Optimal Taxation with Behavioral Agents

Emmanuel Farhi and Xavier Gabaix*

May 22, 2018

Abstract

This paper develops a theory of optimal taxation with behavioral agents. We use a general behavioral framework that encompasses a wide range of behavioral biases such as misperceptions and internalities. We revisit the three pillars of optimal taxation: Ramsey (linear commodity taxation to raise revenues and redistribute), Pigou (linear commodity taxation to correct externalities) and Mirrlees (nonlinear income taxation). We show how the canonical optimal tax formulas are modified and lead to a rich set of novel economic insights. We also show how to incorporate nudges in the optimal taxation frameworks, and jointly characterize optimal taxes and nudges. (JEL: D03, H21).

1 Introduction

This paper develops a systematic theory of optimal taxation with behavioral agents. Our framework allows for a wide range of behavioral biases (for example, misperception of taxes or internalities), structures of demand, externalities, and population heterogeneity, as well as tax instruments. We derive a behavioral version of the three pillars of optimal taxation: Ramsey (1927) (linear commodity taxation to raise revenues and redistribute), Pigou (1920) (linear commodity taxation to correct for externalities), and Mirrlees (1971) (nonlinear income taxation).

Our results take the form of optimal tax formulas that generalize the canonical formulas derived by Diamond (1975), Sandmo (1975), and Saez (2001). Our formulas are expressed in terms of similar sufficient statistics and share a common structure.

The sufficient statistics can be decomposed into two classes: traditional and behavioral. Traditional sufficient statistics, which arise in non-behavioral models, include: social marginal utilities

*Affiliations: Harvard, NBER, and CEPR. Emails: efarhi@fas.harvard.edu, xgabaix@fas.harvard.edu. For excellent research assistance we thank D. Basak, J. Bloesch, V. Chau, A. Coppola, C. Wang, L. Wu and for helpful comments we thank the editor and referees, seminar participants at various institutions and H. Allcott, R. Chetty, P. Diamond, S. DellaVigna, A. Frankel, M. Gentzkow, E. Glaeser, O. Hart, E. Kamenica, L. Kaplow, W. Kopczuk, D. Laibson, B. Lockwood, U. Malmendier, C. Phelan, E. Saez, B. Salanié, J. Schwarzstein, A. Shleifer, T. Stralezcki, and D. Taubinsky. Gabaix thanks INET, the NSF (SES-1325181) and the Sloan Foundation for support.

of income and of public funds, compensated demand elasticities, marginal externalities, and equilibrium demands. Behavioral sufficient statistics are wedges that arise when agents do not fully optimize, and thus appear only in behavioral models. The behavioral tax formulas differ from their traditional counterparts not only because the behavioral sufficient statistics enter the tax formulas directly, but also because the presence of behavioral biases alters the values of traditional sufficient statistics.

We also propose a model of nudges as unconventional instruments that influence behavior without budgetary incidence. We show how to integrate nudges in canonical public finance models and derive optimal nudge formulas.

The value of our framework is three-fold. First, it unifies existing results in one single framework and identifies the key concepts that permeate many specialized behavioral public finance problems. Second, it allows to show how the forces arising in isolation interact. Third, it delivers concrete new insights on some of the cornerstone results of public economics (of course, these results require specific assumptions, which we make explicit as we derive them).

1. The Ramsey inverse elasticity rule states that optimal taxes to raise revenues are inversely proportional to the elasticity of demand. We show that when agents have limited attention to the tax, this principle is modified as follows: optimal taxes increase and scale with the inverse of the square of the attention where attention (between 0 and 1) is defined as the ratio of the perceived tax to the true tax.

2. The Pigou dollar for dollar principle requires that corrective taxes be set to the dollar value of the externality that they correct. When agents have limited attention, optimal taxes increase and must be set according to the dollar value of the externality divided by the attention to the tax.

3. When agents have heterogeneous attention, tax instruments become imperfect because they generate misallocation across agents: optimal Ramsey and Pigou taxes decrease with the variance of attention. Pigouvian taxes can no longer attain the first best and may be dominated by quantity restrictions, even though these blunter interventions prevent agents from expressing the intensity of their preferences. The principle of targeting no longer holds and it may be optimal to tax complements or subsidize substitutes of externality-generating goods.

4. Pigouvian taxes are not only attractive to correct for externalities but also internalities. However, to the extent that internalities are more prevalent among the poor, these taxes have adverse distributive consequences leading to a tradeoff between internality correction and redistribution. Nudges are an attractive intervention to circumvent this tradeoff and target internalities while avoiding reverse redistribution.

5. A fundamental result of the Mirrlees nonlinear income tax model is that optimal marginal tax rates are weakly positive. We show that if the poor do not fully recognize the future

benefits of work, perhaps because of myopia or hyperbolic discounting, then it is optimal to introduce negative marginal tax rates for low incomes. In addition, if the top marginal tax rate is particularly salient and contaminates perceptions of other marginal tax rates, then it should be lower than prescribed in the traditional analysis.

Finally, we hope that this paper will also motivate more empirically-minded economists to take on the task of measuring some of the key quantities identified by the theory, such as, for example: behavioral wedges; the variance of attention to taxes and the variance of nudgeability; the joint co-variance between attention to taxes, internalities and nudgeability; and the elasticity of attention to taxes.

Relation to the Literature We rely on recent progress in behavioral public finance and basic behavioral modelling. We build on earlier behavioral public finance theory.[1] Chetty (2009) and Chetty et al. (2009) analyze tax incidence and welfare with misperceiving agents; however, they do not analyze optimal taxation in this context. An emphasis of previous work is on the correction of "internalities," i.e. misoptimization because of self-control or limited foresight, which can lead to optimal "sin taxes" on cigarettes or fats (Gruber and Kőszegi (2001), O'Donoghue and Rabin (2006)).

Mullainathan et al. (2012) offer an overview of behavioral public finance. In particular, they derive optimality conditions for linear taxes, in a framework with a binary action and a single good. Baicker et al. (2015) further develop those ideas in the context of health care. Allcott et al. (2014) analyze optimal energy policy when consumers underestimate the cost of gas with two goods (e.g. cars and gas) and two linear tax instruments. The Ramsey and Pigou models in our paper generalize those two analyses by allowing for multiple goods with arbitrary patterns of own and cross elasticities and for multiple tax instruments. We derive a behavioral version of the Ramsey inverse elasticity rule.

Liebman and Zeckhauser (2004) study a Mirrlees framework when agent misperceive the marginal tax rate for the average tax rate. Two recent, independent papers by Gerritsen (2016) and Lockwood and Taubinsky (2017) study a Mirrlees problem in a decision vs. experienced utility model. Our behavioral Mirrlees framework is general enough to encompass, at a formal level, these models as well as many other relying on alternative behavioral biases.

We also take advantage of recent advances in behavioral modeling. We use a general framework that reflects previous analyses, including misperceptions and internalities. When modelling consumer demand with inattention to prices, we rely on part of the framework in Gabaix (2014), which builds on the burgeoning literature on inattention (Bordalo et al. (2013), Caplin and Dean (2015), Chetty et al. (2009), Gabaix (2017a), Gabaix and Laibson (2006), Khaw et al. (2017), Kőszegi and

[1] Numerous studies now document inattention to prices, e.g. Abaluck and Gruber (2011), Allcott and Taubinsky (2015), Allcott and Wozny (2014) (see also Busse et al. (2013)), Anagol and Kim (2012), Brown et al. (2010), Chetty (2015), DellaVigna (2009), and Ellison and Ellison (2009).

Szeidl (2013), Schwartzstein (2014), Sims (2003)). The behavioral agent in this framework misperceives prices while respecting the budget constraint in a way that gives a tractable behavioral version of basic objects of consumer theory, e.g. the Slutsky matrix and Roy's identity. Second, we also use the "decision utility" paradigm, in which the agent maximizes the wrong utility function. We unify those two strands in a general, agnostic framework that can be particularized to various situations. We also make some progress on the modelling of nudges.

The rest of the paper is organized as follows. Section 2 develops the general theory, with heterogeneous agents, arbitrary utility and decision functions. Section 3 shows a number of examples. We explain how they connect to the general theory, but we also make an effort to exposit them in a relatively self-contained manner. Section 4 studies the Mirrlees (1971) optimal nonlinear income tax problem. The online appendix contains more proofs and extensions.

For the readers who are mainly interested in applications, we have made an effort to ensure that the main applications in Section 3 are relatively self-contained and use a small amount of formalism. They also contain examples linking our theory to the existing empirical literature, and identify a number of challenges and opportunities for future measurement.

2 Optimal Linear Commodity Taxation

In this section, we introduce our general model of behavioral biases. We then describe how the basic results of price theory are modified in the presence of behavioral biases. Armed with these results, we then analyze the problem of optimal linear commodity taxation without externalities (Ramsey) where the objective of the government is to raise revenues and redistribute, and with externalities (Pigou) where an additional objective is to correct externalities. We also propose a model of nudges, show how to incorporate nudges in the optimal taxation framework, and characterize the joint optimal use of taxes and nudges. This analysis is performed at a general and rather abstract level. In the next section, we will derive a number of concrete results using simple examples, which are simple particularizations of the general model and results. The main proofs are in the appendix (Section 8).

2.1 Some Behavioral Price Theory

We start by describing a convenient "behavioral price theory" formalism to capture general behavioral biases using the central notion of "behavioral wedge". Our primitive is a demand function $\boldsymbol{c}(\boldsymbol{q}, w)$ where $\boldsymbol{q}$ is the price vector and w is the budget of the consumer. The demand function incorporates all the behavioral biases that the agent might be subject to (internalities, misperceptions, etc.). The only restriction that we impose on this demand function is that it exhausts the agent's budget so that $\boldsymbol{q} \cdot \boldsymbol{c}(\boldsymbol{q}, w) = w$. We evaluate the welfare of this agent according to a utility function $u(\boldsymbol{c})$, which represents the agent's true or "experienced" utility. The resulting indirect

utility function given by $v(\boldsymbol{q}, w) = u(\boldsymbol{c}(\boldsymbol{q}, w))$. Crucially, the demand function $\boldsymbol{c}(\boldsymbol{q}, w)$ is not assumed to result from the maximization of the utility function $u(\boldsymbol{c})$ subject to the budget constraint $\boldsymbol{q} \cdot \boldsymbol{c} = w$.

A central object is the "behavioral wedge", defined by:

$$\boldsymbol{\tau}^b(\boldsymbol{q}, w) = \boldsymbol{q} - \frac{u_{\boldsymbol{c}}(\boldsymbol{c}(\boldsymbol{q}, w))}{v_w(\boldsymbol{q}, w)}, \tag{1}$$

where b refers to a wedge due to $b$ehavioral biases. It is the difference between the price and marginal utility vectors (expressed in a money metric, as captured by $v_w(\boldsymbol{q}, w)$).[2] In the traditional model without behavioral biases, $\boldsymbol{\tau}^b(\boldsymbol{q}, w) = 0$. The wedge $\boldsymbol{\tau}^b(\boldsymbol{q}, w)$ encodes the welfare effects of a marginal reduction in the consumption of different goods, expressed in a money metric. We will see below how specific behavioral models lead to different values of the behavioral wedge.

This behavioral wedge plays a key role in a basic question that pervades this paper: how does an agent's welfare change when the price q_j of good j changes by dq_j? The answer is that it changes by $v_{q_j}(\boldsymbol{q}, w)\, dq_j$, where $v_{q_j}(\boldsymbol{q}, w)$ is given by the following behavioral version of Roy's identity:[3]

$$\frac{v_{q_j}(\boldsymbol{q}, w)}{v_w(\boldsymbol{q}, w)} = -c_j(\boldsymbol{q}, w) - \boldsymbol{\tau}^b(\boldsymbol{q}, w) \cdot \boldsymbol{S}_j^C(\boldsymbol{q}, w), \tag{2}$$

where $\boldsymbol{S}^C(\boldsymbol{q}, w)$ is the "income-compensated" Slutsky matrix, whose column j (corresponding to the consumption response to a compensated change in the price q_j) is defined as

$$\boldsymbol{S}_j^C(\boldsymbol{q}, w) = \boldsymbol{c}_{q_j}(\boldsymbol{q}, w) + \boldsymbol{c}_w(\boldsymbol{q}, w) c_j(\boldsymbol{q}, w).$$

The term $\boldsymbol{\tau}^b(\boldsymbol{q}, w) \cdot \boldsymbol{S}_j^C(\boldsymbol{q}, w)$ in equation (2) is a new term that arises with behavioral agents. The intuition is the following: a change dq_j in the price of good j changes welfare by $v_{q_j}(\boldsymbol{q}, w)\, dq_j = u_{\boldsymbol{c}}(\boldsymbol{c}(\boldsymbol{q}, w))\, \boldsymbol{c}_{q_j}(\boldsymbol{q}, w) dq_j$, a change which can be decomposed into an income effect $-u_{\boldsymbol{c}}(\boldsymbol{c}(\boldsymbol{q}, w))\, \boldsymbol{c}_w(\boldsymbol{q}, w) c_j(\boldsymbol{q}, w) dq_j = -v_w(\boldsymbol{q}, w)\, c_j(\boldsymbol{q}, w) dq_j$ and a substitution effect $u_{\boldsymbol{c}}(\boldsymbol{c}(\boldsymbol{q}, w)) \cdot \boldsymbol{S}_j^C(\boldsymbol{q}, w) dq_j$. In the traditional model with no behavioral biases, the income-compensated price change that underlies the substitution effect does not lead to any change in welfare—an application of the envelope theorem. The traditional version of Roy's identity follows. With behavioral biases, income-compensated price changes lead to changes in welfare—the envelope theorem no longer applies. The behavioral version of Roy's identity accounts for the associated welfare effects.

To gain some intuition, imagine that the agent is a net buyer of good j ($c_j > 0$) and that we are considering an increase dq_j in the price of good j. If the agent were rational, his welfare in monetary unit would be reduced by the usual term $-c_j dq_j < 0$. Now suppose that the agent is

[2] The behavioral wedge is independent of the particular cardinalization chosen for experienced utility (i.e., it is invariant to an increasing transformation $u \mapsto \phi \circ u$).

[3] We refer the reader to Section 7 in the appendix for the detailed derivations.

subject to behavioral biases such that the behavioral wedge is positive for good j ($\tau_j^b > 0$) and that behavioral wedges are negative for other goods ($\tau_i^b \leq 0$ for $i \neq j$). In addition, assume that good j is substitute with all the other goods ($S_{ij}^C \geq 0$ for $i \neq j$) and that the usual own-elasticity sign holds ($S_{jj}^C < 0$). In this case, the usual term $-c_j dq_j < 0$ overestimates the welfare loss for the agent because he was over-consuming good j to begin with.

As an example, consider the case of a smoker, calibrated using numbers from Gruber and Kőszegi (2004).[4] He smokes $c_j = 1$ pack a day. Suppose that the government increases the price of a pack of cigarettes by a dollar, causing the smoker to reduce his daily consumption of cigarettes by $-S_{jj}^C = 0.14$ packs. The traditional Roy's identity says that if the smoker is rational, his utility is reduced by a dollar a day. Now suppose that the smoker is behavioral and smokes "too much" because he does not take into account part of the health cost of smoking, with a corresponding internality of $\tau_j^b = 10.5$ dollars per pack. Then the behavioral Roy's identity says that his utility is improved by $-1 + 10.5 \times 0.14 = 0.47$ dollars a day. Taking into account that the agent is behavioral therefore flips the welfare effect of increasing the price of cigarettes because it helps the agent curb his excessive smoking.

We now present two useful concrete instantiations of the general formalism: decision vs. experienced utility and misperceptions.

Decision vs. Experienced Utility Model The demand function arises from the maximization of a "decision utility" $u^s(\boldsymbol{c})$ (the $s$ubjectively perceived utility), so that

$$\boldsymbol{c}(\boldsymbol{q}, w) = \arg\max_{\boldsymbol{c}} u^s(\boldsymbol{c}) \text{ s.t. } \boldsymbol{q} \cdot \boldsymbol{c} \leq w.$$

However, the true "experienced" utility remains $u(\boldsymbol{c})$ which can be different from $u^s(\boldsymbol{c})$. In this case, the behavioral wedge is simply given by the wedge between the decision and experienced marginal utilities

$$\boldsymbol{\tau}^b(\boldsymbol{q}, w) = \frac{u_c^s(\boldsymbol{c}(\boldsymbol{q}, w))}{v_w^s(\boldsymbol{q}, w)} - \frac{u_c(\boldsymbol{c}(\boldsymbol{q}, w))}{v_w(\boldsymbol{q}, w)}. \tag{3}$$

Intuitively, if a good entails a negative internality, then the agent over-consumes it at the margin, and the corresponding behavioral wedge is positive. The Slutsky matrix $\boldsymbol{S}^C(\boldsymbol{q}, w)$ is the Slutsky matrix of an agent with utility $u^s(\boldsymbol{c})$. Together with equation (3), it gives the expressions for behavioral

[4] Gruber and Kőszegi (2004) estimate that the total future health cost of a pack of cigarettes is $h = 35$ dollars and report a demand elasticity of below-median-income smokers of $\psi = 0.7$. If the smoker is a hyperbolic $\beta - \delta$ discounter with quasilinear utility, then he only internalizes a fraction $\beta = 0.7$ of these costs, and so, as we shall see shortly in the decision vs. experienced utility model, the internality for a pack of cigarettes is $\tau_j^b = (1 - \beta) h = 10.5$ dollars per pack. With a price $q_j = 5$ dollars per pack and a consumption of $c_j = 1$ pack a day, the Slutsky term encoding the sensitivity of the demand for cigarettes to its price is $S_{jj}^C = -\frac{\psi c_j}{q_j} = -0.14$ packs per dollar per day (here S_{jj}^C is the jth diagonal element of the Slutsky matrix $\boldsymbol{S}^C$). Hence, assuming that the behavioral wedges are zero for all goods but cigarettes ($\tau_i^b = 0$ for $i \neq j$), the behavioral term in the Roy's identity (2) is $-\boldsymbol{\tau}^b \cdot \boldsymbol{S}_j^C = -\tau_j^b S_{jj}^C = 1.47$ packs per day.

wedges and Slutsky matrices for the misperception particularization of the general model.[5]

To make things concrete, take a β-δ model à la Laibson (1997) with two periods and log utility. Then, $u(c_0, c_1) = \ln c_0 + \delta \ln c_1$ and $u^s(c_0, c_1) = \ln c_0 + \beta\delta \ln c_1$, where $\beta \in (0, 1)$ indicates excessive discounting of the future. We normalize the price of good 0 to be 1. The behavioral wedges are given by $\tau_0^b = \frac{\delta(1-\beta)}{1+\delta}$ and $\tau_1^b = -\frac{1-\beta}{\beta(1+r)(1+\delta)}$, where r is the interest rate.[6] The signs indicate that the agent over-consumes in the present, and under-consumes in the future. For example, if period 0 is work life and period 1 is retirement, then the agent does not save enough for retirement.

