} + \frac{\partial \psi^h}{\partial t}}. (49)
$$

**Planner’s Problem**

The social planner’s problem is to choose the tax system $(t, \alpha)$ that maximizes the welfare function

$$
W(t, \alpha) \equiv \int u^h(\ell^h) \, dh + v(g), \quad (50)
$$

where $g \equiv t \int z^h \, dh - a(\alpha)$ is public expenditure and $a(\alpha)$ is the administrative cost associated with some choice of the bias alteration policy $\alpha \geq 0$. Again, we impose the same conditions on the $v$ and $a$ functions as in section 2.1.

**Planner’s First-Order Conditions**

The social planner’s first-order conditions are

$$
W_t = -\int z^h \, dh + v_g \cdot \left( \int z^h \, dh + t \int z_t^h \, dh \right) + \int \tau^{b,h} \ell^h_t \, dh = 0 \quad (51)
$$

$$
W_\alpha = v_g \cdot \left( t \int z_\alpha^h \, dh - a_\alpha \right) + \int \tau^{b,h} \ell^h_\alpha \, dh = 0 \quad (52)
$$

or alternatively,

$$
(v_g - 1) \mathbb{E} \left[ z^h \right] + v_g t \mathbb{E} \left[ z_t^h \right] + \mathbb{E} \left[ \tau^{b,h} \right] \mathbb{E} \left[ \ell^h_t \right] + \text{Cov} \left( \tau^{b,h}, \ell^h_t \right) = 0 \quad (53)
$$

$$
v_g \cdot \left( t \mathbb{E} \left[ z_\alpha^h \right] - a_\alpha \right) + \mathbb{E} \left[ \tau^{b,h} \right] \mathbb{E} \left[ \ell^h_\alpha \right] + \text{Cov} \left( \tau^{b,h}, \ell^h_\alpha \right) = 0 \quad (54)
$$
Optimal Policy Formulae

Let \( \tilde{\tau}^{b,h} = \frac{\tau^{b,h}}{w^{b,h}} \) be agent \( h \)'s marginal internality of taxable income measured in terms the marginal value of public expenditure. If \( v_g > 1 \) at the optimal tax rate, then the rate satisfies:

\[
\frac{t + E[\tilde{\tau}^{b,h}] + Cov(\tilde{\tau}^{b,h}, \frac{z^h}{E[z^h]})}{1 - t} \left( \frac{v_g - 1}{v_g} \right) \frac{1}{E_{z}^{1-t}} = 0
\]

(55)

where \( E_{z}^{1-t} = -\frac{1-t}{E[z^h]} \) is the aggregate ETI.\(^{57}\) Thus, the manner in which the optimal rate differs from the representative agent case depends on the covariance term \( Cov(\tilde{\tau}^{b,h}, \frac{z^h}{E[z^h]}) \). This term measures how the response to taxation covaries with the marginal internality of taxable income across individuals. The formula has a quite intuitive interpretation: if the tax disproportionately decreases the taxable income of individuals with lower behavioral wedges, it should be raised to a higher level than what the average wedge \( (E[\tilde{\tau}^{b,h}]) \) would imply. If the opposite holds, it should be lower.

Similarly, the optimal bias alteration policy satisfies:\(^{58}\)

\[
\left[ t + E[\tilde{\tau}^{b,h}] + Cov(\tilde{\tau}^{b,h}, \frac{z^h}{E[z^h]}) \right] E[\frac{z^h}{E[z^h]}] = a_{\alpha}
\]

(56)

or equivalently

\[
E_{z}^{\alpha} = \frac{\alpha a_{\alpha}}{t + E[\tilde{\tau}^{b,h}] + Cov(\tilde{\tau}^{b,h}, \frac{z^h}{E[z^h]}) E[\frac{z^h}{E[z^h]}]}
\]

(57)

where \( E_{z}^{\alpha} = \frac{\alpha E[z^h]}{E[z^h]} \) is the aggregate bias alteration elasticity (BAE) of taxable income.\(^{59}\) Again, a fairly intuitive covariance term alters the optimal policy formula relative to the representative agent framework.