Misperception Model We turn to a model where the agent misperceives after-tax prices. There are two primitives: a utility function $u(\boldsymbol{c})$ and a perception function indicating the *s*ubjective price $\boldsymbol{q}^s(\boldsymbol{q}, w)$ perceived by the agent, as a function of the true price $\boldsymbol{q}$ and his income w.[7] Given true prices $\boldsymbol{q}$, perceived prices $\boldsymbol{q}^s$, and budget w, the demand $\boldsymbol{c}^s(\boldsymbol{q}, \boldsymbol{q}^s, w)$ is the consumption vector $\boldsymbol{c}$ satisfying $u_{\boldsymbol{c}}(\boldsymbol{c}) = \lambda^s \boldsymbol{q}^s$ for some $\lambda^s > 0$ such that $\boldsymbol{q} \cdot \boldsymbol{c} = w$.[8] Then the primitive demand function $\boldsymbol{c}(\boldsymbol{q}, w)$ of the general model is given by

$$\boldsymbol{c}(\boldsymbol{q}, w) = \boldsymbol{c}^s(\boldsymbol{q}, \boldsymbol{q}^s(\boldsymbol{q}, w), w).$$

With this formulation, the usual "trade-off" intuition applies in the space of perceived prices: marginal rates of substitution are equal to relative perceived prices $\frac{u_{c_i}}{u_{c_j}} = \frac{q_i^s}{q_j^s}$. The adjustment factor λ^s ensures that the budget constraint holds, despite the fact that agents misperceive prices.

The behavioral wedge is then given by the discrepancy between true prices and renormalized perceived prices:

$$\boldsymbol{\tau}^b(\boldsymbol{q}, w) = \boldsymbol{q} - \frac{\boldsymbol{q}^s(\boldsymbol{q}, w)}{\boldsymbol{q}^s(\boldsymbol{q}, w) \cdot \boldsymbol{c}_w(\boldsymbol{q}, w)}. \tag{4}$$

To derive the Slutsky matrix, we start by defining the Hicksian matrix of marginal perceptions $\boldsymbol{M}^H(\boldsymbol{q}, w)$, with elements $M_{ij}^H(\boldsymbol{q}, w) = \frac{\partial q_i^s(\boldsymbol{q}, w)}{\partial q_j} - \frac{\partial q_i^s(\boldsymbol{q}, w)}{\partial w} \frac{v_{q_j}}{v_w}$. Next, we define $\boldsymbol{S}^r(\boldsymbol{q}, w)$ to be the Slutsky matrix of a rational agent who faces prices $\boldsymbol{q}^s(\boldsymbol{q}, w)$ and achieves utility $v(\boldsymbol{q}, w)$: it simply records the derivatives of the expenditure function of the rational agent at that point.

The Slutsky matrix in the model with misperceptions is given by

$$\boldsymbol{S}^C(\boldsymbol{q}, w) = \left(\boldsymbol{I} - \boldsymbol{c}_w(\boldsymbol{q}, w) \left(\boldsymbol{\tau}^b(\boldsymbol{q}, w)\right)'\right) \boldsymbol{S}^r(\boldsymbol{q}, w) \boldsymbol{M}^H(\boldsymbol{q}, w). \tag{5}$$

In the rest of the paper, we will consider only the case where $\boldsymbol{q}_w^s = \boldsymbol{0}$, so that $\boldsymbol{M}^H = \boldsymbol{M}$, where

[5] For example, these expressions can be plugged back into equation (2).

[6] The derivation is in Section 11.1 in the online appendix.

[7] Our leading example will be as follows. There is a pre-tax price p_i, a tax τ_i so that the full price is $q_i = p_i + \tau_i$. However, the consumer perceives $q_i^s = p_i + m_i \tau_i$, where $m_i \in [0, 1]$ is the attention to the tax. See Sections 2.7-3.3 for applications of this setup.

[8] This is the formulation advocated for in Gabaix (2014), who discusses it extensively and uses it to derive a behavioral version of classical consumer and equilibrium theory. The problem has a solution under the usual Inada conditions. If there are several such λ^s, we take the lowest one, which is also the utility-maximizing one.

$\boldsymbol{M} = \boldsymbol{q}^s_{\boldsymbol{q}}$ is the matrix of marginal misperceptions. It shows how a change in the price q_j of good j creates a change $M_{kj}(\boldsymbol{q}, w) = \frac{\partial q^s_k(\boldsymbol{q},w)}{\partial q_j}$ in the perceived price q^s_k of a generic good k. The term $\boldsymbol{S}^r(\boldsymbol{q}, w)$ encodes how this change in the perceived price changes the demand for goods.[9] The term $\boldsymbol{c}_w(\boldsymbol{q}, w)\left(\boldsymbol{\tau}^b(\boldsymbol{q}, w)\right)'$ is a correction for wealth effects.

Equations (4) and (5) give the expressions for behavioral wedges and Slutsky matrices for the misperception particularization of the general model.[10] In Section 2.7, we work out in detail a tractable special case with quasilinear utility and a mixture of decision vs. experienced utility and misperceptions.

To make things concrete, consider a misperception model with two goods and quasilinear utility $u(c_0, c_1) = c_0 + U(c_1)$. Good 0 is the untaxed numéraire, the pre-tax price of good 1 is p_1, the post-tax price of good 1 is $q_1 = p_1 + \tau_1$ where τ_1 is the tax. The tax is not fully salient so that the perceived tax is $m_1\tau_1$, where $m_1 \in [0, 1]$ is the attention to the tax, and the perceived price is $q^s_1 = p_1 + m_1\tau_1$. In this case the behavioral wedges are $\tau^b_0 = 0$ and $\tau^b_1 = (1 - m_1)\tau_1$. If the tax is positive, then the agent over-consumes good 1. For example, if the tax of a retail item is not included in the posted price (as in Chetty et al. (2009)), then it is not fully salient, and consumers over-consume the item.

2.2 Optimal Taxation to Raise Revenues and Redistribute: Ramsey

There are H agents indexed by h. Each agent is competitive (price taker) as described in Section 2.1. All the functions describing the behavior and welfare of agents are allowed to depend on h. We assume perfectly elastic supply with fixed producer prices $\boldsymbol{p}$.[11]

The government sets a tax vector $\boldsymbol{\tau}$, so that the vector of after-tax prices is $\boldsymbol{q} = \boldsymbol{p} + \boldsymbol{\tau}$. Good 0 is constrained to be untaxed: $\tau_0 = 0$.[12] We introduce a social welfare function $W\left(v^1, ..., v^H\right)$ and a marginal value of public funds λ. We omit the dependence of all functions on $(\boldsymbol{q}, w)$, unless an ambiguity arises.

[9] There always exists a representation of the general model as a misperception model, but not as a decision vs. experienced utility model (see Lemma 12.1 in the online appendix). But the converse is not true, as a decision vs. experienced utility generates a symmetric Slutsky matrix.

[10] For example, these expressions can be plugged back into equation (2).

[11] The traditional justification for this assumption is the result, established by Diamond and Mirrlees (1971), that with a full set of commodity taxes, optimal tax formulas are independent of production elasticities. In the NBER working paper version of this paper we showed that this result extends to the environments with behavioral agents under the stronger assumption that there is a full set of completely salient taxes perceived as prices (potentially in addition to other taxes). We also generalized our optimal tax formulas to the case where these assumptions are not verified.

[12] Think about leisure for instance, which cannot be taxed. This assumption rules out the replication of lump-sum taxes via uniform ad valorem taxes on all goods. Indeed, just like lump sum taxes, uniform ad valorem taxes raise revenues without introducing distortions since they do not change relative prices.

The planning problem is[13]

$$\max_{\boldsymbol{\tau}} L(\boldsymbol{\tau}),$$

where

$$L(\boldsymbol{\tau}) = W\left(\left(v^h\left(\boldsymbol{p}+\boldsymbol{\tau}, w^h\right)\right)_{h=1,\ldots,H}\right) + \lambda \sum_h \boldsymbol{\tau}\cdot \boldsymbol{c}^h\left(\boldsymbol{p}+\boldsymbol{\tau}, w^h\right), \tag{6}$$

and $w^h = \boldsymbol{p}\cdot\boldsymbol{e}^h$ is the value of the initial endowment $\boldsymbol{e}^h$ of agent h.

Following Diamond (1975), for every agent h we define $\beta^h = W_{v^h} v^h_w$ to be the social marginal welfare weight and $\gamma^h = \beta^h + \lambda\boldsymbol{\tau}\cdot\boldsymbol{c}^h_w$ to be the social marginal utility of income. The difference $\lambda\boldsymbol{\tau}\cdot\boldsymbol{c}^h_w$ between β^h and γ^h captures the marginal impact on tax revenues of a marginal increase in the income of agent h. We also renormalize the behavioral wedge to take into account the welfare weight attached to each agent. Letting $\boldsymbol{\tau}^{b,h}$ be the behavioral wedge for agent h, we define

$$\widetilde{\boldsymbol{\tau}}^{b,h} = \frac{\beta^h}{\lambda}\boldsymbol{\tau}^{b,h}. \tag{7}$$

We now characterize the optimal tax system.[14]

Proposition 2.1 (Behavioral many-person Ramsey formula) *If commodity i can be taxed, then at an interior optimum*

$$\frac{\partial L(\boldsymbol{\tau})}{\partial \tau_i} = \sum_h \left[\left(\lambda-\gamma^h\right)c^h_i + \lambda\left(\boldsymbol{\tau}-\widetilde{\boldsymbol{\tau}}^{b,h}\right)\cdot \boldsymbol{S}^{C,h}_i\right] = 0. \tag{8}$$

An intuition for this formula can be given along the following lines. The impact of a marginal increase in $d\tau_i$ on social welfare is the sum of three effects: a mechanical effect, a substitution effect, and a misoptimization effect.

Let us start with the mechanical effect $\sum_h(\lambda-\gamma^h)c^h_i d\tau_i$. If there were no changes in behavior besides income effects, then the government would reduce the utility of agent h by $\beta^h c^h_i d\tau_i$ and collect additional revenues $(1-\boldsymbol{\tau}\cdot\boldsymbol{c}^h_w)c^h_i d\tau_i$ valued at $\lambda(1-\boldsymbol{\tau}\cdot\boldsymbol{c}^h_w)c^h_i d\tau_i$, leading to a total effect on the government objective of $(\lambda-\gamma^h)c^h_i d\tau_i$.

Let us turn to the substitution effect $\sum_h \lambda\boldsymbol{\tau}\cdot\boldsymbol{S}^{C,h}_i d\tau_i$. The change in consumer prices resulting from the tax change $d\tau_i$ induces a change in behavior $\boldsymbol{S}^{C,h}_i d\tau_i$ of agent h over and above the income effect accounted for in the mechanical effect. The resulting change $\boldsymbol{\tau}\cdot\boldsymbol{S}^{C,h}_i d\tau_i$ in tax revenues is a fiscal externality which is valued by the government as $\lambda\boldsymbol{\tau}\cdot\boldsymbol{S}^{C,h}_i d\tau_i$.

Finally, let us analyze the misoptimization effect $-\sum_h \lambda\widetilde{\boldsymbol{\tau}}^{b,h}\cdot\boldsymbol{S}^{C,h}_i d\tau_i$. Noting that $-\lambda\widetilde{\boldsymbol{\tau}}^{b,h}\cdot$

[13] If the government needs to raise a given amount of revenues from taxes, then λ is endogenous and equal to the Lagrange multiplier on the government budget constraint.

[14] Suppose that there is uncertainty, possibly heterogeneous beliefs, several dates for consumption, and complete markets. Then, our formula (8) applies without modifications, interpreting goods as a state-and-date contingent goods. See Spinnewijn (2015) for an analysis of unemployment insurance when agents misperceive the probability of finding a job, and Dávila (2017) for an analysis of a Tobin tax in financial markets with heterogeneous beliefs.

$\boldsymbol{S}_i^{C,h} d\tau_i = -\beta^h \boldsymbol{\tau}^{b,h} \cdot \boldsymbol{S}_i^{C,h} d\tau_i$, this effect can be understood as a manifestation of the failure of the envelope theorem encoded in the behavioral version of Roy's identity in equation (2). Basically, if the agent over-consumes a good i, then, everything else equal, taxing good i is more attractive at the margin.

All in all, adding behavioral agents introduces the following differences. First, it modifies behavioral responses which endogenously changes the values of β^h, γ^h, and $\boldsymbol{S}_i^{C,h}$. Second, it leads to the new term $-\lambda \widetilde{\boldsymbol{\tau}}^{b,h} \cdot \boldsymbol{S}_i^{C,h}$.

One way to think about the optimal tax formula (8) is as a system of equations indexed by i in the optimal taxes τ_j for the different commodities:

$$\frac{-\sum_{j,h} S_{ji}^{C,h} \tau_j}{c_i} = 1 - \frac{\bar{\gamma}}{\lambda} - cov\left(\frac{\gamma^h}{\lambda}, \frac{H c_i^h}{c_i}\right) - \frac{\sum_{j,h} \widetilde{\tau}_j^{b,h} S_{ji}^{C,h}}{c_i}, \tag{9}$$

where $c_i = \sum_h c_i^h$ is total consumption of good i and $\bar{\gamma} = \frac{1}{H}\sum_h \gamma^h$ is the average social marginal utility of income. The term on the left-hand-side encodes the extent to which the consumption of good i is discouraged by the overall tax system. The first right-hand-side term $1 - \frac{\bar{\gamma}}{\lambda} - cov(\frac{\gamma^h}{\lambda}, \frac{Hc_i^h}{c_i})$ captures the revenue raising and redistributive objectives of taxation: at the optimum, good i is more discouraged if the need for government revenues is large and if agents with low social marginal utility of income consume relatively more of good i. The second right-hand-side term $\frac{-\sum_{j,h} \tilde{\tau}_j^{b,h} S_{ji}^{C,h}}{c_i}$ captures the corrective objective of taxation: at the optimum, good i is more discouraged if good i and complements to i have large behavioral wedges.[15]

We can view this system of equations as a linear system of equations in the τ_j's indexed by i with *endogenous* coefficients given by $\sum_{j,h} S_{ji}^{C,h}/c_i$ and *endogenous* forcing terms given by the right-hand-side of (9). Solving this system allows us to express the taxes as functions of these endogenous objects which we refer to as *sufficient statistics* since they mediate the dependence of optimal taxes on primitives of the model and of observables at the optimum. Since these observables in turn depend on taxes, one can view this mapping as a nonlinear fixed-point equation.

To be clear, the sufficient statistics must be computed at the optimum. In certain parametric models, these objects will be constant leading to a closed-form solution for taxes. Indeed this will be the case for many of the examples explored in Section 3, which require specific functional forms (e.g. isoelastic, quasi-linear), in which elasticities or key derivatives are independent of the tax. In general however, these sufficient statistics are not constant. It would then be incorrect to use estimates obtained away from the optimum to infer optimal taxes. Instead, they can be used to test for optimality of an observed tax system, and in case of sub-optimality, to determine the direction of welfare-improving marginal tax reforms. Alternatively, meta-analyses of empirical estimates of

[15] Suppose that in addition to linear commodity taxes, the government can use a lump-sum tax or rebate, identical for all agents (a "negative income tax"). Then optimal commodity taxes are characterized by the exact same conditions. But there is now an additional optimality condition corresponding to the optimal choice of the lump-sum rebate yielding $\bar{\gamma} = \lambda$.

these sufficient statistics can be used to determine a plausible range for optimal taxes.

The generality of the formula is useful for several reasons. First, it identifies the basic objects that matter for optimal taxes in different contexts. Second, it unifies an otherwise disparate set of insights. Third, it allows to identify tractable special cases while at the same time clarifying the forces that are being eliminated to get tractability.

2.3 Optimal Taxation with Externalities: Pigou

We now introduce externalities and study the consequences for the optimal design of commodity taxes with behavioral agents. The utility of agent h is now $u^h\left(\boldsymbol{c}^h, \xi\right)$, where $\xi = \xi\left(\left(\boldsymbol{c}^h\right)_{h=1,\ldots,H}\right)$ is a one-dimensional externality (for simplicity) that depends on the consumption vectors of all agents and is therefore endogenous to the tax system.[16] All individual functions encoding the behavior and welfare of agents now depend on the externality ξ.

The planning problem becomes $\max_{\boldsymbol{\tau}} L\left(\boldsymbol{\tau}\right)$, where

$$L\left(\boldsymbol{\tau}\right) = W\left(\left(v^h\left(\boldsymbol{p}+\boldsymbol{\tau}, w^h, \xi\right)\right)_{h=1,\ldots,H}\right) + \lambda \sum_h \boldsymbol{\tau}\cdot\boldsymbol{c}^h\left(\boldsymbol{p}+\boldsymbol{\tau}, w^h, \xi\right)$$

and $\xi = \xi\left(\left(\boldsymbol{c}^h\left(\boldsymbol{p}+\boldsymbol{\tau}, w^h, \xi\right)\right)_{h=1,\ldots,H}\right)$. We call $\Xi = \frac{\sum_h\left[\beta^h \frac{v_\xi^h}{v_w^h} + \lambda\boldsymbol{\tau}\cdot\boldsymbol{c}_\xi^h\right]}{1-\sum_h \xi_{\boldsymbol{c}^h}\cdot\boldsymbol{c}_\xi^h}$ the social marginal value of the externality. This concept includes all the indirect effects of the externality on consumption and the associated effects on tax revenues (the term $\sum_h \lambda\boldsymbol{\tau}\cdot\boldsymbol{c}_\xi^h$ in the numerator) as well as the associated multiple round effects on the externality (the "multiplier" term encapsulated in the denominator). With this convention, Ξ is negative for a bad externality, like pollution. We also define the (agent-specific) Pigouvian wedge

$$\boldsymbol{\tau}^{\xi,h} = -\frac{\Xi\xi_{\boldsymbol{c}^h}}{\lambda}.$$

It represents the social dollar value of the externality created by one more unit of consumption by agent h. We finally define the externality-augmented social marginal utility of income $\gamma^{\xi,h} = \gamma^h + \Xi\xi_{\boldsymbol{c}^h}\cdot\boldsymbol{c}_w^h = \beta^h + \lambda\left(\boldsymbol{\tau}-\boldsymbol{\tau}^{\xi,h}\right)\cdot\boldsymbol{c}_w^h$.[17] The next proposition generalizes Proposition 2.1.

Proposition 2.2 (Behavioral many-person Pigou formula) *If commodity i can be taxed, then at an*

[16] For example, to capture an externality (e.g. second hand smoke) from the consumption of good 1, we could specify $\xi = \frac{\xi_*}{H}\sum_h c_1^h$ and $u^h\left(\boldsymbol{c}^h, \xi\right) = u^h\left(\boldsymbol{c}^h\right) - \xi$.

[17] This definition captures the fact that, as one dollar is given to the agent, his direct social utility increases by γ^h, but the extra dollar changes consumption by $\boldsymbol{c}_w^h$, and, hence, the total externality by $\xi_{\boldsymbol{c}^h}\cdot\boldsymbol{c}_w^h$, with a welfare impact $\Xi\xi_{\boldsymbol{c}^h}\cdot\boldsymbol{c}_w^h$.

interior optimum

$$\frac{\partial L(\boldsymbol{\tau})}{\partial \tau_i} = \sum_h \left[\left(\lambda - \gamma^{\xi,h} \right) c_i^h + \lambda \left(\boldsymbol{\tau} - \widetilde{\boldsymbol{\tau}}^{b,h} - \boldsymbol{\tau}^{\xi,h} \right) \cdot \boldsymbol{S}_i^{C,h} \right] = 0. \tag{10}$$

Formally, misoptimization and externality wedges ($\widetilde{\boldsymbol{\tau}}^{b,h}$, $\boldsymbol{\tau}^{\xi,h}$) enter symmetrically in the optimal tax formula. We will see later that in some particular cases, behavioral biases can be alternatively modeled as externalities. But this is not true in general. For example, misperceptions of prices typically give rise to non-symmetric Slutsky matrices $\boldsymbol{S}^{C,h}$ which cannot be captured with a traditional externality model. Moreover, even with a quasilinear utility function and separable utility (so that the Slutsky matrix is diagonal and hence symmetric), the misperception model would require externalities that directly depend on price wedge $\boldsymbol{q} - \boldsymbol{q}^s$, which is not covered in the traditional externalities literature.