\(^{56}\)The result is derived from equation (53) via the following intermediate steps:

\[
\left( \frac{v_g - 1}{v_g} \right) E[z^h] + \left[ t + \frac{E[\tilde{\tau}^{b,h}]}{v_g} + Cov(\tilde{\tau}^{b,h}, \frac{z^h}{E[z^h]}) \right] E[\frac{z^h}{E[z^h]}] = 0
\]

\[
\left( \frac{v_g - 1}{v_g} \right) E_{z}^{1-t} = 0
\]

\(^{57}\)Note that this is not the average ETI, which is \( E[\frac{1-t}{E[z^h]}] \).

\(^{58}\)The result is derived from equation (54) via the following intermediate steps:

\[
-a_{\alpha} + \left[ t + E[\tilde{\tau}^{b,h}] + Cov(\tilde{\tau}^{b,h}, \frac{z^h}{E[z^h]}) \right] E[\frac{z^h}{E[z^h]}] = 0
\]

\[
-a_{\alpha} + \left[ t + E[\tilde{\tau}^{b,h}] + Cov(\tilde{\tau}^{b,h}, \frac{z^h}{E[z^h]}) \right] E[\frac{z^h}{E[z^h]}] E_{z}^{\alpha} = 0
\]

\(^{59}\)Note that this is not the average BAE, which is \( E[\frac{z^h}{E[z^h]}] \).
Insofar as the policy induces disproportionately large behavioral responses by individuals with higher behavioral wedges, the internality benefits (costs) of bias alteration policy will be larger (smaller) than those implied by the average wedge ($\mathbb{E} \left[ \tau^{b,h} \right]$) and aggregate BAE ($E_2^\alpha$).

**The Optimal Tax System without Administrative Costs**

To understand how heterogeneity alters the nature of the jointly optimal choice of tax rates and bias alteration, consider the case with costless bias alteration policy: $a(\alpha) = 0$, $\forall \alpha$. In such a case equation (56) implies that, at the optimum, either: $\mathbb{E} \left[ z^h_\alpha \right] = 0$, so bias alteration has no effect on tax revenue; or, $t + \mathbb{E} \left[ \tilde{z}^{b,h} \right] + \text{Cov} \left( \tilde{z}^{b,h}, \frac{z^h_\alpha}{\mathbb{E} [z^h_\alpha]} \right) = 0$. As with the optimal system discussed in section 2, the latter condition provides a description of how biases would ideally be altered, while the former condition reflects the fact that any given policy instrument may prove ineffective in achieving that ideal.

Let us consider the case where $t + \mathbb{E} \left[ \tilde{z}^{b,h} \right] + \text{Cov} \left( \tilde{z}^{b,h}, \frac{z^h_\alpha}{\mathbb{E} [z^h_\alpha]} \right) = 0$. It can be shown that the optimal tax system $(t^*, \alpha^*)$ satisfies

$$\mathbb{E} \left[ \frac{\tau^{b,h}}{u^h} \right] = -t^* v_g (g^*) - \text{Cov} \left( \frac{\tau^{b,h}}{u^h}, \frac{z^h_\alpha}{\mathbb{E} [z^h_\alpha]} \right),$$

(58)

$$g^* = t^* \int z^h dh \left[ v_g \right]^{-1} \left( 1 + \text{Cov} \left( \tau^{b,h}, \frac{z^h_\alpha}{\mathbb{E} [z^h_\alpha]} \right) - \text{Cov} \left( \tilde{z}^{b,h}, \frac{z^h_\alpha}{\mathbb{E} [z^h_\alpha]} \right) \right) \mathbb{E} \left[ z^h_\alpha \right]$$

(59)

Unless $\text{Cov} \left( \frac{z^h_\alpha}{u^h}, \frac{z^h_\alpha}{\mathbb{E} [z^h_\alpha]} \right) = \text{Cov} \left( \frac{z^h_\alpha}{u^h}, \frac{z^h_\alpha}{\mathbb{E} [z^h_\alpha]} \right)$, this result indicates that the first-best welfare outcome will not be achieved in the optimal tax system. Rather, the optimal level of public expenditure depends on the difference in the covariance terms in equation (59). Once again, it may be the case that $v_g < 1$ at the optimum. Assuming that $\mathbb{E} \left[ z^h_\alpha \right] < 0$, we have

$$v_g > 1 \iff \text{Cov} \left( \frac{\tau^{b,h}}{u^h}, \frac{z^h_\alpha}{\mathbb{E} [z^h_\alpha]} \right) > \text{Cov} \left( \frac{\tau^{b,h}}{u^h}, \frac{z^h_\alpha}{\mathbb{E} [z^h_\alpha]} \right).$$