2.4 Optimal Nudges

We turn our attention to another type of instrument with no counterpart in the traditional theory: nudges (Thaler and Sunstein (2008)). The concept of nudge captures many different forms of interventions ranging from shocking pictures (for example the picture of a cancerous lung on a pack of cigarettes), to default options (for example in 401(k) retirement savings accounts). There is no agreed-upon model to capture these interventions. The goal of this section is to make an attempt at proposing a general formalism that captures some of the common elements of these different nudges, and a specific specialization of this general model which we think is useful to capture the psychology of nudges.

At an abstract level, we assume that a nudge influences consumption but does not enter the budget constraint—this is the key difference between a nudge and a tax. The demand function $\boldsymbol{c}^h(\boldsymbol{q}, w, \chi)$ satisfies the budget constraint $\boldsymbol{q} \cdot \boldsymbol{c}^h(\boldsymbol{q}, w, \chi) = w$, where χ is the nudge vector. In general, a nudge may also affect the agents' utility $u^h(\boldsymbol{c}, \chi)$.[18]

For example, in the decision vs. experienced utility model, we capture the dependence of behavior on nudges by allowing them to influence the decision utility $u^{s,h}(\boldsymbol{c}, \chi)$ of the agents. Similarly, in the misperception model, they modify the perceived prices $\boldsymbol{q}^s(\boldsymbol{q}, w, \chi)$. As we have already noted in footnote 9, it is always possible to represent the behavior arising from a decision vs. experienced utility model (and in fact any behavioral model considered in this paper) via a misperception model, and the converse is not true. Hence, for a given context, there are two possibilities for the dependence of behavior with nudges: (i) it can be modeled in both ways; (ii) it can only be modeled with misperceptions.[19]

[18] Glaeser (2006) and Loewenstein and O'Donoghue (2006) discuss the idea that nudges have a psychic cost.

[19] For example, if the Slutsky matrices encoding elasticities of substitution in the presence of nudges are not symmetric, then nudges cannot be captured using the decision vs. utility framework. In contrast, a misperception model can generate asymmetric Slutsky matrices.

In case (i) two reasonable attitudes are possible. First, one can take the view what really matters is to represent behavior accurately, and it is not intrinsically problematic if behavior can be rationalized with two different but isomorphic models. Second, one can argue that one model better captures the underlying psychological mechanisms than the other. In case (ii), this dilemma does not arise and one is forced to adopt the more general model.

Let us now sketch concrete models. We propose the following model of a "nudge as a psychological tax", which is one useful specialization of the general formalism. We assume that in the absence of a nudge, the agent has decision utility $u^{s,h}$ and perceived price $\boldsymbol{q}^{s,h,*}$. We imagine that a nudge χ applied to good i changes the perceived price of good to $q_j^{s,h} = q_j^{s,h,*} + \chi\eta^h$ if $j = i$ and leaves $q_j^{s,h} = q_j^{s,h,*}$ otherwise unchanged, where $\eta^h \geq 0$ captures the nudgeability of the agent so that $\eta^h = 0$ corresponds to a non-nudgeable agent. Hence, $\boldsymbol{c}$ satisfies $u_c^{s,h}(\boldsymbol{c}) = \Lambda^h \boldsymbol{q}^{s,h}$ for some Λ^h such that $\boldsymbol{q} \cdot \boldsymbol{c} = w$. A straightforward example of such nudge is a public campaign against cigarettes ($\chi > 0$) or for recycling ($\chi < 0$).[20]

Let us next lay down a specific model of "nudge as a change in perceived utility". The nudge changes the perceived utility to $u_c^{s,h}(\boldsymbol{c}, \chi) = u_c^{s,h,*}(\boldsymbol{c}) - \frac{\chi\eta^h}{\Lambda^h} c_i$. This captures that the nudge reduces the perceived marginal utility of good i.[21]

When utility is quasilinear the two concrete formulations above are isomorphic, with $\bar{\Lambda}^h = \Lambda^h = 1$. Indeed, they lead to the same first order conditions and share the same budget constraint. When utility is not quasilinear, this isomorphism is only valid up to the first order in nudges around a zero nudge benchmark. In both formulations, the extent to which these nudges are intrinsically aversive can be captured with an aversiveness parameter ι^h and an experienced utility of the form $u^h(\boldsymbol{c}, \chi) = u^h(\boldsymbol{c}) - \iota^h \chi c_i$.

The planning problem is $\max_{\boldsymbol{\tau}, \chi} L(\boldsymbol{\tau}, \chi)$, where

$$L(\boldsymbol{\tau}, \chi) = W\left(\left(v^h\left(\boldsymbol{p} + \boldsymbol{\tau}, w^h, \xi, \chi\right)\right)_{h=1,\ldots,H}\right) + \lambda \sum_h \boldsymbol{\tau} \cdot \boldsymbol{c}^h\left(\boldsymbol{p} + \boldsymbol{\tau}, w^h, \xi, \chi\right),$$

with $v^h\left(\boldsymbol{p} + \boldsymbol{\tau}, w^h, \xi, \chi\right) = u^h\left(\boldsymbol{c}^h\left(\boldsymbol{p} + \boldsymbol{\tau}, w^h, \xi, \chi\right), \xi, \chi\right)$.

We now characterize the formula for optimal nudges in the general case.

[20] More generally, one could think of examples of nudges that alter the perceived budget constraint in a nonlinear fashion, so that the agent perceives the budget set $B^{s,h}(\boldsymbol{q}, \boldsymbol{c}, \chi) \leq w$, so that his consumption $\boldsymbol{c}$ satisfies $u_c^{s,h}(\boldsymbol{c}) = \Lambda^h B_{\boldsymbol{c}}^{s,h}(\boldsymbol{q}, \boldsymbol{c}, \chi)$ for some $\Lambda^h > 0$ such that the true budget constraint $\boldsymbol{q} \cdot \boldsymbol{c} = w$ holds. In some cases, it might even make sense to consider non-differentiable perceived budget sets $B^{s,h}(\boldsymbol{q}, \boldsymbol{c}, \chi) = \boldsymbol{q}^{s,h,*} \cdot \boldsymbol{c} + \eta^h |c_i - \chi|$ to capture, for example, default options in retirement plans (see e.g. Carroll et al. (2009)), so that the agent experiences an extra psychological penalty if he deviates from the default quantity χ recommended by the nudge. In such a case, one would expect, in an heterogeneous population, to observe a discrete mass of agents bunched at the default.

[21] Alternatively, we could assume that the nudge increases the perceived marginal utility of goods. For instance, a nudge could discourage the consumption of hamburgers or encourage the consumption of vegetables.

Proposition 2.3 (Optimal nudge formula) *At an interior optimum, nudges satisfy*

$$\frac{\partial L\left(\boldsymbol{\tau},\chi\right)}{\partial\chi}=\sum_{h}[\lambda\left(\boldsymbol{\tau}-\boldsymbol{\tau}^{\xi,h}-\widetilde{\boldsymbol{\tau}}^{b,h}\right)\cdot\boldsymbol{c}_{\chi}^{h}+\beta^{h}\frac{u_{\chi}^{h}}{v_{w}^{h}}]=0. \tag{11}$$

The optimality conditions for taxes (10) *are unchanged.*

This formula has four terms corresponding to the potentially conflicting goals of nudges. The first term, $\lambda\boldsymbol{\tau}\cdot\boldsymbol{c}_{\chi}^{h}$, captures the fact that the changes in behavior induced by nudges directly change tax revenues. The second term, $-\lambda\boldsymbol{\tau}^{\xi,h}\cdot\boldsymbol{c}_{\chi}^{h}$, captures the fact that the changes in behavior induced by nudges affect welfare and tax revenues through their effect on externalities. The third term, $-\lambda\widetilde{\boldsymbol{\tau}}^{b,h}\cdot\boldsymbol{c}_{\chi}^{h}$, captures the fact that the changes in behavior induced by nudges affect welfare because agents misoptimize. The fourth term, $\beta^{h}\frac{u_{\chi}^{h}}{v_{w}^{h}}$, captures the potential direct effects of nudges on utility.[22]

2.5 Discussion

We now discuss a few limitations and potential extensions of our approach, some of which we plan to investigate in future work.

Paternalism In our model, agents make mistakes that the government can identify, which in practice is a difficult task.[23] This approach, which is common but not uncontroversial, departs from the revealed preferences welfare paradigm and has elements of paternalism: the government tries to respect the agents' "true" preferences but recognizes that agents sometimes do not act in their own best interest (see Bernheim and Rangel (2009) for an in-depth discussion of this approach).

There are several important objections to this approach. For example, when agents behave in ways that do not fit economists' models, it may be that we do not understand their motives or constraints well enough. Then paternalism may simply be a misguided approach. In addition, governments may not be benevolent, or fully optimizing themselves, and face various forms of political economy and institutional constraints. While we acknowledge these objections, they are beyond the scope of this paper, which is to establish the benchmark model with a benevolent, knowledgeable government—leaving its relaxation to future work.[24]

[22] We note in passing that to date, the empirical literature (reviewed briefly below) has measured the impact of nudges on decisions ($\boldsymbol{c}_{\chi}^{h}$), but not (to the best of our knowledge) the impact on utility (u_{χ}^{h}).

[23] Arguably, agents' mistakes can be persistent. For example, Slemrod (2006) argues that Americans overestimate on average the odds their estate will be taxed. Similarly, people seem to perceive average for marginal tax rates (Liebman and Zeckhauser (2004)), and to overestimate the odds they'll move to a higher tax bracket (Bénabou and Ok (2001)). Second, our framework applies to situations where consumers do not maximize experienced utility. There, learning may be quite slow. For instance, people may persistently smoke too much, perhaps because of hyperbolic discounting (Laibson (1997)).

[24] In fact, the government need not be benevolent. For example, the government could be modeled as an agent and put a large weight on himself and small or even negative weights on other agents. Moreover, while we have assumed

Other Biases Despite our model's generality, there are categories of behavioral biases that it does not accommodate. First, our model only allows for intrapersonal but not for interpersonal behavioral deviations from the traditional model. For example, it leaves aside issues of fairness, relative comparisons, social norms, and social learning. Second, it is not ideally suited to capture information-based behavioral phenomena, such as self and social signaling as a motivation for behavior, or the potential signaling effects of taxes and nudges (see e.g. Bénabou and Tirole (2006) and references therein).

"Lucas Critique" A difficulty confronting all behavioral policy approaches is a form of the Lucas critique: how do the underlying biases change with policy? The empirical evidence is limited, but we try to bring it to bear in two places: when we analyze how past taxes influence the perception of current taxes (see Section 3.1) and when we discuss the endogeneity of attention to taxes (Section 3.5). We hope that more empirical evidence on this will become available as the field of behavioral public finance develops.

2.6 Measurement

Operationalizing our optimal tax formula requires taking a stand on the relevant sufficient statistics: social marginal value of public funds, social marginal utilities of income, elasticities, internalities, and externalities. For example, in the general Ramsey model, the optimal tax formula features the social marginal value of public funds λ, the social marginal utilities of income γ^h, consumption vectors $\boldsymbol{c}^h$, Slutsky matrices $\boldsymbol{S}^{C,h}$, and behavioral wedges $\tilde{\boldsymbol{\tau}}^{b,h}$.[25]

All these sufficient statistics are present in the optimal tax formula of the traditional model with no behavioral biases, with the exception of behavioral wedges $\tilde{\boldsymbol{\tau}}^{b,h}$. In principle, they can be estimated with rich enough data on observed choices.[26]

As already pointed out in Section 2.2, there are several ways to use estimates of these sufficient statistics. In general, estimates for a given tax system can be used to test for optimality of this tax system or to identify the direction of welfare-improving local tax reforms. With extra assumptions about functional forms, these sufficient statistics reflect constant deep parameters (e.g. demand

that the arguments of the social welfare function are the experienced utilities $v^h(\boldsymbol{q}, w) = u^h\left(\boldsymbol{c}^h(\boldsymbol{q}, w)\right)$ of the agents, we could alternatively have used other indirect utility functions divorced from the actual utilities experienced by the agents, and the analysis would still go through. For example, the true experienced utility of the agent might value whisky, but a "puritanical" government could place a negative value on whisky, for moral rather than instrumental (e.g. health) reasons.

[25] Sometimes a given bias can be modeled using two distinct particularizations (decisions vs. experienced utility and misperceptions). For example, in the absence of wealth effects, it is possible to capture non-salient taxes either using the decision vs. experienced utility model (as Mullainathan et al. (2012)) or using the misperception model (as we do here). These different approaches have the same implications for optimal taxation since they rationalize the same behavior (demands and elasticities) and capture the same mistakes (behavioral wedges).

[26] This is true except for the "social constructs" such as the social welfare function and its impact on γ^h. A possible approach is to vary these parameters to trace out the whole constrained Pareto frontier.

elasticities for isoelastic utility functions) and then local estimates can be used to compute globally optimal tax system, as we shall see below in Section 3.

In practice, this remains a momentous task, as the data and sources of exogenous variations are limited. 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 amount of time the tax system has been in place (allowing agents to become familiar with it); the presence of nudges, etc.[27]

The behavioral wedges $\tilde{\boldsymbol{\tau}}^{b,h}$, which summarize the effects of behavioral biases at the margin are arguably even harder to measure because estimating welfare is inherently challenging. This poses a problem similar to the more traditional problem of estimating marginal externalities $\boldsymbol{\tau}^{\xi,h}$ to calibrate corrective Pigouvian taxes in the traditional model with no behavioral biases. The common challenge is that these statistics are not easily recoverable from observations of private choices. In both cases, it is possible to use a structural model, but more reduced-form approaches are also feasible in the case of behavioral biases. This difficulty is arguably harder to navigate in the case of behavioral wedges, because of the higher dimensionality of the problem and the newness of the task.

Existing approaches to measuring behavioral wedges $\tilde{\boldsymbol{\tau}}^{b,h}$ can be divided in three broad categories. In Section 3 when we consider specific examples, we will attempt to draw from the existing empirical evidence to give a concrete sense of how to implement these principles.

1. Comparing choices in clear vs. confusing environments. A common strategy involves comparing choices in environments where behavioral biases are attenuated and environments resembling those of the tax system under consideration. Choices in environments where behavioral biases are attenuated can be thought of as rational, allowing the recovery of experienced utility u^h as a utility representation of these choices, with associated indirect utility function v^h.[28] Differences in choices in environments where behavioral biases are present would then allow to measure the marginal internalities $\boldsymbol{\tau}^{b,h} = \boldsymbol{q} - \frac{u^h_{\boldsymbol{c}}}{v^h_w}$.

For example, if the biases arise from the misperception of taxes so that $\boldsymbol{\tau}^{b,h} = \boldsymbol{\tau} - \boldsymbol{\tau}^{s,h}$, then perceived taxes $\boldsymbol{\tau}^{s,h}$ could be estimated by comparing consumption behavior in the environment under consideration where taxes might not be fully salient to consumption behavior in an environment where taxes are very salient (see e.g. Chetty et al. (2009), Allcott et al. (2014), and Feldman et al. (2016)). We flesh out the details regarding the implementation of this strategy in the quantitative illustration at the end of Section 3.1.

Another example is when agents may not fully understand the utility consequences of their choices, which can be captured with a decision vs. experienced utility model. For instance, Allcott and

[27] For example, Slutsky matrices might depend on a host a factors, e.g. the salience of taxes and the complexity of the environment, which can vary with the context in which the underlying data is gathered.

[28] Choices are more likely to reveal true preferences if agents have a lot of time to decide, taxes and long run effects are salient, and information about costs and benefits is readily available, etc.

Taubinsky (2015) study the purchases of energy-saving light bulbs with or without an intervention which gives information on potential savings in a field experiment. By comparing purchase decisions with and without treatment, they recover $\boldsymbol{\tau}^{b,h} = \frac{u_c^s}{v_w^s} - \frac{u_c}{v_w}$.[29]

2. Surveys. Another strategy, if behavioral biases arise from misperceptions, is to use surveys to directly elicit perceived taxes $\boldsymbol{\tau}^{s,h}$. See e.g. De Bartolomé (1995), Liebman and Zeckhauser (2004), and Slemrod (2006) for examples implementing this method.

3. Structural models. Finally, it is sometimes possible to use a calibrated structural model. For example, Lockwood and Taubinsky (2017) combine an assessment of the health consequences of soda consumption with a hyperbolic discounting model (Laibson (1997)) to estimate the associated internality. See Section 3.4 for a more detailed explanation.

2.7 A Useful Simple Specification

We close this section by working out a useful particularization of the general model which yields simple optimal tax formulas. This simple case will prove useful to construct many of our examples in Section 3.

We use a hybrid model with both decision vs. experienced utility and misperceptions. We make several simplifying assumptions: we assume that decision and experienced utility are quasi-linear so that the marginal utility of wealth is constant; we allow for a simple convenient form for misperceptions of taxes; we assume that externalities ξ are separable from consumption.

Formally, we decompose consumption $\boldsymbol{c} = (c_0, \boldsymbol{C})$ with $\boldsymbol{C} = (c_1, ..., c_n)$ and we normalize $p_0 = q_0 = 1$, as good 0 is assumed to be untaxed. The experienced utility of agent h is quasilinear

$$u^h(c_0, \boldsymbol{C}, \xi) = c_0 + U^h(\boldsymbol{C}) - \xi,$$

where $\xi = \xi((\boldsymbol{C}^h)_{h=1,...,H})$ is an externality. Agent h is subject to two sets of biases. First, taking ξ as given he maximizes a decision utility

$$u^{s,h}(c_0, \boldsymbol{C}, \xi) = c_0 + U^{s,h}(\boldsymbol{C}) - \xi,$$

which differs from his experienced utility, but remains quasilinear. Second, while the true after-tax price is $\boldsymbol{q} = \boldsymbol{p} + \boldsymbol{\tau}$, he perceives prices to be

$$\boldsymbol{q}^{s,h} = \boldsymbol{p} + \boldsymbol{M}^h \boldsymbol{\tau}, \tag{12}$$

where $\boldsymbol{M}^h$ is a constant matrix of marginal perceptions (which in practice will be diagonal $\boldsymbol{M}^h =$

[29] Consider yet another example: if the biases arise because of temptation, then standard choices would reveal decision utility $u^{s,h}$. To the extent that agents are sophisticated and understand that they are subject to these biases, experienced utility u^h could be recovered by confronting agents with the possibility of restricting their later choice sets. In the terminology of Bernheim and Rangel (2009), this strategy uses refinements to uncover true preferences.

$diag\left(m_i^h\right)_{i=1,\ldots,n}$). The corresponding perception function is $\boldsymbol{q}^{s,h}(\boldsymbol{q}) = \boldsymbol{p} + \boldsymbol{M}^h(\boldsymbol{q} - \boldsymbol{p})$.[30]

The demand $\boldsymbol{c}^h(\boldsymbol{q}, w, \xi) = \left(c_0^h(\boldsymbol{q}, w), \boldsymbol{C}^h(\boldsymbol{q})\right)$ of agent h is such that $\boldsymbol{C}^h(\boldsymbol{q}) = \boldsymbol{C}^{s,h}\left(\boldsymbol{q}^{s,h}(\boldsymbol{q})\right)$ and $c_0^h(\boldsymbol{q}, w) = w - \boldsymbol{q} \cdot \boldsymbol{C}^h(\boldsymbol{q})$, where $\boldsymbol{C}^{s,h}\left(\boldsymbol{q}^{s,h}\right) = \arg\max_{\boldsymbol{C}} U^{s,h}(\boldsymbol{C}) - \boldsymbol{q}^{s,h} \cdot \boldsymbol{C}$. Because decision utility is quasilinear, there are no income effects and we have $\boldsymbol{S}^{C,h}(\boldsymbol{q}, w) = \boldsymbol{S}^{r,h}\left(\boldsymbol{q}^{s,h}(\boldsymbol{q})\right)\boldsymbol{M}^h$, where $\boldsymbol{S}^{r,h}\left(\boldsymbol{q}^{s,h}\right) = \frac{\partial \boldsymbol{C}^{s,h}\left(\boldsymbol{q}^{s,h}\right)}{\partial \boldsymbol{q}^{s,h}}$ is the rational Slutsky matrix.

We also define the internality wedge $\boldsymbol{\tau}^{I,h} = U_{\boldsymbol{C}}^{s,h}(\boldsymbol{C}) - U_{\boldsymbol{C}}^h(\boldsymbol{C})$ and the internality/externality wedge $\boldsymbol{\tau}^{X,h} = \frac{\beta^h}{\lambda}\boldsymbol{\tau}^{I,h} + \boldsymbol{\tau}^{\xi,h}$.[31] Because there are no wealth effects in consumption, we have $\gamma^{\xi,h} = \gamma^h = \beta^h$. We now characterize optimal taxes.

Proposition 2.4 (Optimal tax formula with constant marginal utility of wealth and constant misperceptions) *In the constant marginal utility of wealth and constant misperceptions specification of the general model, optimal taxes satisfy*

$$\boldsymbol{\tau} = -\left[\sum_h \boldsymbol{M}^{h\prime}\boldsymbol{S}^{r,h}\left(\boldsymbol{I} - \left(\boldsymbol{I} - \boldsymbol{M}^h\right)\frac{\gamma^h}{\lambda}\right)\right]^{-1}\sum_h\left[\left(1 - \frac{\gamma^h}{\lambda}\right)\boldsymbol{C}^h - \boldsymbol{M}^{h\prime}\boldsymbol{S}^{r,h}\boldsymbol{\tau}^{X,h}\right]. \tag{13}$$

This formula is a direct application of the tax formulas in Propositions 2.1 and 2.2, obtained by particularizing the general model, and by solving the system of linear equations in taxes $\boldsymbol{\tau}$ formed by these tax formulas. Optimal taxes are expressed as a function of sufficient statistics. As explained above, these sufficient statistics are endogenous since they must be computed at the optimum.

A complete closed-form solution as a function of primitives is unavailable in general. Such closed forms with explicit comparative statics are available in two special cases that we will put to use in our concrete examples: when utility is isoelastic and when it is quadratic. The examples in Section 3.1-3.4 are exact applications of this formula (13).

Closed-form solutions can also be obtained as an approximation in the limit of small taxes often emphasized in public finance, and to which we now turn. We assume a utilitarian welfare functions with exogenous Pareto weights b^h. Since utility is quasi-linear, we have $\gamma^{\xi,h} = \gamma^h = \beta^h = b^h$. To consider the limit of small taxes, we assume that $b^h - \lambda$ is small, and that $\boldsymbol{\tau}^{I,h}$ and $\boldsymbol{\tau}^{\xi,h}$ are small when taxes are equal to 0 (and hence that they remain small for small taxes). To be formal, we introduce a small disturbance parameter $\eta = \sum_h\left(\left|b^h - \lambda\right| + \|\boldsymbol{\tau}^{I,h}\| + \|\boldsymbol{\tau}^{\xi,h}\|\right)$ and we compute a Taylor expansions in η around 0.

[30] In all those definitions, we omit the row and columns corresponding to good 0, which has no taxes and no misperceptions.

[31] The wedge $\boldsymbol{\tau}^{I,h}$ is closely related to the behavioral wedge $\boldsymbol{\tau}^{b,h}$ according to $\boldsymbol{\tau}^{b,h} = \boldsymbol{\tau}^{I,h} + \left(\boldsymbol{I} - \boldsymbol{M}^h\right)\boldsymbol{\tau}$. Basically, $\boldsymbol{\tau}^{b,h}$ captures two forms of misoptimization: those arising from the difference between decision and experienced utility ($\boldsymbol{\tau}^{I,h}$) and those arising from the misperception of taxes ($\left(\boldsymbol{I} - \boldsymbol{M}^h\right)\boldsymbol{\tau}$). In this example, we find it useful to separate them.

Optimal taxes can then be solved in terms of fundamentals, up to the second order in η:

$$\boldsymbol{\tau} = -\left[\sum_h \boldsymbol{M}^{h\prime}\boldsymbol{S}^{r,h}(\boldsymbol{p},w)\,\boldsymbol{M}^h\right]^{-1}\sum_h\left[\left(1-\frac{b^h}{\lambda}\right)\boldsymbol{C}^h(\boldsymbol{p},w)\right] + \left[\sum_h \boldsymbol{M}^{h\prime}\boldsymbol{S}^{r,h}(\boldsymbol{p},w)\,\boldsymbol{M}^h\right]^{-1}\sum_h\left[\boldsymbol{M}^{h\prime}\boldsymbol{S}^{r,h}(\boldsymbol{p},w)\left(\boldsymbol{\tau}^{I,h}(\boldsymbol{p},w)+\boldsymbol{\tau}^{\xi,h}(\boldsymbol{p},w)\right)\right]. \quad (14)$$

We emphasize that in this formula, the objects on the right-hand side are evaluated at the zero-tax equilibrium, and can therefore be taken to be primitives (independent of taxes). It can be broken down in the different motives for taxation: the revenue-raising and redistributive motives (the first term on the right-hand side), and the internality-externality corrective motives (the second term on the right-hand side). In the appendix we derive a similar formulation (46) for the optimal tax, without assuming quasilinear utility.

Equation (14) delivers some explicit comparative statics. In response to a change $d\boldsymbol{\tau}^{I,h}(\boldsymbol{p},w)$ in the internalities, we have the following average comparative static result (up to the third order in η):[32]

$$d\boldsymbol{\tau}'\sum_h \boldsymbol{M}^{h\prime}\boldsymbol{S}^{r,h}(\boldsymbol{p},w)\,d\boldsymbol{\tau}^{I,h}(\boldsymbol{p},w)\leq 0. \quad (15)$$

This means that on average, the change in optimal perceived taxes co-moves positively with the change in internalities. For example, imagine that for all agents, the attention matrix $\boldsymbol{M}^h$ is diagonal and positive and that all other goods are substitutes with good i. Suppose that for all agents, we increase the internality of good i and that we decrease the internalities for the other goods so that $d\tau_i^{I,h}(\boldsymbol{p},w)\geq 0$ and $d\tau_j^{I,h}(\boldsymbol{p},w)\leq 0$ for all $j\neq i$. As is intuitive, the optimal tax system then redirects consumption away from good i and towards the other goods.[33]

Moreover, formula (14) allows us to anticipate some important insights that will come out of our simple examples in Sections 3.1 and 3.2. For instance, it clarifies the scaling of the different tax motives with inattention $\left(\boldsymbol{M}^h\right)^{-1}$: the revenue-raising and redistributive terms scale quadratically in inattention, whereas the internality-externality corrective terms scale linearly in inattention. We now turn to developing these intuitions in depth.

[32] Indeed (14) gives $d\boldsymbol{\tau}=Q^{-1}d\boldsymbol{x}$ with $d\boldsymbol{x}=\sum_h \boldsymbol{M}^{h\prime}\boldsymbol{S}^{r,h}(\boldsymbol{p},w)\,d\boldsymbol{\tau}^{I,h}(\boldsymbol{p},w)$ and $Q=\sum \boldsymbol{M}^{h\prime}\boldsymbol{S}^{r,h}(\boldsymbol{p},w)\,\boldsymbol{M}^h$ is a negative definite matrix, as rational Slutsky matrices are negative semi-definite. So, $d\boldsymbol{\tau}'d\boldsymbol{x}=d\boldsymbol{\tau}'Qd\boldsymbol{\tau}\leq 0$, i.e. (15).

[33] Indeed, the attention-weighted discouragement of good i increases and is given up to the first order in η, by:$-\left(\left[\sum_h \boldsymbol{M}^{h\prime}\boldsymbol{S}^{r,h}(\boldsymbol{p},w)\,\boldsymbol{M}^h\right]d\boldsymbol{\tau}\right)_i/c_i=-\left(\sum_h \boldsymbol{M}^{h\prime}\boldsymbol{S}^{r,h}(\boldsymbol{p},w)\,d\boldsymbol{\tau}^{I,h}(\boldsymbol{p},w)\right)_i/c_i\geq 0$.

3 Examples

3.1 Basic Ramsey Problem: Raising Revenues with Behavioral Agents

Inverse Elasticity Rule: A Behavioral Version We start by developing a behavioral version of the canonical Ramsey inverse elasticity rule. The government must raise revenues using linear commodity taxes $\boldsymbol{\tau}$ with marginal utility of public funds λ. Following the tradition, we start with a homogeneous population of agents (so that we can drop the h superscript), with welfare weight γ. We define $\Lambda = 1 - \frac{\gamma}{\lambda}$ so that a higher Λ corresponds to a higher relative benefit of raising revenues. Utility is $c_0 + \sum_{i=1}^{n} \frac{c_i^{1-1/\psi_i}-1}{1-1/\psi_i}$. The only bias is that the agent perceives the tax τ_i as $\tau_i^s = m_i \tau_i$, where $m_i \in (0,1]$ captures the attention to the tax. The rational case corresponds to $m_i = 1$.

This setup is a particular case of that of Section 2.7, and the behavioral Ramsey formula in Proposition 3.1 can be derived by specializing our tax formula (13).[34] However, we find it useful to also provide a short self-contained rendition. The Ramsey planning problem is

$$\max_{\{\tau_i\}} \gamma \sum_{i=1}^{n} \left[\frac{[c_i(\tau_i)]^{1-1/\psi_i} - 1}{1 - 1/\psi_i} - (p_i + \tau_i) c_i(\tau_i) \right] + \lambda \sum_{i=1}^{n} \tau_i c_i(\tau_i), \tag{16}$$

where $c_i(\tau_i) = (p_i + m_i \tau_i)^{-\psi_i}$ is the demand of the consumer perceiving the price to be $p_i + m_i \tau_i$. We can then derive the optimal tax formula by taking first-order conditions in this planning problem.

Proposition 3.1 (Modified Ramsey inverse elasticity rule) *The optimal tax on good i is*

$$\frac{\tau_i}{p_i} = \frac{\Lambda}{m_i^2 \psi_i} \cdot \frac{1}{1 + \Lambda \left(\frac{1 - m_i - 1/\psi_i}{m_i} \right)}. \tag{17}$$

When $m_i = 1$ so that the tax is fully salient, we recover the traditional Ramsey inverse elasticity rule which states that taxes decrease with the elasticity ψ_i of the demand for the good and increase with Λ. When $m_i < 1$ so that the tax is less than fully salient, the tax is higher. Mullainathan et al. (2012) discuss intuitively that taxes should be higher when they are underperceived, but do not derive a formal mathematical behavioral counterpart to the Ramsey inverse elasticity rule.

To gain intuition for equation (17), we now follow an instructive tradition of public finance and consider the limit of small taxes (i.e. the small Λ limit) where the distortions from taxation take the form of Harberger triangles. Up to the first order in Λ, optimal taxes are then given by the first term $(\frac{\Lambda}{m_i^2 \psi_i})$ in equation (17), and the second term only introduces second order corrections.

In the limit of small taxes, the traditional Ramsey inverse elasticity rule prescribes that the

[34] Simply take $\boldsymbol{M} = diag\,(m_i)_{i=1,...,n}$ (which is the diagonal matrix of with entries m_i for $i = 1, ..., n$), $\boldsymbol{S}^r = -diag(\frac{c_i \psi_i}{q_i^s})_{i=1,...,n}$, and $\boldsymbol{\tau}^X = 0$.

optimal tax should be $\frac{\tau_i^R}{p_i} = \frac{\Lambda}{\psi_i}$. With inattention, optimal taxes are higher at[35]

$$\frac{\tau_i}{p_i} = \frac{\Lambda}{m_i^2 \psi_i}. \tag{19}$$

Loosely speaking, this is because inattention makes agents less elastic. Given partial attention $m_i \leq 1$, the effective elasticity of the demand for good i is $m_i \psi_i$, rather than the parametric elasticity ψ_i. In the spirit of the traditional Ramsey formula, a lower elasticity leads to higher optimal taxes.[36] However, a naive application of the Ramsey rule would lead to the erroneous conclusion that $\frac{\tau_i}{p_i} = \frac{\Lambda}{m_i \psi_i}$ rather than $\frac{\tau_i}{p_i} = \frac{\Lambda}{m_i^2 \psi_i}$. The discrepancy arises because it is the perceived tax, and not the true tax, that should be inversely proportional to the effective demand elasticity: $\frac{\tau_i^s}{p_i} = \frac{\Lambda}{m_i \psi_i}$.[37]

The upshot of this analysis is that optimal taxes τ_i tend to increase relatively fast with inattention m_i. In the limit of small taxes, the increase is exactly quadratic so that taxes scale with $\frac{1}{m_i^2}$.[38]

Heterogeneity in Attention We now turn our attention to the case where perceptions of taxes are heterogeneous.[39]

We suppose that type h has attention m_i^h to the tax on good i. The optimal tax is again a

[35] It is enlightening to walk through a self-contained derivation of optimal taxes in that limit. The objective function of the government admits the following second order approximation: $\mathcal{L}(\boldsymbol{\tau}) - \mathcal{L}(0) = L(\boldsymbol{\tau}) + o\left(\|\boldsymbol{\tau}\|^2\right) + O\left(\Lambda \|\boldsymbol{\tau}\|^2\right)$, with

$$L(\boldsymbol{\tau}) = \frac{-1}{2} \sum_{i=1}^{n} \left(\frac{\tau_i^s}{p_i}\right)^2 \psi_i y_i + \Lambda \sum_{i=1}^{n} \frac{\tau_i}{p_i} y_i, \tag{18}$$

where $\tau_i^s = m_i \tau_i$ is the perceived tax, y_i expenditure on good i at zero taxes. This approximation neatly separates the benefits of taxation in the form of increased revenues $\Lambda \sum_{i=1}^{n} \frac{\tau_i}{p_i} y_i$ from the distortionary cost of taxation and $\frac{-1}{2} \sum_{i=1}^{n} \left(\frac{\tau_i^s}{p_i}\right)^2 \psi_i y_i$ as the area of Harberger triangles (the latter was also derived by Chetty et al. (2009)). The key observation is that the cost of taxation depends on perceived taxes while the revenues depend on true taxes. Optimal taxes can be derived by solving $L'(\boldsymbol{\tau}) = 0$.

[36] Finkelstein (2009) finds evidence for this effect. When highway tolls are paid automatically thus are less salient, people are less elastic to them, and the government reacts by increasing the toll (i.e., the tax rate).

[37] To gain intuition, consider the effect of a marginal increase in $\frac{\tau_i}{p_i}$. The marginal benefit in terms of increased tax revenues is Λy_i, the marginal cost in terms of increased distortions is $\frac{\tau_i^s}{p_i} m_i \psi_i y_i$, where y_i is the expenditure on good i when there are no taxes. At the optimum, the marginal cost and the marginal benefit are equalized. The result is that $\frac{\tau_i^s}{p_i} = \frac{\Lambda}{m_i \psi_i}$, i.e. it is the perceived tax $\frac{\tau_i^s}{p_i}$ that is inversely related to the effective elasticity $m_i \psi_i$. This in turns implies $\frac{\tau_i}{p_i} = \frac{\tau_i^s / p_i}{m_i} = \frac{\Lambda}{m_i^2 \psi_i}$.

[38] The specific formulation of misperception that we have used in this section assumes that the budget adjustments required when agents misperceive taxes are all absorbed by the consumption a good (good 0) with a constant marginal utility. This renders these adjustments relatively painless. Section 9.3.3 of the online appendix considers an alternative, budget adjustments are concentrated on a "shock absorber" good with a sharply decreasing marginal utility. Then, the optimal tax τ_i is lower than in the baseline case (17), particularly for less salient taxes with a small m_i.

[39] For instance, the poor might pay more attention to the price of the goods they currently buy, while perhaps paying less attention to some future consequences of their actions. For explorations of the demographic correlates of attention, see Mani et al. (2013),Taubinsky and Rees-Jones (2017).

particular case of formula (13). With isoelastic utility, no closed-form solution is available, and so we directly place ourselves in the limit of small taxes to derive analytical insights.[40] We confirm the validity of these intuitions in our quantitative illustration at the end of this section, where we do not rely on this approximation.

Optimal taxes are now given by[41]

$$\frac{\tau_i}{p_i} = \frac{\Lambda}{\psi_i \mathbb{E}\left[m_i^{h\,2}\right]} = \frac{\Lambda}{\psi_i \left(\mathbb{E}\left[m_i^h\right]^2 + var\left[m_i^h\right]\right)}, \tag{20}$$

where here and elsewhere $\mathbb{E}$ and var denote respectively the average and the variance computed over the different types h of agents. Controlling for average attention $\mathbb{E}\left[m_i^h\right]$ (which determines the effective elasticity of total demand to the tax), an increase in the heterogeneity of attention $var\left[m_i^h\right]$ reduces the optimal tax. The intuition is that heterogeneity in attention introduces a further cost of taxation in the form of misallocation across consumers who do not all perceive the same post-tax price.

Endogenous Social Cost of Public Funds So far, our treatment of the basic Ramsey problem is for a given value of the social cost of public funds λ. Here we briefly show how to account for the potential endogeneity of λ. For simplicity, we confine ourselves to the case of a representative agent.

We assume that the government solves the following planning problem:

$$\max_{\{\tau_i, G\}} \gamma \sum_{i=1}^{n} \left[\frac{\left[c_i(\tau_i)\right]^{1-1/\psi_i} - 1}{1 - 1/\psi_i} - (p_i + \tau_i) c_i(\tau_i) \right] + V(G), \tag{21}$$

subject to the revenue constraint $G = \sum_{i=1}^{n} \tau_i c_i(\tau_i)$, where $V(G)$ is a concave function of spending on public goods. The social cost of public funds is then given by $\lambda = V'(G)$ where G is computed at the optimum.

Consider the limit of small taxes (i.e., for small $V'(0) - \gamma$). Up to the first order, optimal taxes are still given by equation (19) but now Λ is endogenously given by

$$\Lambda = \frac{V'(0) - \gamma}{\gamma - V''(0) \sum_{i=1}^{n} \frac{p_i c_i(0)}{\psi_i m_i^2}}. \tag{22}$$

[40] For an exact closed-form derivation with quadratic utility, see the online appendix (Section 9.1.2).

[41] This can be directly seen by maximizing the second order approximation of the objective function of the government, valid for small Λ and small taxes:

$$\frac{1}{H} L(\boldsymbol{\tau}) = \frac{-1}{2} \sum_{i=1}^{n} \mathbb{E}\left[m_i^{h\,2}\right] \left(\frac{\tau_i}{p_i}\right)^2 \psi_i y_i + \Lambda \sum_{i=1}^{n} \frac{\tau_i}{p_i} y_i.$$

The exogenous Λ case arises when the marginal utility of government spending is constant so that $V''(0) = 0$. The more inattentive agents are to taxes, the less costly it is for the government to raise a given amount of revenues, the more the government spends on public goods, and the higher the taxes that it decides to set.

When there is decreasing marginal utility of government spending ($V''(0) < 0$) , the marginal utility of government spending is lower for any given level of government spending, and so the government sets lower taxes, raises less revenues, and spends less on public goods whether or not there is inattention, compared to the case where the marginal utility of government spending is constant. Interestingly, in this case, at the optimum, the social cost of public funds Λ, which is equal to $(V'(G) - \gamma)/\gamma$, decreases with inattention. This indicates that just like in the exogenous Λ case, the government still raises more revenues and spends more on public goods when agents are more inattentive (but less so than in the exogenous Λ case).