This compliments the results we presented in section 2.4 by demonstrating that heterogeneity, like administrative costs of bias alteration, can reduce the social welfare attainable in the optimal tax system with bias alteration. This finding echoes earlier discussions about the implications of heterogeneity in the behavioral public finance literature, most prominently Taubinsky and Rees-Jones (2017), who note that the welfare cost of taxation is increasing in the variance of tax salience.
Identifying Welfare Improving Tax System Changes

Do our results about the possibility of identifying welfare improving tax system changes without needing to estimate behavioral wedges generalize to the heterogeneous agent context? Not in a straightforward fashion, but given some prior knowledge about the likely structure of biases and the impact of bias alteration policy, we can obtain some potentially useful analogous results.

Consider a marginal change to the tax system \((dt, d\alpha)\) such that \(\frac{dt}{d\alpha} = -\frac{E[z^h]}{E[z^t]}\) and \(d\alpha > 0\). Its impact on welfare is

\[
\begin{align*}
\text{d}W & = -W_t \frac{E[z^h]}{E[z^t]} + W_\alpha \\
& = \left[ -(v_g - 1)E[z^h] \frac{E[z^h]}{E[z^t]} - v_g a_\alpha \right] \\
& \quad + \left[ \text{Cov} \left( \tau_{b,h} w_h, z^h \right) - \text{Cov} \left( \tau_{b,h} w_h, z^t \right) \frac{E[z^h]}{E[z^t]} \right]
\end{align*}
\]

As our earlier results anticipate, the welfare impact of this change contains one term \((i)\) which is quite similar to the welfare impact of an analogous tax system change we discussed in the representative agent case (see section 2.4). Estimating the \((i)\) term does not require estimating behavioral wedges. However, accounting for heterogeneous agents has introduced a new term \((ii)\) which includes the covariances between agents’ behavioral responses to each policy and their behavioral wedges. This precludes a straightforward extension of our results about the possibility of identifying welfare improving tax system changes without the need to directly estimate features of agent biases. We note however, that if prior knowledge generates a clear prediction about the sign of \((ii)\), then the possibility of identifying welfare improving policy changes without knowledge of behavioral wedges remains.
C Additional Results for Tax Compliance Model

C.1 Comparative Statics

Our taxpayer’s FOCs can be represented as:

\[
\begin{bmatrix}
\tau^{b,\ell} - (1 - t) w + \psi_\ell + w c_y \\
\tau^{h,e} - t + c_e
\end{bmatrix}
= \begin{bmatrix} 0 \\ 0 \end{bmatrix}
\]

Let \( f_1 \equiv -(1 - t) w + \psi_\ell + w c_y \) and \( f_2 \equiv -t + c_e \). Further, let \( \kappa \in \{ t, \alpha, p \} \) be the policy parameter of interest. By the implicit function theorem, we have the following comparative statics results:

\[
\ell_\kappa = \lambda^{-1} \left[ -\left( \frac{\partial \tau^{b,\ell}}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} \right) \left( \frac{\partial \tau^{h,e}}{\partial e} + c_e \right) + \left( \frac{\partial \tau^{b,e}}{\partial \kappa} + \frac{\partial f_2}{\partial \kappa} \right) \left( \frac{\partial \tau^{h,e}}{\partial e} + \psi_\ell + w^2 c_y \right) \right]
\]

\[
e_\kappa = \lambda^{-1} \left[ \left( \frac{\partial \tau^{b,\ell}}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} \right) \left( \frac{\partial \tau^{h,e}}{\partial e} + w c_y \right) + \left( \frac{\partial \tau^{b,e}}{\partial \kappa} + \frac{\partial f_2}{\partial \kappa} \right) \left( \frac{\partial \tau^{h,e}}{\partial e} + \psi_\ell + w^2 c_y \right) \right]
\]