Quantitative Illustration To gauge the real-world importance of these effects, we calibrate the model with heterogeneity in misperceptions, based on the findings of Taubinsky and Rees-Jones (2017) for sales taxes. Sales taxes are not included in the tag price. To elicit their salience, Taubinsky and Rees-Jones design an online experiment and elicit the maximum tag price that agents would be willing to pay when there are no taxes or when there are standard taxes corresponding to their city of residence (in the latter case, they are not reminded what the tax rate is). In our notation, the ratio of these two prices is $1 + m^h \frac{\tau}{p}$, where p is the maximum tax price when there are no taxes (we focus a a given good, and suppress the index i). This allows the estimation of tax salience m^h.

Taubinsky and Rees-Jones (2017) find (in their standard tax treatment)[42] $\mathbb{E}\left[m^h\right] = 0.25$ and $var\left(m^h\right) = 0.13$, so that heterogeneity is very large, $\frac{var\left(m^h\right)}{\mathbb{E}\left[m^h\right]^2} = \frac{0.13}{0.25^2} = 2.1$.[43] In our calibration, we take $\psi = 1$ (as in the Cobb-Douglas case, which is often a good benchmark for the elasticity between broad categories of goods) and $\Lambda = 1.25\%$, which is consistent with the baseline tax in their setup, at $\tau = 7.3\%$.[44] If the tax became fully salient, the optimal tax would be divided by 5.7. If heterogeneity disappeared (but keeping mean attention constant), the optimal tax would be

[42] They actually provide a lower bound on variance, and for simplicity we take it here to be a point estimate.

[43] In the standard tax vs. no tax treatment, participants react to the tax rate as if they are 25% of their real sizes. But in triple tax vs. no tax treatment, this ratio rises to 50%. But as suggested by multiple-choice questions before their purchase decisions, participants know the tax rate well. The estimate of mean attention is broadly consistent with the results of Chetty et al. (2009) using a field experiment, who finds a mean attention of between 0.06 (by computing the ratio of the semi-elasticities for sales taxes, which are not included in the sticker price, vs. excise taxes, which are included in the sticker price) and 0.35 (computing the ratio of the semi-elasticities for sales taxes vs. more salient sticker prices).

[44] For illustration purposes, we calibrate Λ imagining that the government sets taxes optimally, given behavioral elasticities and inattention. We use a two-point distribution with rational and behavioral agents to match the mean and dispersion of attention.

multiplied by 2.8.[45]

We conclude that the extant empirical evidence and our simple Ramsey model indicate that the mean and dispersion of attention have a sizable impact on optimal taxes.

3.2 Basic Pigou Problem: Externalities, Internalities, and Inattention

Dollar for Dollar Principle: A Behavioral Version The analysis in this section is a direct application of formula (13). However, to help build intuition, we start with an elementary and self-contained analysis of the basic Pigou problem. We then use formula (13) to derive more complex generalizations.

We continue to assume a quasilinear utility function. We assume that there is only one taxed good $n = 1$. The representative agent maximizes $u(c_0, c) = c_0 + U(c)$ subject to $c_0 + pc \leq w$. Here c stands for the consumption of good 1 (we could call it c_1, but expressions are cleaner by calling it c). If the representative agent were rational, he would solve

$$\max_c U(c) - pc. \tag{23}$$

However, there is a negative externality that depends on the aggregate consumption of good 1 (think for example of second-hand smoke), so that total utility is $c_0 + U(c) - \xi_* c$. Alternatively, ξ_* could be an internality (think for example of the temptation to smoke): a divergence between decision utility $c_0 + U(c)$ and experienced utility $c_0 + U(c) - \xi_* c$.

To focus on the corrective role of taxes, we assume that $\Lambda = 0$ and that the government can rebate tax revenues lump-sum to consumers. The objective function of the government is therefore

$$U(c) - (p + \xi_*) c. \tag{24}$$

To attempt to correct the externality/internality, the government can set a tax τ. Consider first an agent who correctly perceives taxes and solves

$$\max_c U(c) - (p + \tau) c.$$

The optimal tax is then $\tau = \xi_*$, ensuring that the agent maximizes the same objective function as that of the government. This is the classic Pigouvian prescription: a dollar of externality/internality must be corrected with a dollar of tax so that the agent fully internalizes the externality/internality.

[45] The numbers we report in the main text use formula (8) without any approximation. To get a feel for these magnitudes, however, it is useful to consider the small tax approximation. Then, if the tax became fully salient, the optimal tax would be divided by 5 (multiplied by $\mathbb{E}\left[m^h\right]^2 + var\left(m^h\right) \simeq 0.2$). If heterogeneity disappeared (but keeping mean attention constant), the optimal tax would be multiplied by $\frac{\mathbb{E}\left[m^h\right]^2 + var\left(m^h\right)}{\mathbb{E}[m^h]^2} \simeq 3$).

Now consider an agent who only perceives a fraction m of the tax. Then he solves

$$\max_c U(c) - (p + m\tau) c. \tag{25}$$

The optimal Pigouvian corrective tax required to ensure that agents correctly internalize the externality/internality is now $\tau = \frac{\xi_*}{m}$. A dollar of externality must now be corrected with $\frac{1}{m^*}$ dollars of tax. We record this simple modification of the "dollar-for-dollar" principle.[46]

Proposition 3.2 (Modified Pigou formula) *In the basic Pigou problem with misperceptions, the optimal Pigouvian corrective tax is modified by inattention according to* $\tau = \frac{\xi_*}{m}$.

It is interesting to contrast this result with the modified optimal Ramsey tax (Proposition 3.1), for which $\frac{\tau_i}{p_i} = \frac{\Lambda}{\psi_i}\frac{1}{m_i^2}$ in the limit of small taxes. Partial attention m_i leads to a multiplication of the traditional tax by $\frac{1}{m_i}$ in the Pigou case and by $\frac{1}{m_i^2}$ in the Ramsey case.

The intuition is as follows. At the optimum in the Ramsey case, the perceived tax $m_i\tau_i$ is proportional to the inverse of the demand elasticity $m_i\psi_i$ which is itself reduced by inattention. At the optimum in the Pigou case, the perceived tax $m_i\tau_i$ is set equal to the externality ξ_* which is independent of inattention.

If different consumers have heterogeneous perceptions, then Proposition 3.2 suggests that no uniform tax can perfectly correct all of them. Hence, heterogeneity in attention prevents the implementation of the first best.[47]

Heterogeneity in Attention, Externality, Internality We now explore this issue more thoroughly. We assume that there are several consumers, indexed by $h = 1, ..., H$, all with the same welfare weight $\gamma^h = \beta^h = \lambda$. Agent h maximizes $u^h\left(c_0^h, c^h\right) = c_0^h + U^h\left(c^h\right)$. The associated externality/internality is $\xi^h c^h$. To be more precise, in the internality case, $U^{s,h}\left(c^h\right) - U^h\left(c^h\right) = \xi^h c^h$, and in the externality case, the externality is $\xi = \frac{1}{H}\sum_h \xi^h c^h$. Agent h pays an attention m^h to the tax so that perceived taxes are $\tau_h^s = m^h\tau$.

To get closed forms solutions, we specify utility to be quadratic:

$$U^h(c) = \frac{a^h c - \frac{1}{2}c^2}{\Psi}, \tag{26}$$

which implies a demand function $c^h(q^s) = a^h - \Psi q^s$.[48] We call $c^{*h} = \arg\max_{c^h} U^h\left(c^h\right) - \left(p + \xi^h\right) c^h$ the quantity consumed by the agent at the first best.

[46] The intuition that Pigouvian taxes should be higher when they are not fully salient is also discussed in Mullainathan et al. (2012) and could be formalized using their framework.

[47] If the budget adjustment is concentrated on a "shock absorber" good with a sharply decreasing marginal utility, then we obtain another force making Pigouvian taxes more distortionary, resulting in lower optimal Pigouvian taxes. This is developed in Section 9.3.3 of the online appendix.

[48] The expressions in the rest of this section are exact with this quadratic utility specification. For general utility functions, they hold provided that they are understood as the leading order terms in a Taylor expansion around an economy with no heterogeneity.

With heterogeneous externality or attention, to reach the first best, we would need a person-specific Pigouvian tax, $\frac{\xi^h}{m^h}$. However, under our maintained assumption of a single uniform tax, the first best cannot be implemented except in the knife-edge case where $\frac{\xi^h}{m^h}$ is the same across agents. A direct application of formula (13) yields the optimal Pigouvian tax:[49]

$$\tau^* = \frac{\mathbb{E}\left[\xi^h m^h\right]}{\mathbb{E}\left[m^{h2}\right]} = \frac{\mathbb{E}\left[\xi^h\right]\mathbb{E}\left[m^h\right] + cov\left(\xi^h, m^h\right)}{\mathbb{E}\left[m^h\right]^2 + var\left[m^h\right]}. \tag{27}$$

As in the Ramsey case, an increase in the heterogeneity of inattention $var\left(m^h\right)$ reduces the optimal tax. The intuition is that heterogeneity in attention introduces a further cost of taxation in the form of misallocation across consumers. In addition, the optimal tax is higher if the tax is better targeted in the sense that agents with a higher externality/internality ξ^h pay more attention to the tax, as measured by $cov\left(\xi^h, m^h\right)$. See Allcott et al. (2015) for a study where subsidies to weatherization are hampered by the fact that people who benefit the most pay the least attention.

Inattention and Tax vs. Quantity Regulation We continue to use the assumptions of heterogenous consumers with quasilinear utilities and utilitarian government with no motives of raising revenues or redistributing. The fact that the first best is generally not achievable in the presence of heterogeneity opens up a potential role for quantity regulations. Suppose the government imposes a uniform quantity restriction, mandating $c^h = c^*$. A simple calculation reveals that the optimal quantity restriction is given by the intuitive formula $c^* = \mathbb{E}\left[c^{h*}\right]$.

The following proposition compares optimal Pigouvian regulation and optimal quantity regulation. We consider a situation where the planner implements either an optimal Pigouvian tax, or an optimal quantity regulation, but not both policies.

Proposition 3.3 (Pigouvian tax vs Quantity regulation) *Consider a Pigouvian tax or a quantity restriction in the basic Pigou problem with misperceptions and heterogeneity. Quantity restrictions are superior to corrective taxes if and only if*

$$\frac{1}{2\Psi} var\left(c^{h*}\right) < \Psi \frac{\mathbb{E}\left[\xi^{h2}\right]\mathbb{E}\left[m^{h2}\right] - \left(\mathbb{E}\left[\xi^h m^h\right]\right)^2}{2\mathbb{E}\left[m^{h2}\right]}. \tag{28}$$

where the left-hand side is the welfare loss under optimal quantity regulation, and the right-hand side the welfare loss under optimal Pigouvian taxation.

Consider first the case with homogeneous attention ($m^h = m$). Then, the right-hand side of (28) is $\Psi \frac{var\left(\xi^h\right)}{2}$. Quantity restrictions tend to dominate taxes if heterogeneity in externalities/internalities is high compared to the heterogeneity in preferences. A higher demand elasticity

[49] This is a direct application of formula (13), with one non-quasilinear good, $\boldsymbol{M}^h = m^h$, $\boldsymbol{S}^{r,h} = -\Psi$, $\boldsymbol{\tau}^{X,h} = \xi^h$.

(high Ψ) favors quantity restrictions, because agents suffer less from a given deviation from their optimal quantity and more from a given price distortion. This generalizes the results in Weitzman (1974) who provided a treatment in the case with full (and hence homogeneous) attention.

Let us turn to the case with homogeneous externalities ($\xi^h = \xi$). Then, the right-hand side of (28) is $\Psi\xi^2 \frac{var\left(m^h\right)}{2\mathbb{E}\left[m^{h^2}\right]}$. Whether or not quantity restrictions dominate taxes is determined by similar principles as in the previous case, with heterogeneity now in attention instead of externalities. Heterogeneity of attention renders taxes less attractive because they introduce misallocation across consumers but do not affect the effectiveness of quantity restrictions, and this difference in effectiveness is magnified by high elasticies of substitution.[50] Note however a difference: with homogeneous attention, the common level of attention is irrelevant whereas with homogeneous externalities, the common level of the externality is relevant and higher levels favor quantity restrictions.

Now consider the case where both externalities/internalities and attention are heterogeneous. There is then an interaction effect: the tax is more attractive to the extent that it is better targeted in the sense that $cov\left(\xi^h, m^h\right)$ is higher.

One might naively have thought that the optimality criterion (28) for taxes vs. quantity restrictions could be derived by simply taking the full attention criterion $\frac{1}{2\Psi} var\left(c^{h*}\right) < \Psi \frac{var\left(\xi^h\right)}{2}$ and replacing the full-attention ideal person-specific tax ξ^h by its generalization ξ^h/m^h in the presence of inattention. This heuristic would lead to an erroneous criterion.[51]

Quantitative Illustration To get a sense of magnitudes, we use again the empirical findings of Taubinsky and Rees-Jones (2017) regarding the mean and dispersion of attention ($\mathbb{E}\left[m^h\right] = 0.25$ and $var\left(m^h\right) = 0.13$). We consider the case where the internality/externality ξ is the same across agents.[52] We saw that that optimal Pigouvian tax is $\tau^* = \xi \frac{\mathbb{E}\left[m^h\right]}{\mathbb{E}\left[m^h\right]^2 + var\left(m^h\right)}$. In the baseline case with heterogeneity, their numbers lead to $\tau^* = 1.3\xi$. If the tax became fully salient (i.e. $m^h = 1$), it would be divided by 1.3. If heterogeneity disappeared (i.e. $m^h = 0.25$), the optimal tax would be multiplied by $\frac{\mathbb{E}\left[m^h\right]^2 + var\left(m^h\right)}{\mathbb{E}\left[m^h\right]^2} = 3$. As in the Ramsey case, the effects of attention and its heterogeneity on optimal taxes are important.

[50] Very heterogeneous attention will not always lead to preferring quantity regulations—it will do so if and only if the losses from quantity regulation are less than those of under zero Pigouvian tax (i.e. $\frac{1}{2\Psi} var\left(c^{h*}\right) \leq \Psi \frac{\mathbb{E}\left[\xi^{h^2}\right]}{2}$). This will be the case if preference heterogeneity is small enough. The proof is as follows: in the worse-case scenario for attention, Pigouvian taxes lose their potency (the maximization of the right-hand side of (28) corresponds to full attention for a fraction π of the population, 0 attention for the rest, and letting π go to 0), and the loss is then the loss under laissez-faire, $\Psi \frac{\mathbb{E}\left[\xi^{h^2}\right]}{2}$.

[51] To gain intuition, note that (28) can be also written as $\frac{1}{2\Psi} var\left(c^{h*}\right) < \Psi \frac{var_{m^{h^2}}\left(\xi^h/m^h\right)}{2} \mathbb{E}\left[m^{h^2}\right]$, where the notation $var_{m^{h^2}}(\cdot)$ indicates a variance operator that weighs agents in proportion to the square of their attention rather than uniformly. The main reason the optimality principle takes this form is that more inattentive agents are also less affected by any given tax.

[52] When the externality/internality is heterogeneous across agents, it becomes important to measure $cov\left(\xi^h, m^h\right)$. We are not aware of empirical evidence on this quantity.

3.3 Correcting Internalities/Externalities: Relaxation of the Principle of Targeting

The classical "principle of targeting" can be stated as follows. If the consumption of a good entails an externality, the optimal policy is to tax it, and not to subsidize substitute goods or tax complement goods. For example, if fuel pollutes, then optimal policy requires taxing fuel but not taxing fuel inefficient cars or subsidizing solar panels (see Salanié (2011) for such an example). Likewise, if fatty foods are bad for consumers, and they suffer from an internality, then fatty foods should be taxed, but lean foods should not be subsidized. As we shall see, misperceptions of taxes lead to a reconsideration of this principle of targeting.

We use the specialization of the general model developed in Section 2.7. We assume that $\gamma^h = \beta^h = \lambda$, so there is no revenue-raising motive and no redistribution motive. We also assume that agents are identical except for their attention to taxes.

We consider the case with $n = 2$ taxed goods (in addition to the untaxed good 0), where the consumption of good 1 features an internality/externality so that $\boldsymbol{\tau}^X = (\xi_*, 0)$ with $\xi_* > 0$. This can be generated as follows in the model in Section 2.7. In the externality case, we simply assume that $\xi(\left(\boldsymbol{C}^h\right)_{h=1,\dots,H}) = \xi_* \frac{1}{H} \sum_h C_1^h$. In the internality case, we assume that $U^h(\boldsymbol{C}) = U^{s,h}(\boldsymbol{C}) - \xi_* C_1^h$. For example, in the externality case, good 1 could be fuel and good 2 a solar panel. In the internality example, good 1 could be fatty beef and good 2 lean turkey. In addition, we assume that the attention matrices are diagonal so that $\boldsymbol{M}^h = diag(m_1^h, m_2^h)$. Goods 1 and 2 are substitutes (respectively complements) if at all points $S_{12}^r(\boldsymbol{q}, w) > 0$ (respectively < 0).

Proposition 3.4 (Modified principle of targeting) *Suppose that the consumption of good* 1 *(but not good* 2*) entails a negative internality/externality. If agents perceive taxes correctly ($m^h = 1$ for all h), then good* 1 *should be taxed, but good* 2 *should be left untaxed—the classical principle of targeting holds. If agents' misperceptions of the tax on good* 1 *are heterogeneous (var $\left(m_1^h\right) > 0$), and if the price of good 2 is homogeneously perceived or if the misperceptions m_1^h and m_2^h of the taxes on the two goods are not too correlated (i.e. if $\mathbb{E}[m_2^h - \frac{\mathbb{E}\left[m_1^h m_2^h\right]}{\mathbb{E}\left[\left(m_1^h\right)^2\right]} m_1^h] > 0$), then good* 2 *should be subsidized (respectively taxed) if and only if goods* 1 *and* 2 *are substitutes (respectively complements).*[53]

Proposition 3.4 shows that if people have heterogeneous attention to a fuel tax, then solar panels should be subsidized (Allcott et al. (2014) derived a similar result in a different context with 0 or 1 consumption). The reason is that the tax on good 1 is an imperfect instrument in the presence of attention heterogeneity.[54] It should therefore be supplemented with a subsidy on substitute goods

[53] In particular, the conclusion of the proposition applies if agents do not misperceive the tax on good 2 but have heterogeneous misperceptions of good 1: it is then optimal to tax good 1 and to subsidize good 2. This makes clear that the key driving force is the imperfection of the tax on good 1 as a corrective instrument in the face of heterogeneous misperceptions of that tax.

[54] Heterogeneity is key for this result. Indeed, if attention to good 1 were uniform at $m_1 > 0$, the first best could be attained by taxing good 1 with tax $\tau_1 = \frac{\xi_*}{m_1}$. This ceases to be true only in the knife-edge case where $m_1 = 0$.

and a tax on complement goods. A fuel tax, for example, should be supplemented with a subsidy on solar panels and tax on fuel inefficient cars. Similarly, a fat tax should be supplemented with a subsidy on lean foods.

A similar logic applies in the traditional model with no behavioral biases, if there is an externality, and if this externality is heterogeneous across agents. Our result should therefore be interpreted as an additional and potentially important reason for why the principle of targeting might fail in the presence of behavioral biases: heterogeneous perceptions of corrective taxes.

3.4 Correcting Internalities via Taxes or Nudges with Distributive Concerns

Suppose that the poor consume "too much" sugary soda. This brings up a difficult policy trade-off. On the one hand, taxing sugary soda corrects this internality. On the other hand, taxing sugary soda redistributes away from the poor. These were the arguments regarding a recent proposal in New York City. In independent work, Lockwood and Taubinsky (2017) examine a related problem, in the context of a Mirrleesian income tax.[55]

To gain insights on how to balance these two conflicting objectives, we use the specialization of the general model developed in Section 2.7. For simplicity, we assume that good 1 is solely consumed by a class of agents, h^* but not by other agents $h \neq h^*$. As a concrete example, h^* could stand for "poor" and good 1 for "sugary soda". We also assume that utility is separable in good 1, $U^{s,h^*}(\boldsymbol{C}) = U_1^{s,h^*}(c_1) + U_2^{s,h^*}(\boldsymbol{C}_2)$, where $\boldsymbol{C}_2 = (c_i)_{i\geq 2}$ and $U^{s,h}(\boldsymbol{C}) = U_2^{s,h}(\boldsymbol{C}_2)$ for $h \neq h^*$. We assume that experienced utility for good 1 is $U_1^{h^*}(c_1) = \frac{c_1^{1-1/\psi_1}-1}{1-1/\psi_1}$ and that the internality is $U_1^{s,h^*}(c_1) - U_1^{h^*}(c_1) = \xi^{h^*} c_1$, where ξ^{h^*} is a positive constant. Taxes are correctly perceived. Applying formula (13) yields the following.

Proposition 3.5 (Taxation with both redistributive and corrective motives) *Suppose that good* 1 *is consumed only by agent* h^**, and entails an internality (captured by the internality wedge* $\tau_1^{I,h^*} = \xi^{h^*}$*). Then the optimal tax on good 1 is*

$$\tau_1 = \frac{\frac{\gamma^{h^*}}{\lambda}\xi^{h^*} + \left(1 - \frac{\gamma^{h^*}}{\lambda}\right)\frac{p_1}{\psi_1}}{1 + \left(\frac{\gamma^{h^*}}{\lambda} - 1\right)\frac{1}{\psi_1}}. \tag{29}$$

The sign of the tax τ_1 is ambiguous because there are two forces at work, corresponding to the two terms in the numerator of the right-hand side. The first term $\frac{\gamma^{h^*}}{\lambda}\xi^{h^*}$ corresponds to the internality-corrective motive of taxes and is unambiguously positive. The second term $(1 - \frac{\gamma^{h^*}}{\lambda})\frac{p_1}{\psi_1}$ corresponds to the redistributive objective of taxes, and is negative if the government wants to

[55] See also O'Donoghue and Rabin (2006) and Cremer and Pestieau (2011) for a related approach in the context of sin goods and savings, respectively, and Allcott et al. (2018) for a recent development.

redistribute towards the agent (i.e., if $\frac{\gamma^{h^*}}{\lambda} > 1$). This is because good 1 is consumed only by agent h^* and therefore taxing good 1 redistributes away from agent h^*.

Concretely, if the redistribution motive is small ($\frac{\gamma^{h^*}}{\lambda}$ close to 1), soda should be taxed. If the redistribution motive is large ($\frac{\gamma^{h^*}}{\lambda} \to \infty$) soda should taxed if and only if $\xi^{h^*} > \frac{p}{\psi_1}$, i.e. if the internality correction motive is large enough or if the demand elasticity is large enough. The former is intuitive, the latter arises because if demand is very elastic, then a given tax increase leads to a larger reduction in consumption and hence to a larger reduction in the amount of fiscal revenues extracted from the agents, thereby mitigating the associated adverse redistributive consequences.

Quantitative Illustration We now offer a simple calibration in the context of taxes on sugary sodas. It is challenging to estimate the internality coming from misunderstanding of future health costs ξ^{h^*}. One methodology is that of Lockwood and Taubinsky (2017). They estimate that the marginal health cost for an 8-oz can of soda is $C^{\$} = \$0.8$.[56] They next assume a hyperbolic $\beta - \delta$ model with short-run discount factor of $\beta = 0.7$, which translates into an internality $\xi^{h*} = (1-\beta)C^{\$} = 0.24$.

We use our formula (29). We take the cost of a can of soda to be \$1. First, if there is no redistribution motive ($\frac{\gamma^{h^*}}{\lambda} = 1$) then the optimal tax is given by the traditional Pigouvian formula $\tau_1 = \xi^{h^*} = \$0.24$, independently of the demand elasticity ψ_1. Suppose now that the government has a strong desire to redistribute towards these agents ($\frac{\gamma^{h^*}}{\lambda} = 1.5$). Then, the optimal tax depends on the demand elasticity ψ_1, over which there is considerable uncertainty. We consider three plausible values of ψ_1 : 0.25, 1, and 4. The optimal tax is then respectively $-\$0.54$, $-\$0.09$ and \$0.21.

Is it Better to Tax or to Nudge? In this environment there is a tension between the redistributive and corrective objectives of the government. Correcting for the internality of good 1 calls for a tax, but this tax redistributes revenues away from the agents of type h^* consuming the good. In this context, a nudge is attractive because it allows the government to correct the internality without increasing the tax bill of these agents. The following proposition formalizes this intuition.[57]

Proposition 3.6 (Optimal nudge vs. tax) *If* $\frac{\gamma^{h^*}}{\lambda} > 1$ *and* $\xi^{h^*} > \left(1 - \frac{\lambda}{\gamma^h}\right)\frac{p_1}{\psi_1}$, *then a nudge is better than a tax. If* $\frac{\gamma^{h^*}}{\lambda} = 1$, *a tax and a nudge are equally good and each achieve the first best. If* $\frac{\gamma^{h^*}}{\lambda} < 1$, *a tax is better than a nudge.*

[56] Building on Long et al. (2015), Lockwood and Taubinsky (2017) calculate that a 20% reduction on one-year sugar-sweetened beverages (SSBs) consumption roughly increases 0.0021 "quality adjusted life years" (QALYs) per person. They also report that the average annual consumption of SSBs is 5475 oz. From Hirth et al. (2000), a frequently used estimated value for one QALY is \$50,000. Under the estimate of \$50,000/QALY, the marginal cost of 1 oz of SSB is \$0.1.

[57] Galle (2014) provides a nuanced discussion of nudges vs. taxes.

Formula (11) shows that the optimal nudge is given by $\chi = \frac{\xi^{h^*}}{\eta}$, where η is the nudgeability of these agents.[58] This nudge is independent of the redistributive attitude of the government as captured by $\frac{\gamma^{h^*}}{\lambda}$. It perfectly corrects the internality of the agent but has no budgetary impact.

The intuition for this proposition is as follows. Suppose $\frac{\gamma^{h^*}}{\lambda} > 1$ so that the government wants to redistribute towards agents of type h^*. If the internality is strong enough so that $\xi^{h^*} > \left(1 - \frac{\lambda}{\gamma^{h^*}}\right) \frac{p_1}{\psi_1}$, then the optimal tax τ_1 is positive as shown by (29). A nudge can always be designed to achieve the same level of consumption of good 1, simply by taking $\chi = \frac{\tau}{\eta}$. Compared to the optimal tax, this nudge leaves more income to agents of type h^*, allowing them to increase their consumption of good 0, which is desirable. Because a (possibly suboptimal) nudge does better than the optimal tax, this guarantees that the optimal nudge does better than the tax. That the optimal tax does better than the optimal nudge when $\frac{\gamma^{h^*}}{\lambda} < 1$ can be proved along the same lines. In this case there is no conflict between the redistributive and corrective goals of the government, a tax helps achieve both goals while a nudge only addresses the latter.

3.5 Endogenous Attention and Salience

We now allow for endogenous attention to taxes and analyze its impact on optimal taxes. We also discuss tax salience as a policy choice in the design of the optimal tax system. We illustrate the discussion in the context of the general analysis of Section 2.

Attention as a Good To capture attention and its costs, we propose the following reinterpretation of the general framework. We imagine that we have the decomposition $\boldsymbol{c} = (\boldsymbol{C}, \boldsymbol{m})$, where $\boldsymbol{C}$ is the vector of traditional goods (champagne, leisure), and $\boldsymbol{m}$ is the vector of attention (e.g. m_i is attention to good i). We call $I^{\boldsymbol{C}}$ (respectively I^m) the set of indices corresponding to traditional goods (respectively attention). Then, all the analysis and propositions apply without modification.

This flexible modeling strategy allows to capture many potential interesting features of attention. The framework allows (but does not require) attention to be chosen and to react endogenously to incentives in a general way (optimally or not). It also allows (but does not require) attention to be produced, purchased and taxed. We find it most natural to consider the case where attention is not produced, cannot be purchased, and cannot be taxed.

It is useful to consider two benchmarks. The first benchmark is "no attention cost in welfare", where attention is endogenous (given by a function $\boldsymbol{m}(\boldsymbol{q}, w)$), but its cost is assumed not to directly affect welfare so that $u(\boldsymbol{C}, \boldsymbol{m}) = U(\boldsymbol{C})$. For instance, applying the decision vs. experienced utility framework to the example in the previous paragraph, we could have $\boldsymbol{m}(\boldsymbol{q}, w) =$

[58] Section 9.1.3 of the online appendix discusses optimal nudges to correct externalities/internalities with heterogeneous nudgeability. The optimal nudge is $\chi = \frac{\mathbb{E}\left[\xi^h \eta^h\right]}{\mathbb{E}\left[\eta^{h\,2}\right]}$, so that it is stronger when it is well-targeted, in the sense that nudgeable agents are also those with high internality/externality (higher $cov\left(\xi^h, \eta^h\right)$). It is weaker when there is more heterogeneity in nudgeability (higher $var\left[\eta^{h\,2}\right]$).

$\arg\max_{\boldsymbol{m}} u^s(\boldsymbol{C}(\boldsymbol{q},w,\boldsymbol{m}),\boldsymbol{m})$, where $u^s(\boldsymbol{C},\boldsymbol{m}) = U(\boldsymbol{C}) - g(\boldsymbol{m})$, but still $u(\boldsymbol{C},\boldsymbol{m}) = U(\boldsymbol{C})$. In that view, people use decisions heuristics that can respond to incentives, but the cost of these decision heuristics is not counted in the utility function. In this benchmark, we have $\tau_i^b = 0$ for $i \in I^m$.

The second benchmark is "attention cost in welfare". For simplicity, we outline this case under the extra assumption (which is easily relaxed) that attention is allocated optimally. We suppose that there is a primitive choice function $\boldsymbol{C}(\boldsymbol{q},w,\boldsymbol{m})$ for traditional goods that depends on attention $\boldsymbol{m}=(m_1,...,m_A)$ so that $\boldsymbol{c}(\boldsymbol{q},w,\boldsymbol{m}) = (\boldsymbol{C}(\boldsymbol{q},w,\boldsymbol{m}),\boldsymbol{m})$. Attention $\boldsymbol{m}=\boldsymbol{m}(\boldsymbol{q},w)$ is then chosen to maximize $u(\boldsymbol{C}(\boldsymbol{q},w,\boldsymbol{m}),\boldsymbol{m})$. This generates a function $\boldsymbol{c}(\boldsymbol{q},w) = (\boldsymbol{C}(\boldsymbol{q},w,\boldsymbol{m}(\boldsymbol{q},w)),\boldsymbol{m}(\boldsymbol{q},w))$. In this benchmark, attention costs are incorporated in welfare. For instance we might consider a separable utility function $u(\boldsymbol{C},\boldsymbol{m}) = U(\boldsymbol{C}) - g(\boldsymbol{m})$ for some cost function $g(\boldsymbol{m})$. A non-separable u might capture that attention is affected by consumption (e.g., of coffee) and attention affects consumption (by needing aspirin).

Illustration in the Basic Ramsey Case Section 9.2 in the online appendix gives more details for these different benchmarks and derives different versions of the corresponding optimal tax formulas. Here we simply illustrate these notions in the basic Ramsey case of Section 3.1 with just one taxed good (good 1, whose index we drop, and whose pre-tax price is p). Then, optimal attention is

$$m(\tau) = \arg\max_m u(c(p+m\tau)) - (p+\tau)c(p+m\tau) - g(m),$$

where $c(q) = q^{-\psi}$.[59] In the interest of space, we only present the results in the "no attention in welfare" case, and refer the reader to the online appendix (Section 9.2) for a treatment of the "attention cost in welfare" case.

The optimal tax formula with endogenous attention takes a form similar to formula (17), the only difference being that ψ must be replaced by $\psi(1+\tau\frac{m'(\tau)}{m(\tau)})$ to account for the increase in the elasticity of demand arising from endogenous attention.[60] We have the following.

Proposition 3.7 *Consider two economies. The first economy features endogenous attention with "no attention cost in welfare", and an optimal tax rate τ^* such that $m(\tau^*)$ and $m'(\tau^*)$ are strictly positive. The second economy has exogenous attention fixed at $m(\tau^*)$. Then the optimal tax in the*

[59] This is, attention maximizes consumption utility, minus the cost $g(m)$. Here, we choose the "ex post" allocation of attention to the tax $m(\tau)$, where system 1 (in Kahneman (2011)'s terminology—roughly, intuition) chooses attention given τ before system 2 (roughly, analytic thinking) chooses consumption given $\tau^s = m\tau$. One could alternatively choose attention "ex ante", based on the expected size of the tax (as in $m\left(\mathbb{E}\left[\tau^2\right]^{1/2}\right)$), imagining the tax as drawn from the distribution of taxes. See Gabaix (2014) for discussion of this.

[60] Indeed, demand is $D(\tau) = (q^s(\tau))^{-\psi}$ with $q^s(\tau) = p + m(\tau)\tau$, so that the quasi-elasticity of demand is:

$$-q^s(\tau)\frac{D'(\tau)}{D(\tau)} = \psi(m(\tau) + \tau m'(\tau)) = m(\tau)\psi\left(1+\tau\frac{m'(\tau)}{m(\tau)}\right).$$

second economy is higher than in the first one.

A partial intuition is that consumers are less elastic in the second economy (with fixed attention) than in the first one (with variable attention), so that the optimal tax is higher in the second economy.

Quantitative Illustration We rely again on Taubinsky and Rees-Jones (2017). They compare a standard tax regime and a high-tax regime where the tax is tripled. They find that mean attention is doubled in the high-tax regime (from 0.25 to 0.5). To match this evidence, we calibrate a locally constant elasticity of attention $\tau \frac{m'(\tau)}{m(\tau)} = \alpha$ to the tax, and find an elasticity $\alpha = \frac{\ln 2}{\ln 3} \simeq 0.6$.[61] For simplicity, we focus on the homogeneous attention case. Our theoretical results above imply that accounting for the endogeneity of attention reduces the optimal tax by a factor $1+\tau \frac{m'(\tau)}{m(\tau)} \simeq 1.6$.

Salience as a Policy Choice Governments have a variety of ways of making a particular tax more or less salient. For example, Chetty et al. (2009) present evidence that sales taxes that are included in the posted prices that consumers see when shopping have larger effects on demand. It is therefore not unreasonable to think of salience as a characteristic of the tax system that can be chosen or at least influenced by the government. This begs the natural question of the optimal salience of the tax system.[62]

We investigate this question in the context of two simple examples, the basic Ramsey and Pigou models developed in Sections 3.1 and 3.2. We start by assuming away heterogeneity in attention and introduce it only later.

We start with the basic Ramsey model. Imagine that the government can choose between two tax systems with different degrees of salience m and m' with $m'_i < m_i$ for all i, with homogeneous attention. Then it is optimal for the government to choose the lowest degree of salience because the government then raises more revenues for any given perceived tax.[63] The basic Pigou model yields a very different result. The salience of taxes is irrelevant to welfare since the first best can always be reached by adjusting taxes according to Proposition 3.2.

[61] I.e. we take $m(\tau) = \min(k\tau^{\alpha}, 1)$, which can be rationalized by an appropriate cost function $g(m)$.

[62] The optimal choice of the salience of a particular tax instrument could be analyzed using the general formalism of nudges and taxes by considering the salience of the tax as a nudge (as if m were a function of χ). However, this indirect way of proceeding is not as well suited to analyze the optimal use of different taxes with the same budgetary implications but with different salience. Therefore, we do not pursue this analogy further.

[63] The proof is very simple. Suppose that we start with the more salient tax system with attention m_i. Let τ_i be the optimal taxes and c_i be the optimal consumptions. Now consider the less salient tax system with attention $m'_i < m_i$. It is always possible to set taxes in such a way that the perceived tax is the same as at the optimum of the salient tax system by simply choosing $\tau'_i = \frac{m_i}{m'_i}\tau_i > \tau_i$. The consumption of good $i > 0$ by the agent is the same but that of good 0 is lower reflecting the fact that the government collects more revenues $\frac{m_i - m'_i}{m'_i}\tau_i c_i$. The improvement in welfare $\frac{m_i - m'_i}{m'_i}\tau_i c_i(\lambda - \gamma) > 0$ constitutes a lower bound for the welfare gains from moving to a fully optimal less salient tax system.

In discussing salience as a policy choice, we have so far maintained the assumption of homogeneous attention. Heterogeneity can alter the optimal degree of salience.[64] In the basic Ramsey model and in the limit of small taxes, optimal welfare is given by $\frac{H}{2}\sum_i \frac{\Lambda^2}{\psi_i}\frac{1}{\mathbb{E}\left[m_i^h\right]^2\left[1+\frac{var\left[m_i^h\right]}{\mathbb{E}\left[m_i^h\right]^2}\right]}y_i$ up to an additive constant (see Footnote 41). It is therefore possible for a tax system with a lower average salience $\mathbb{E}\left[m_i^{h\prime}\right]^2 < \mathbb{E}\left[m_i^h\right]^2$ to be dominated if it's associated with enough of an increase in attention heterogeneity $\frac{var\left[m_i^{h\prime}\right]}{\mathbb{E}\left[m_i^{h\prime}\right]^2} > \frac{var\left[m_i^h\right]}{\mathbb{E}\left[m_i^h\right]^2}$. The same reasoning holds for the Pigou case.[65]

4 Nonlinear Income Taxation: Mirrlees Problem

4.1 Setup

We next give a behavioral version of the celebrated Mirrlees (1971) income tax problem. To help the readers, we provide here the major building blocks and intuitions. Many details are spelled out in the online appendix (Section 10). We focus on the intensive margin of labor supply, and refer the reader to the online appendix (Section 10.3) for an analysis of the extensive margin.

Agent's Behavior There is a continuum of agents indexed by skill n with density $f(n)$ (we use n, the conventional index in that literature, rather than h). Agent n has a utility function $u^n(c,z)$, where c is his one-dimensional consumption, z is his pre-tax income, and $u_z \leq 0$.[66] The total income tax for income z is $T(z)$, so that disposable income is $R(z) = z - T(z)$.

We call $g(z)$ the social marginal welfare weight (the counterpart of β^h in section 2.2) and $\gamma(z)$ the social marginal utility of income (the counterpart of γ^h). Just like in the Ramsey model, we define the "behavioral wedge" $\tau^b(z) = -\frac{(1-T'(z))u_c(c,z)+u_z(c,z)}{v_w}$, where v_w is the marginal utility of a dollar received lump-sum.[67] If the agent works too much—perhaps because he underperceives taxes (see Feldman et al. (2016) for recent evidence on confusion about marginal tax rates) or overperceives the benefits of working—then τ^b is positive. We also define the renormalized behavioral wedge $\widetilde{\tau}^b(z) = g(z)\tau^b(z)$.

[64] One might expect the heterogeneity of attention to be an inverted U-shaped function of average attention, as it should be 0 in the fully attentive and fully inattentive cases.