where \( \lambda = \left( \frac{\partial \tau^{b,\ell}}{\partial \kappa} + \psi_\ell \right) \left( \frac{\partial \tau^{h,e}}{\partial e} + w c_y \right) - \left( \frac{\partial \tau^{b,e}}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} \right) \left( \frac{\partial \tau^{h,e}}{\partial e} + \psi_\ell \right) \psi_\ell + w^2 c_y \). And because \( z_\kappa = w \ell_\kappa - e_\kappa \), we also have

\[
z_\kappa = \lambda^{-1} \left[ -\left( \frac{\partial \tau^{b,\ell}}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} \right) \left( \frac{\partial \tau^{b,e}}{\partial e} + w c_y \right) + \left( \frac{\partial \tau^{b,e}}{\partial \kappa} + \frac{\partial f_2}{\partial \kappa} \right) \left( \frac{\partial \tau^{h,e}}{\partial e} + \psi_\ell \right) \right] \ldots
\]

To obtain comparative statics for our parameters of interest, simply insert the relevant partial derivatives of \( f_1 \) and \( f_2 \):

\[
\frac{\partial f_1}{\partial t} = w; \quad \frac{\partial f_1}{\partial p} = w c_y p; \quad \frac{\partial f_1}{\partial \alpha} = 0
\]

\[
\frac{\partial f_2}{\partial t} = -1; \quad \frac{\partial f_2}{\partial p} = c_e p; \quad \frac{\partial f_2}{\partial \alpha} = 0
\]
Connecting Behavioral Responses

Here we derive a useful connection between the agent’s behavioral response to the tax rate \( t \) and their response to some non-rate policy instrument \( \kappa \in \{ p, \alpha \} \). It can be shown that:

\[
\left( \frac{\partial \tau_{b,\ell}}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} \right) \ell_t + \left( \frac{\partial \tau_{b,e}}{\partial \kappa} + \frac{\partial f_2}{\partial \kappa} \right) e_t = \left( w + \frac{\partial \tau_{b,\ell}}{\partial t} \right) \ell_{\kappa} + \left( -1 + \frac{\partial \tau_{b,e}}{\partial t} \right) e_{\kappa}.
\]

This implies that the responsiveness of taxable income to changes in \( \kappa \) is closely connected to the responsiveness of labor supply and evasion to changes in the tax rate \( t \):

\[
z_\kappa = \left( \frac{\partial \tau_{b,\ell}}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} \right) \ell_t + \left( \frac{\partial \tau_{b,e}}{\partial \kappa} + \frac{\partial f_2}{\partial \kappa} \right) e_t - \frac{\partial \tau_{b,\ell}}{\partial t} \ell_{\kappa} - \frac{\partial \tau_{b,e}}{\partial t} e_{\kappa}.
\]

Note that, if \( \frac{\partial \tau_{b,\ell}}{\partial t} = \frac{\partial \tau_{b,e}}{\partial t} = 0 \), then \( z_\kappa \) can be written as a combination of \( \ell_t \) and \( e_t \). We use this result in the main text to derive the alternative optimal tax formula presented in section 3.4 (equation 38).

---

\(^{61}\)To see this, let \( X = \begin{bmatrix} \frac{\partial \tau_{b,\ell}}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} + w c_{\ell,\ell} + \frac{\partial \tau_{b,e}}{\partial \kappa} + w c_{\ell,e} \\ \frac{\partial \tau_{b,\ell}}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} + w c_{e,\ell} + \frac{\partial \tau_{b,e}}{\partial \kappa} + w c_{e,e} \end{bmatrix} \), \( a = \begin{bmatrix} \frac{\partial \tau_{b,\ell}}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} + \frac{\partial \tau_{b,e}}{\partial \kappa} + \frac{\partial f_2}{\partial \kappa} \\ \frac{\partial \tau_{b,\ell}}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} + \frac{\partial f_2}{\partial \kappa} + \frac{\partial \tau_{b,e}}{\partial \kappa} + \frac{\partial f_2}{\partial \kappa} \end{bmatrix} \), and \( b = \begin{bmatrix} \frac{\partial \tau_{b,\ell}}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} + \frac{\partial \tau_{b,e}}{\partial \kappa} + \frac{\partial f_2}{\partial \kappa} \\ \frac{\partial \tau_{b,\ell}}{\partial \kappa} + \frac{\partial f_1}{\partial \kappa} + \frac{\partial f_2}{\partial \kappa} + \frac{\partial \tau_{b,e}}{\partial \kappa} + \frac{\partial f_2}{\partial \kappa} \end{bmatrix} \). Because