[65] It could also be interesting to allow the government to combine different tax instruments with the same tax base but different degrees of salience. Our general model could be extended to allow for this possibility. We would start with a function $c\left(w,\boldsymbol{p},\boldsymbol{\tau}^1,\boldsymbol{\tau}^2,...,\boldsymbol{\tau}^K\right)$, where $\boldsymbol{\tau}^\kappa$ are tax vectors with different degrees of salience. Each tax instrument κ corresponds to a Slutsky matrix $S_{ij}^{C,\kappa}$ which depends on the tax instrument indexed by κ. In optimal tax formula (8), the term $(\boldsymbol{\tau}-\widetilde{\boldsymbol{\tau}}^{b,h})\cdot\boldsymbol{S}_i^{C,h}$ is then replaced by $(\bar{\boldsymbol{\tau}}-\widetilde{\boldsymbol{\tau}}^{b,h})\cdot\boldsymbol{S}_i^{C,\kappa,h}$ with $\bar{\boldsymbol{\tau}}=\sum_{\kappa=1}^K\boldsymbol{\tau}^\kappa$. The intuition is that the different tax instruments lead to different substitution effects captured by different Slutsky matrices $S_{ij}^{C,\kappa}$. As an extreme example, differential salience could replicate lump-sum taxes (see Goldin (2015) and Section 9.3.2 of the online appendix).

[66] If the agent's pre-tax wage is n, L is his labor supply, and utility is $U(c,L)$, then $u^n(c,z) = U\left(c,\frac{z}{n}\right)$. Note that this assumes that the wage is constant (normalized to one).

[67] Formally, this is $(1-T'(z),1)\cdot\boldsymbol{\tau}^b$, where $\boldsymbol{\tau}^b$ is the vector behavioral wedge defined earlier.

Planning Problem The objective of the planner is to design the tax schedule $T(z)$ in order to maximize the following objective function: $\int_0^\infty W(v(n)) f(n) dn + \int_0^\infty (z(n) - c(n)) f(n) dn$, where $v(n)$ is the utility attained by agent of type n.

Traditional and Behavioral Elasticity Concepts We call ζ^c the compensated elasticity of labor supply—a traditional elasticity concept. We also define a new elasticity concept, which we shall call "behavioral cross-influence" and denote by $\zeta^c_{Q_{z^*}}(z)$: it is the elasticity of the earnings of an agent at earnings z to the marginal retention rate $(1 - T'(z^*))$ at income $z^* \neq z$. In the traditional model with no behavioral biases, $\zeta^c_{Q_{z^*}}(z) = 0$. But this is no longer true with behavioral agents.[68] For instance, in Liebman and Zeckhauser (2004), people mistake average tax rates for marginal tax rates, so inframarginal rates (at $z^* < z$) affect labor supply, and $\zeta^c_{Q_{z^*}}(z) > 0$.

Following Saez (2001), we call $h(z)$ the density of agents with earnings z at the optimum and $H(z) = \int_0^z h(z') dz'$. We also introduce the virtual density $h^*(z) = \frac{q(z)}{1-T'(z)+\zeta^c z T''(z)} h(z)$.

4.2 Optimal Income Tax Formula

We next present the optimal income tax formula.

Proposition 4.1 *Optimal taxes satisfy the following formulas (for all z^*)*[69]

$$\frac{T'(z^*) - \widetilde{\tau}^b(z^*)}{1 - T'(z^*)} = \frac{1}{\zeta^c(z^*)} \frac{1 - H(z^*)}{z^* h^*(z^*)} \int_{z^*}^\infty (1 - \gamma(z)) \frac{h(z)}{1 - H(z^*)} dz - \int_0^\infty \frac{\zeta^c_{Q_{z^*}}(z)}{\zeta^c(z^*)} \frac{T'(z) - \tilde{\tau}^b(z)}{1 - T'(z)} \frac{z h^*(z)}{z^* h^*(z^*)} dz. \tag{30}$$

The first term $\frac{1}{\zeta^c(z^*)} \frac{1-H(z^*)}{z^* h^*(z^*)} \int_{z^*}^\infty (1 - \gamma(z)) \frac{h(z)}{1-H(z^*)} dz$ on the right-hand side of the optimal tax formula (30) is a simple reformulation of Saez's formula. The second term $-\frac{1}{z^*} \int_0^\infty \frac{\zeta^c_{Q_{z^*}}(z)}{\zeta^c(z^*)} \frac{T'(z) - \tilde{\tau}^b(z)}{1-T'(z)} z \frac{h^*(z)}{h^*(z^*)} dz$ on the right-hand side is new and, together with the term $\frac{-\widetilde{\tau}^b(z^*)}{1-T'(z^*)}$ on the left-hand side, captures misoptimization effects.

The intuition is as follows. First, suppose that $\zeta^c_{Q_{z^*}}(z) > 0$. Then increasing the marginal tax rate at z^* leads the agents at another income z to perceive higher taxes on average, which leads them to decrease their labor supply and reduces tax revenues. *Ceteris paribus*, this consideration pushes towards a lower tax rate (hence the minus sign in front of the last integral in (30)), compared to the Saez optimal tax formula. Second, suppose that $\tilde{\tau}^b(z) < 0$ (perhaps because the agent

[68] Hence, normatively irrelevant tax rates may affect choices, a bit like in the behavioral literature on menu and decoy effects (e.g., Kamenica (2008), Bordalo et al. (2013), Bushong et al. (2016)).

[69] This formula can also be expressed as a modification of the Saez (2001) formula. The modified Saez formula (see equation (88) in the online appendix) uses the concept of the social marginal welfare weight $g(z)$ rather than the social marginal utility of income $\gamma(z)$.

underperceives the benefits of working), then increasing the marginal tax rate at z^* further reduces welfare. This, again, pushes towards a lower tax rate.

The formula is expressed in terms of endogenous objects or "sufficient statistics": social marginal welfare weights $g(z)$, elasticities of substitution $\zeta^c(z)$, income elasticities $\eta(z)$, and income distribution $h(z)$ and $h^*(z)$. With behavioral agents, there are two additional sufficient statistics, namely the behavioral wedge $\tilde{\tau}^b(z)$ and the behavioral cross-elasticities $\zeta^c_{Q_{z^*}}(z)$.

4.3 Implications

This formula has a number of consequences. We highlight two of them, at the bottom and the top of the income tax schedule.

Possibility of Negative Marginal Income Tax Rate and EITC In the traditional model with no behavioral biases, negative marginal income tax rates can never arise at the optimum. To see this, consider an example using the decision vs. experienced utility model. Let decision utility u^s be quasilinear so that there are no income effects $u^s(c,z) = c - \phi(z)$. We take experienced utility to be $u(c,z) = \theta c - \phi(z)$. Then $\tilde{\tau}^b(z) = -g(z)\phi'(z)\frac{\theta-1}{\theta}$, $\gamma = g$, and $\zeta^c_{Q_{z^*}} = 0$. When $\theta > 1$, we have $\widetilde{\tau}^b(z^*) < 0$, and it is possible for this formula to yield $T'(z^*) < 0$. This occurs if agents undervalue the benefits or overvalue the costs from higher labor supply. For example, it could be the case that working more leads to higher human capital accumulation and higher future wages, but that these benefits are underperceived by agents, which could be captured in reduced form by $\theta > 1$. Such biases could be particularly relevant at the bottom of the income distribution (see Chetty et al. (2013) for a review of the evidence). If these biases are strong enough, the modified Saez formula could predict negative marginal income tax rates at the bottom of the income distribution. This could provide a behavioral rationale for the EITC (Earned Income Tax Credit) program. In parallel and independent work, Gerritsen (2016) and Lockwood (2017) derive a modified Saez formula in the context of decision vs. experienced utility model. Lockwood (2017) provides an empirical analysis documenting significant present-bias among EITC recipients, showing that a calibrated version of the model goes a long way towards rationalizing the negative marginal tax rates associated with the EITC program.[70]

Taxes at the Top of the Income Distribution We start by revisiting the classic result that if the income distribution is bounded at $z_{\max}$, then the top marginal income tax rate should be zero. In our model, this need not be the case. One simple way to see that is to consider the case of decision vs. experienced utility. Tax formula (30) then prescribes $T'(z_{\max}) = \tilde{\tau}^b(z_{\max})$ which is

[70] This differs from alternative rationales for negative marginal income tax rates that have been put forth in the traditional literature. For example, Saez (2002) shows that if the Mirrlees model is extended to allow for an extensive margin of labor supply, then negative marginal income tax rates can arise at the optimum. We refer the reader to the online appendix (section 10.3) for a behavioral treatment of the Saez (2002) extensive margin of labor supply model.

positive or negative depending on whether top earners over or under perceive the benefits of work (under or over perceive the costs of work).

The online appendix (Section 10.2.3) also derives a formula for the marginal rate at very high incomes when the income distribution is unbounded at the top (Proposition 10.2). Compared to the traditional model, the optimal tax is higher if people at the top overestimate the benefits of earning more money (as reflected in the $\tilde{\tau}^b$ term). It is also lower if the top marginal tax rate is particularly salient to all agents, and affects not only top earners, but also the tax perceived by agents at all points of the income distribution (as reflected in the $\zeta^c_{Q_{z^*}}$ term). It would be interesting to measure the size of these effects.[71]

5 Conclusion

We have generalized the main results of the traditional theory of optimal taxation to allow for a large class of behavioral biases. To do so, we have generalized some of the classical objects and relations of price theory when agents are behavioral. We have also proposed specializations of the model to show how some classical optimal taxation results are modified in the presence of specific behavioral frictions. They were in part summarized in points 1-5 of the introduction.

In the NBER working paper version of this paper we derive additional results. In particular, we show that the conditions for productive efficiency in Diamond and Mirrlees (1971) leading to the suboptimality of taxes on intermediate goods become more stringent: they require a full set of completely salient taxes perceived as prices, instead of simply a full set of taxes (constant returns or fully taxed profits are also required in both cases). The same conditions are required for the associated result that optimal tax formulas are independent of production elasticities. We generalize our optimal tax formulas to the case where these assumptions are not verified. We also show that the separability conditions required for the uniform commodity taxation result in Atkinson and Stiglitz (1976) are no longer sufficient in the presence of behavioral biases. Finally we present a modest attempt at modelling mental accounts with an application to vouchers. We plan to collect these results in a separate paper.

Natural extensions of our approach would be to consider behavioral biases that cannot be captured by our model, such as interpersonal behavioral biases, or to relax the focus on a benevolent and well-informed government. We leave these issues to future work.

References

Abaluck, Jason and Jonathan Gruber, "Heterogeneity in Choice Inconsistencies Among the Elderly: Evidence from Prescription Drug Plan Choice," *The American Economic Review Papers*

[71] Concretely, think of the recent case of France where increasing the top rate to 75% might have created an adverse general climate with the perception that even earners below the top income bracket would pay higher taxes.

and Proceedings, 2011, *101* (3), 377–381.

Aguiar, Victor H. and Roberto Serrano, "Slutsky matrix norms: The size, classification, and comparative statics of bounded rationality," *Journal of Economic Theory*, 2017, *172*, 163 – 201.

Allcott, Hunt and Dmitry Taubinsky, "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market," *The American Economic Review*, 2015, *105* (8), 2501–2538.

_ **and Nathan Wozny**, "Gasoline Prices, Fuel Economy, and the Energy Paradox," *Review of Economics and Statistics*, 2014, *96* (5), 779–795.

_ **, Benjamin Lockwood, and Dmitry Taubinsky**, "Ramsey Strikes Back: Optimal Commodity Tax and Redistribution in the Presence of Salience Effects," *AEA Papers and Proceedings*, 2018, *108*, 88–92.

_ **, Christopher Knittel, and Dmitry Taubinsky**, "Tagging and targeting of energy efficiency subsidies," *The American Economic Review*, 2015, *105* (5), 187–191.

_ **, Sendhil Mullainathan, and Dmitry Taubinsky**, "Energy Policy with Externalities and Internalities," *Journal of Public Economics*, 2014, *112*, 72–88.

Anagol, Santosh and Hugh Hoikwang Kim, "The Impact of Shrouded Fees: Evidence from a Natural Experiment in the Indian Mutual Funds Market," *The American Economic Review*, 2012, *102* (1), 576–593.

Atkinson, Anthony B. and Joseph E. Stiglitz, "The Structure of Indirect Taxation and Economic Efficiency," *Journal of Public Economics*, 1972, *1* (1), 97–119.

_ **and** _ , "The Design of Tax Structure: Direct versus Indirect Taxation," *Journal of Public Economics*, 1976, *6* (1-2), 55–75.

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein, "Behavioral Hazard in Health Insurance," *Quarterly Journal of Economics*, 2015, *130* (4), 1623–1667.

Bénabou, Roland and Efe A. Ok, "Social Mobility and the Demand for Redistribution : The POUM Hypothesis," *Quarterly Journal of Economics*, 2001, *116* (2), 447–487.

_ **and Jean Tirole**, "Incentives and Prosocial Behavior," *The American Economic Review*, 2006, *96* (5), 1652–1678.

Bernheim, B. Douglas and Antonio Rangel, "Beyond Revealed Preference: Choice Theoretic Foundations for Behavioral Welfare Economics," *Quarterly Journal of Economics*, 2009, *124* (1), 51–104.

Beshears, John, James J. Choi, David Laibson, Brigitte C Madrian, and Sean Yixiang Wang, "Who is Easier to Nudge?," *Working Paper*, 2016.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Salience and Consumer Choice," *Journal of Political Economy*, 2013, *121* (5), 803–843.

Brown, Jennifer, Tanjim Hossain, and John Morgan, "Shrouded Attributes and Information Suppression: Evidence from the Field," *Quarterly Journal of Economics*, 2010, *125* (2), 859–876.

Bushong, Benjamin, Matthew Rabin, and Joshua Schwartzstein, "A Model of Relative Thinking," *Working Paper*, 2016.

Busse, Meghan R., Christopher R. Knittel, and Florian Zettelmeyer, "Are Consumers Myopic? Evidence from New and Used Car Purchases," *The American Economic Review*, 2013, *103* (1), 220–256.

Caplin, Andrew and Mark Dean, "Revealed Preference, Rational Inattention, and Costly Information Acquisition," *The American Economic Review*, 2015, *105* (7), 2183–2203.

Carroll, Gabriel D., James J. Choi, David Laibson, Brigitte C. Madrian, and Andrew Metrick, "Optimal Defaults and Active Decisions," *Quarterly Journal of Economics*, 2009, *124* (4), 1639–1674.

Chetty, Raj, "The Simple Economics of Salience and Taxation," *NBER Working Paper No. 15246*, 2009.

_ , "Behavioral Economics and Public Policy: A Pragmatic Perspective," *The American Economic Review*, 2015, *105* (5), 1–33.

_ **, Adam Looney, and Kory Kroft**, "Salience and Taxation: Theory and Evidence," *The American Economic Review*, 2009, *99* (4), 1145–1177.

_ **and Emmanuel Saez**, "Teaching the Tax Code: Earnings Responses to an Experiment with EITC Recipients," *American Economic Journal: Applied Economics*, 2013, *5* (1), 1–31.

_ **, John N. Friedman, and Emmanuel Saez**, "Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings," *The American Economic Review*, 2013, *103* (7), 2683–2721.

_ **,** _ **, Soeren Leth-Petersen, Torben Heien Nielsen, and Tore Olsen**, "Active vs. Passive Decisions and Crowd-out in Retirement Savings Accounts: Evidence from Denmark," *Quarterly Journal of Economics*, 2014, *129* (3), 1141–01219.

Cremer, Helmuth and Pierre Pestieau, "Myopia, Redistribution and Pensions," *European Economic Review*, 2011, *55* (2), 165–175.

Dávila, Eduardo, "Optimal Financial Transaction Taxes," *Working Paper*, 2017.

De Bartolomé, Charles A. M., "Which Tax Rate do People Use: Average or Marginal?," *Journal of Public Economics*, 1995, *56* (1), 79–96.

DellaVigna, Stefano, "Psychology and Economics: Evidence from the Field," *Journal of Economic Literature*, 2009, *47* (2), 315–372.

Diamond, Peter A., "A Many-person Ramsey Tax Rule," *Journal of Public Economics*, 1975, *4* (4), 335–342.

_ **and James A. Mirrlees**, "Optimal Taxation and Public Production I: Production Efficiency," *The American Economic Review*, 1971, *61* (1), 8–27.

Ellison, Glenn and Sara Fisher Ellison, "Search, Obfuscation, and Price Elasticities on the Internet," *Econometrica*, 2009, *77* (2), 427–452.

Feldman, Naomi E., Peter Katuscak, and Laura Kawano, "Taxpayer Confusion: Evidence from the Child Tax Credit," *The American Economic Review*, 2016, *106* (3), 807–835.

Finkelstein, Amy, "E-ztax: Tax Salience and Tax Rates," *Quarterly Journal of Economics*, 2009, *124* (3), 969–1010.

Gabaix, Xavier, "A Sparsity-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 2014, *129* (4), 1661–1710.

_ , "Behavioral Inattention," *NBER Working Paper No. 24096*, 2017.

_ , "Behavioral macroeconomics via sparse dynamic programming," *Working Paper*, 2017.

_ **and David Laibson**, "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets," *Quarterly Journal of Economics*, 2006, *121* (2), 505–540.

Galle, Brian, "Tax, Command or Nudge: Evaluating the New Regulation," *Texas Law Review*, 2014, *92*, 837.

Gerritsen, Aart, "Optimal Taxation when People do not Maximize Well-Being," *Journal of Public Economics*, 2016, *144*, 122–139.

Glaeser, Edward L., "Paternalism and Psychology," *University of Chicago Law Review*, 2006, *73* (1), 133–156.

Goldin, Jacob, "Optimal Tax Salience," *Journal of Public Economics*, 2015, *131*, 115–123.

Gruber, Jonathan and Botond Kőszegi, "Is Addiction "Rational"? Theory and Evidence," *Quarterly Journal of Economics*, 2001, *116* (4), 1261–1303.

_ **and** _ , "Tax incidence when individuals are time-inconsistent: the case of cigarette excise taxes," *Journal of Public Economics*, 2004, *88* (9), 1959–1987.

Hastings, Justine S. and Jesse M. Shapiro, "Fungibility and Consumer Choice: Evidence from Commodity Price Shocks," *Quarterly Journal of Economics*, 2013, *128* (4), 1449–1498.

Hastings, Justine S and Jesse M Shapiro, "How Are SNAP Benefits Spent? Evidence from a Retail Panel," *Forthcoming at the American Economic Review*, 2018.

Hirth, Richard A, Michael E Chernew, Edward Miller, A Mark Fendrick, and William G Weissert, "Willingness to pay for a quality-adjusted life year: in search of a standard," *Medical Decision Making*, 2000, *20* (3), 332–342.

Kahneman, Daniel, *Thinking, fast and slow*, Macmillan, 2011.

_ **, Peter P Wakker, and Rakesh Sarin**, "Back to Bentham? Explorations of Experienced Utility," *Quarterly Journal of Economics*, 1997, *112* (2), 375–406.

Kamenica, Emir, "Contextual Inference in Markets: On the Informational Content of Product Lines," *The American Economic Review*, 2008, *98* (5), 2127–2149.

Kaplow, Louis, "Myopia and the Effects of Social Security and Capital Taxation on Labor Supply," *National Tax Journal*, 2015, *68* (1), 7–32.

Khaw, Mel Win, Ziang Li, and Michael Woodford, "Risk Aversion as a Perceptual Bias,"

NBER Working Paper No. 23294, 2017.

Kőszegi, Botond and Adam Szeidl, "A Model of Focusing in Economic Choice," *Quarterly Journal of Economics*, 2013, *128* (1), 53–104.

Laibson, David, "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 1997, *112* (2), 443–478.

Liebman, Jeffrey B. and Richard J. Zeckhauser, "Schmeduling," *Working Paper*, 2004.

Lockwood, Benjamin, "Optimal Income Taxation with Present Bias," *Working Paper*, 2017.