\( X \) is symmetric, \( a' X^{-1} b = b' X^{-1} a \). Finally, note that \( \begin{bmatrix} \ell_{\kappa} \\ e_{\kappa} \end{bmatrix} = X^{-1} a \) and \( \begin{bmatrix} \ell_t \\ e_t \end{bmatrix} = X^{-1} b \), so therefore \( a' \begin{bmatrix} \ell_t \\ e_t \end{bmatrix} = b' \begin{bmatrix} \ell_{\kappa} \\ e_{\kappa} \end{bmatrix} \).
C.2 Derivatives of Social Welfare

For reference, here we present various derivatives of the social welfare function with respect to \((t, p, \alpha)\). Let \(\frac{d_x h_x}{d\kappa} = \frac{\partial h_x}{\partial \kappa} + \frac{\partial h_x}{\partial \ell} \ell + \frac{\partial h_x}{\partial \kappa} \kappa\) for \(x \in \{\ell, e\}\) and \(\kappa \in \{t, p, \alpha\}\).

First derivatives:

\[
W_t = (v_g - 1)z + v_g tz_t + \tau_{b,\ell} \ell_t + \tau_{b,e} e_t
\tag{60}
\]

\[
W_p = -c_p v_g \cdot (tz_p - a_p) + \tau_{b,\ell} \ell_p + \tau_{b,e} e_p
\tag{61}
\]

\[
W_\alpha = v_g \cdot (tz_\alpha - a_\alpha) + \tau_{b,\ell} \ell_\alpha + \tau_{b,e} e_\alpha
\tag{62}
\]

Second derivatives (own):

\[
W_{tt} = (v_g - 1)z_t + v_g z_t + \frac{d\tau_{b,\ell}}{dt} \ell_t + \frac{d\tau_{b,e}}{dt} e_t + v_g tz_{tt} + \tau_{b,\ell} \ell_{tt} + \tau_{b,e} e_{tt} + v_{gg} \cdot (z + tz)^2
\tag{63}
\]

\[
W_{pp} = -c_{pp} - c_{pg} \omega_{lp} - c_{pe} e_p + v_g (tz_{pp} - a_{pp}) + \frac{d\tau_{b,\ell}}{dp} \ell_p + \frac{d\tau_{b,e}}{dp} e_p + \tau_{b,\ell} \ell_{pp} + \tau_{b,e} e_{pp} + v_{gg} (tz_p - a_p)^2
\tag{64}
\]

\[
W_{\alpha\alpha} = v_g (tz_{\alpha\alpha} - a_{\alpha\alpha}) + \frac{d\tau_{b,\ell}}{d\alpha} \ell_\alpha + \frac{d\tau_{b,e}}{d\alpha} e_\alpha + \tau_{b,\ell} \ell_{\alpha\alpha} + \tau_{b,e} e_{\alpha\alpha} + v_{gg} (tz_\alpha - a_\alpha)^2
\tag{65}
\]

Second derivatives (cross):

\[
W_{tp} = (v_g - 1)z_p + \frac{d\tau_{b,\ell}}{dp} \ell_t + \frac{d\tau_{b,e}}{dp} e_t + v_g tz_{tp} + \tau_{b,\ell} \ell_{tp} + \tau_{b,e} e_{tp} + v_{gg} \cdot (z + tz) (tz_p - a_p)
\tag{66}
\]

\[
= -c_{pg} \omega_{lp} - c_{pe} e_t + \frac{d\tau_{b,\ell}}{dp} \ell_t + \frac{d\tau_{b,e}}{dp} e_t + v_g tz_{tp} + \tau_{b,\ell} \ell_{tp} + \tau_{b,e} e_{tp} + v_{gg} \cdot (z + tz) (tz_p - a_p)
\]

\[
= W_{pt}
\]

\[
W_{ta} = (v_g - 1)z_\alpha + \frac{d\tau_{b,\ell}}{d\alpha} \ell_t + \frac{d\tau_{b,e}}{d\alpha} e_t + v_g tz_{ta} + \tau_{b,\ell} \ell_{ta} + \tau_{b,e} e_{ta} + v_{gg} \cdot (z + tz) (tz_\alpha - a_\alpha)
\tag{67}
\]

\[
= v_g z_\alpha + \frac{d\tau_{b,\ell}}{dt} \ell_\alpha + \frac{d\tau_{b,e}}{dt} e_\alpha + v_g tz_{ta} + \tau_{b,\ell} \ell_{ta} + \tau_{b,e} e_{ta} + v_{gg} \cdot (z + tz) (tz_\alpha - a_\alpha)
\]

\[
= W_{at}
\]
\[ W_{pa} = -c_{pp}w^{\ell}_{a} - c_{pe}e_{a} - v_{g}a_{pa} + \frac{d\tau_{b,\ell}}{d\alpha} \ell_{p} + \frac{d\tau_{b,c}}{d\alpha} e_{p} + v_{g}t_{zpa} + \tau_{b,\ell} \ell_{pa} + \tau_{b,c} e_{pa} \ldots \]  
\[ \cdots + v_{gg} \cdot (t_{zp} - a_{p}) (t_{za} - a_{a}) \]
\[ = - v_{g}a_{pa} + \frac{d\tau_{b,\ell}}{dp} \ell_{a} + \frac{d\tau_{b,c}}{dp} e_{a} + v_{g}t_{zpa} + \tau_{b,\ell} \ell_{pa} + \tau_{b,c} e_{pa} + v_{gg} \cdot (t_{zp} - a_{p}) (t_{za} - a_{a}) \]
\[ = W_{\alpha p} \]

**C.3 Welfare Improvements Without Wedges**

In the tax compliance model of section 3, the welfare change associated with a marginal change to the tax system \((dt, dp, da)\) is:

\[ dW = W_{t} dt + W_{p} dp + W_{a} da \]
\[ = (v_{g} - 1) zd t - c_{p} dp - v_{g} \cdot (a_{p} dp + a_{a} da) + v_{g} tdz + \tau_{b,\ell} dz + \tau_{b,\ell} dp + \tau_{b,c} de \]  
(69)

where \(dx = x_{t} dt + x_{p} dp + x_{a} da\) for any \(x \in \{z, \ell, e\}\).

Thus, if want to evaluate the welfare impact of some marginal tax system change \((dt, dp, da)\) without knowledge of \(\tau_{b,\ell}\) and \(\tau_{b,c}\), it is not sufficient for the tax system change to leave taxable income constant \((dz = 0)\), as it was in section 2.4. Instead, it must be the case that the tax system change leaves both labor supply and evasion unchanged \((df = de = 0)\). For a given \(da\), we can solve for the ratios of marginal changes \(\frac{dt}{da}\) and \(\frac{dp}{da}\) that satisfy this requirement using the following system of equations

\[ \ell_{t} \frac{dt}{da} + \ell_{p} \frac{dp}{da} + \ell_{a} = 0 \]
\[ e_{t} \frac{dt}{da} + e_{p} \frac{dp}{da} + e_{a} = 0 \]

The solution is

\[ \frac{dt}{da} = \frac{\ell_{p} e_{a} - \ell_{a} e_{p}}{\ell_{t} e_{p} - \ell_{p} e_{t}} \]  
(70)
\[ \frac{dp}{da} = \frac{\ell_{a} e_{t} - \ell_{t} e_{a}}{\ell_{t} e_{p} - \ell_{p} e_{t}} \]  
(71)

This implies that for some marginal tax system change \((dt, dp, da)\) where \(da > 0\) and equations 70 and 71 are satisfied, \(dW > 0\) if and only if:

\[ (v_{g} - 1) z \left[ \frac{\ell_{p} e_{a} - \ell_{a} e_{p}}{\ell_{t} e_{p} - \ell_{p} e_{t}} \right] - (c_{p} + v_{g} a_{p}) \left[ \frac{\ell_{a} e_{t} - \ell_{t} e_{a}}{\ell_{t} e_{p} - \ell_{p} e_{t}} \right] - v_{g} a_{a} > 0 \]  
(72)
We discuss this result in section 3.4.