_ **and Dmitry Taubinsky**, "Regressive Sin Taxes," *NBER Working Paper No. 23085*, 2017.

Loewenstein, George and Ted O'Donoghue, ""We Can Do This the Easy Way or the Hard Way": Negative Emotions, Self-Regulation, and the Law," *University of Chicago Law Review*, 2006, *73* (1), 183–206.

Long, Michael W, Steven L Gortmaker, Zachary J Ward, Stephen C Resch, Marj L Moodie, Gary Sacks, Boyd A Swinburn, Rob C Carter, and Y Claire Wang, "Cost effectiveness of a sugar-sweetened beverage excise tax in the US," *American journal of preventive medicine*, 2015, *49* (1), 112–123.

Mani, Anandi, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao, "Poverty Impedes Cognitive Function.," *Science*, 2013, *341* (6149), 976–80.

Mirrlees, James A., "An Exploration in the Theory of Optimum Income Taxation," *Review of Economic Studies*, 1971, *38* (2), 175–208.

Moser, Christian and Pedro Olea de Souza e Silva, "Optimal Paternalistic Savings Policies," *Working Paper*, 2017.

Mullainathan, Sendhil, Joshua Schwartzstein, and William J. Congdon, "A Reduced-Form Approach to Behavioral Public Finance," *Annual Review of Economics*, 2012, *4* (1), 511–540.

O'Donoghue, Ted and Matthew Rabin, "Optimal Sin Taxes," *Journal of Public Economics*, 2006, *90* (10-11), 1825–1849.

Pigou, Arthur, *The Economics of Welfare*, Macmillan and Company, 1920.

Piketty, Thomas and Emmanuel Saez, "Income Inequality in the United States, 1913-1998," *Quarterly Journal of Economics*, 2003, *118* (1), 1–39.

Ramsey, Frank P., "A Contribution to the Theory of Taxation," *Economic Journal*, 1927, *37* (145), 47–61.

Saez, Emmanuel, "Using Elasticities to Derive Optimal Income Tax Rates," *Review of Economic Studies*, 2001, *68* (1), 205–229.

_ , "Optimal Income Transfer Programs: Intensive versus Extensive Labor Supply Responses," *Quarterly Journal of Economics*, 2002, *117* (3), 1039–1073.

Salanié, Bernard, *The Economics of Taxation*, MIT press, 2011.

Sandmo, Agnar, "Optimal Taxation in the Presence of Externalities," *Swedish Journal of Economics*, 1975, *77* (1), 86–98.

Schwartzstein, Joshua, "Selective Attention and Learning," *Journal of the European Economic Association*, 2014, *12* (6), 1423–1452.

Sims, Christopher A., "Implications of Rational Inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690.

Slemrod, Joel, "The Role of Misconceptions in Support for Regressive Tax Reform," *National Tax Journal*, 2006, *59* (1), 57–75.

Spinnewijn, Johannes, "Unemployed but Optimistic: Optimal Insurance Design with Biased Beliefs," *Journal of the European Economic Association*, 2015, *13* (1), 130–167.

Taubinsky, Dmitry and Alex Rees-Jones, "Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment," *Forthcoming at the Review of Economic Studies*, 2017.

Thaler, Richard, "Mental Accounting and Consumer Choice," *Marketing Science*, 1985, *4* (3), 199–214.

Thaler, Richard H. and Cass R. Sunstein, *Nudge*, Yale University Press, 2008.

Tversky, Amos and Daniel Kahneman, "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty*, 1992, *5* (4), 297–323.

Weitzman, Martin L., "Prices vs. Quantities," *Review of Economic Studies*, 1974, *41* (4), 477–491.

6 Appendix: Notations

Vectors and matrices are represented by bold symbols (e.g. $\boldsymbol{c}$).

$\boldsymbol{c}$: consumption vector

h: index for household type h

L: government's objective function.

$\boldsymbol{m}, \boldsymbol{M}$: attention vector, matrix

$\boldsymbol{p}$: pre-tax price

$\boldsymbol{q} = \boldsymbol{p} + \boldsymbol{\tau}$: after-tax price

$\boldsymbol{q}^s$: $s$ubjectively perceived after-tax price

$\boldsymbol{S}_j$: Column of the Slutsky matrix when price j changes.

$u(\boldsymbol{c})$: experienced utility

$u^s(\boldsymbol{c})$: $s$ubjectively perceived utility

$v(\boldsymbol{q}, w)$: experienced indirect utility

$v^s(\boldsymbol{q}, w)$: $s$ubjectively perceived indirect utility

w: personal income

W: social utility

γ^h (resp. $\gamma^{\xi,h}$): marginal social utility of income (resp. adjusted for externalities)

η^h: nudgeability of agents of type h

λ: weight on revenue raised in planner's objective

ψ_i: demand elasticity for good i

$\boldsymbol{\tau}$: tax

$\boldsymbol{\tau}^b$: $b$ehavioral wedge

$\boldsymbol{\tau}^s$: $s$ubjectively perceived tax

ξ: externality

χ: intensity of the nudge

7 Appendix: Behavioral Consumer Price Theory

This section expands on the sketch given in Section 2.1. Here we develop behavioral consumer price theory with a nonlinear budget. This nonlinear budget is useful both for conceptual clarity and for the study of Mirrleesian nonlinear taxation. The agent faces a budget constraint $B(\boldsymbol{c},\boldsymbol{q}) \leq w$. When the budget constraint is linear, $B(\boldsymbol{c},\boldsymbol{q}) = \boldsymbol{q}\cdot\boldsymbol{c}$, so that $B_{q_j} = c_j, B_{c_j} = q_j$.

The agent, whose utility is $u(\boldsymbol{c})$, may not completely maximize. Instead, his policy is described by $\boldsymbol{c}(\boldsymbol{q},w)$, which exhausts his budget $B(\boldsymbol{c}(\boldsymbol{q},w),\boldsymbol{q}) = w$. Though this puts very little structure on the problem, some basic relations can be derived, as follows.

7.1 Abstract General Framework

The indirect utility is defined as $v(\boldsymbol{q},w) = u(\boldsymbol{c}(\boldsymbol{q},w))$ and the expenditure function as $e(\boldsymbol{q},\widehat{u}) = \min_w w$ s.t. $v(\boldsymbol{q},w) \geq \widehat{u}$. This implies $v(\boldsymbol{q},e(\boldsymbol{q},\widehat{u})) = \widehat{u}$ (with $\widehat{u}$ a real number). Differentiating with respect to q_j, this implies

$$\frac{v_{q_j}(\boldsymbol{q},w)}{v_w(\boldsymbol{q},w)} = -e_{q_j}. \tag{31}$$

We call $\boldsymbol{S}^C(\boldsymbol{q},w)$ the "income-compensated" Slutsky matrix, whose row j (corresponding to the consumption response to a compensated change in the price q_j) is defined to be:

$$\boldsymbol{S}_j^C(\boldsymbol{q},w) = \boldsymbol{c}_{q_j}(\boldsymbol{q},w) + \boldsymbol{c}_w(\boldsymbol{q},w)\, B_{q_j}(\boldsymbol{c},\boldsymbol{q})_{|\boldsymbol{c}=\boldsymbol{c}(\boldsymbol{q},w)}. \tag{32}$$

The Hicksian demand is: $\boldsymbol{h}(\boldsymbol{q},\widehat{u}) = \boldsymbol{c}(\boldsymbol{q},e(\boldsymbol{q},\widehat{u}))$, and the Hicksian-demand based Slutsky matrix is defined as: $\boldsymbol{S}_j^H(\boldsymbol{q},\widehat{u}) = \boldsymbol{h}_{q_j}(\boldsymbol{q},\widehat{u})$.

The Slutsky matrices represent how demand changes when prices change by a small amount, and the budget is compensated to make the previous basket or the previous utility level available: $\boldsymbol{S}^C(\boldsymbol{q},w) = \partial_{\boldsymbol{x}}\boldsymbol{c}(\boldsymbol{q}+\boldsymbol{x}, B(\boldsymbol{c}(\boldsymbol{q},w),\boldsymbol{q}+\boldsymbol{x}))_{|\boldsymbol{x}=0}$ and $\boldsymbol{S}^H(\boldsymbol{q},w) = \partial_{\boldsymbol{x}}\boldsymbol{c}(\boldsymbol{q}+\boldsymbol{x}, e(\boldsymbol{q}+\boldsymbol{x}, v(\boldsymbol{q},w)))_{|\boldsymbol{x}=0}$,

i.e., using (31),

$$\boldsymbol{S}_j^H(\boldsymbol{q},w)=\boldsymbol{c}_{q_j}(\boldsymbol{q},w)-\boldsymbol{c}_w(\boldsymbol{q},w)\frac{v_{q_j}(\boldsymbol{q},w)}{v_w(\boldsymbol{q},w)}. \tag{33}$$

In the traditional model, $\boldsymbol{S}^C=\boldsymbol{S}^H$, but we shall see that this won't be the case in general.[72]

We have the following elementary facts (with $\boldsymbol{c}(\boldsymbol{q},w),v(\boldsymbol{q},w)$ unless otherwise noted).

$$B_{\boldsymbol{c}}\cdot\boldsymbol{c}_w=1,\qquad B_{\boldsymbol{c}}\cdot\boldsymbol{c}_{q_i}=-B_{q_i},\qquad u_{\boldsymbol{c}}\cdot\boldsymbol{c}_w=v_w. \tag{34}$$

The first two come from differentiating $B(\boldsymbol{c}(\boldsymbol{q},w),\boldsymbol{q})=w$. The third one comes from differentiating $v(\boldsymbol{q},w)=u(\boldsymbol{c}(\boldsymbol{q},w))$ with respect to w.

Proposition 7.1 (Behavioral Roy's identity) *We have*

$$\frac{v_{q_j}(\boldsymbol{q},w)}{v_w(\boldsymbol{q},w)}=-B_{q_j}(\boldsymbol{c}(\boldsymbol{q},w),\boldsymbol{q})+D_j(\boldsymbol{q},w), \tag{35}$$

where

$$D_j(\boldsymbol{q},w)=-\boldsymbol{\tau}^b(\boldsymbol{q},w)\cdot\boldsymbol{c}_{q_j}(\boldsymbol{q},w)=-\boldsymbol{\tau}^b\cdot\boldsymbol{S}_j^H=-\boldsymbol{\tau}^b\cdot\boldsymbol{S}_j^C, \tag{36}$$

and the behavioral wedge is defined to be

$$\boldsymbol{\tau}^b(\boldsymbol{q},w)=B_{\boldsymbol{c}}(\boldsymbol{c}(\boldsymbol{q},w),\boldsymbol{q})-\frac{u_{\boldsymbol{c}}(\boldsymbol{c}(\boldsymbol{q},w))}{v_w(\boldsymbol{q},w)}. \tag{37}$$

When the agent is the traditional rational agent, $\boldsymbol{\tau}^b=0$. In general, $\boldsymbol{\tau}^b\cdot\boldsymbol{c}_w(\boldsymbol{q},w)=0$.

Proof. Relations (34) imply: $\boldsymbol{\tau}^b\cdot\boldsymbol{c}_w=\left(B_{\boldsymbol{c}}-\frac{u_{\boldsymbol{c}}}{v_w}\right)\cdot\boldsymbol{c}_w=1-1=0$. Next, we differentiate $v(\boldsymbol{q},w)=u(\boldsymbol{c}(\boldsymbol{q},w))$

$$\frac{v_{q_i}}{v_w}=\frac{u_{\boldsymbol{c}}\cdot\boldsymbol{c}_{q_i}}{v_w}=\frac{(u_{\boldsymbol{c}}-v_wB_{\boldsymbol{c}}+v_wB_{\boldsymbol{c}})\cdot\boldsymbol{c}_{q_i}}{v_w}=\frac{(u_{\boldsymbol{c}}-v_wB_{\boldsymbol{c}})\cdot\boldsymbol{c}_{q_i}}{v_w}-B_{q_i}\text{ as }B_{\boldsymbol{c}}\cdot\boldsymbol{c}_{q_i}=-B_{q_i}\text{ from (34)}$$

$$=-\boldsymbol{\tau}^b\cdot\boldsymbol{c}_{q_i}-B_{q_i}. \tag{38}$$

Next,

$$\begin{aligned}D_j&=-\boldsymbol{\tau}^b\cdot\boldsymbol{c}_{q_j}=-\boldsymbol{\tau}^b\cdot\left(\boldsymbol{S}_j^H+\boldsymbol{c}_w(\boldsymbol{p},w)\frac{v_{q_j}(\boldsymbol{p},w)}{v_w(\boldsymbol{p},w)}\right)\text{ by (33)}\\&=-\boldsymbol{\tau}^b\cdot\boldsymbol{S}_j^H\text{ as }\boldsymbol{\tau}^b\cdot\boldsymbol{c}_w=0.\end{aligned} \tag{39}$$

Likewise, (32) gives, using again $\boldsymbol{\tau}^b\cdot\boldsymbol{c}_w=0$: $D_j=-\boldsymbol{\tau}^b\cdot\boldsymbol{c}_{q_j}=-\boldsymbol{\tau}^b\cdot\left(\boldsymbol{S}_j^C-\boldsymbol{c}_wB_{q_j}\right)=-\boldsymbol{\tau}^b\cdot\boldsymbol{S}_j^C$.□

[72] See Aguiar and Serrano (2017) for a recent study of Slutsky matrices with behavioral models.

Proposition 7.2 *(Slutsky relation modified) With* $\boldsymbol{c}(\boldsymbol{q}, w)$ *we have*

$$\boldsymbol{c}_{q_j}(\boldsymbol{q}, w) = -\boldsymbol{c}_w B_{q_j} + \boldsymbol{S}_j^H + \boldsymbol{c}_w D_j = -\boldsymbol{c}_w B_{q_j} - \boldsymbol{c}_w\left(\boldsymbol{\tau}^b \cdot \boldsymbol{S}_j^H\right) + \boldsymbol{S}_j^H = -\boldsymbol{c}_w B_{q_j} + \boldsymbol{S}_j^C,$$

and

$$\boldsymbol{S}_j^C - \boldsymbol{S}_j^H = \boldsymbol{c}_w D_j = -\boldsymbol{c}_w\left(\boldsymbol{\tau}^b \cdot \boldsymbol{S}_j^H\right). \tag{40}$$

Proof.

$$\begin{aligned}\boldsymbol{c}_{q_j} &= \boldsymbol{c}_w \frac{v_{q_j}(\boldsymbol{q}, w)}{v_w(\boldsymbol{q}, w)} + \boldsymbol{S}_j^H \text{ by (33)} \\ &= \boldsymbol{c}_w\left(-B_{q_j} + D_j\right) + \boldsymbol{S}_j^H \text{by Proposition 7.1.}\end{aligned}$$

Also, (32) gives: $\boldsymbol{c}_{q_j} = -\boldsymbol{c}_w B_{q_j} + \boldsymbol{S}_j^C$.□

Lemma 7.1 *We have*

$$B_{\boldsymbol{c}} \cdot \boldsymbol{S}_j^C = 0, \qquad B_{\boldsymbol{c}} \cdot \boldsymbol{S}_j^H = -D_j. \tag{41}$$

Proof. Relations (34) imply $B_{\boldsymbol{c}} \cdot \boldsymbol{S}_j^C = B_{\boldsymbol{c}} \cdot \left(\boldsymbol{c}_{q_j} + \boldsymbol{c}_w B_{q_j}\right) = -B_{q_j} + B_{q_j} = 0$. Also, $B_{\boldsymbol{c}} \cdot \boldsymbol{S}_j^H = B_{\boldsymbol{c}} \cdot \left(\boldsymbol{S}_j^C - \boldsymbol{c}_w D_j\right) = -D_j$. □

7.2 Application in Specific Behavioral Models

Decision vs. Experienced Utility Model In the decision-utility model there is an experience utility function $u(\boldsymbol{c})$, and a perceived utility function $u^s(\boldsymbol{c})$. Demand is $\boldsymbol{c}(\boldsymbol{q}, w) = \arg\max_{\boldsymbol{c}} u^s(\boldsymbol{c})$ s.t. $B(\boldsymbol{q}, \boldsymbol{c}) \leq w$.

Consider another agent who is rational with utility u^s. We call $v^s(\boldsymbol{q}, w) = u^s(\boldsymbol{c}(\boldsymbol{q}, w))$ his utility. For that other, rational agent, call $\boldsymbol{S}^{s,r}(\boldsymbol{q}, w) = \boldsymbol{c}_{\boldsymbol{q}}(\boldsymbol{q}, w) + \boldsymbol{c}_w(\boldsymbol{q}, w)' B_{\boldsymbol{q}}$ his Slutsky matrix. Given the previous results, the following Propsition is immediate.

Proposition 7.3 *In the decision vs. experienced utility model,* $\boldsymbol{S}_j^C = \boldsymbol{S}_j^{s,r}$ *is the Slutsky matrix of a rational agent with utility* $u^s(\boldsymbol{c})$*. The behavioral wedge is:*

$$\boldsymbol{\tau}^b = \frac{u_{\boldsymbol{c}}^s(\boldsymbol{c}(\boldsymbol{q}, w))}{v_w^s(\boldsymbol{q}, w)} - \frac{u_{\boldsymbol{c}}(\boldsymbol{c}(\boldsymbol{q}, w))}{v_w(\boldsymbol{q}, w)}.$$

Misperception Model To illustrate this framework, we take the misperception model (i.e., the agent developed in Gabaix (2014)). It comprises a perception function $\boldsymbol{q}^s(\boldsymbol{q}, w)$ (which itself can be endogenized, something we consider later). The demand satisfies:

$$\boldsymbol{c}(\boldsymbol{q}, w) = \boldsymbol{h}^r\left(\boldsymbol{q}^s(\boldsymbol{q}, w), v(\boldsymbol{q}, w)\right),$$

where $\boldsymbol{h}^r(\boldsymbol{q}^s, u)$ is the Hicksian demand of a rational agent with perceived prices $\boldsymbol{q}^s(\boldsymbol{q}, w)$.

Proposition 7.4 *Take the misperception model. Then, with* $\boldsymbol{S}^r(\boldsymbol{q},w)=\boldsymbol{h}^r_{\boldsymbol{q}^s}(\boldsymbol{q}^s(\boldsymbol{q},w),v(\boldsymbol{q},w))$ *the Slutsky matrix of the underlying rational agent, we have:*

$$\boldsymbol{S}^H_j(\boldsymbol{q},w)=\boldsymbol{S}^r(\boldsymbol{q},w)\left(\boldsymbol{q}^s_{q_j}(\boldsymbol{q},w)-\boldsymbol{q}^s_w(\boldsymbol{q},w)\frac{v_{q_j}}{v_w}\right), \tag{42}$$

i.e. $S^H_{ij}=\sum_k S^r_{ik}\left(\frac{\partial q^s_k(\boldsymbol{q},w)}{\partial q_j}-\frac{\partial q^s_k(\boldsymbol{q},w)}{\partial w}\frac{v_{q_j}}{v_w}\right)$*, where* $\frac{\partial q^s_k(\boldsymbol{q},w)}{\partial q_j}-\frac{\partial q^s_k(\boldsymbol{q},w)}{\partial w}\frac{v_{q_j}}{v_w}$ *is the Hicksian marginal perception matrix. Also*

$$\boldsymbol{\tau}^b=B_{\boldsymbol{c}}(\boldsymbol{c},\boldsymbol{q})-\frac{B_{\boldsymbol{c}}(\boldsymbol{c},\boldsymbol{q}^s)}{B_{\boldsymbol{c}}(\boldsymbol{c},\boldsymbol{q}^s)\cdot\boldsymbol{c}_w(\boldsymbol{q},w)}. \tag{43}$$

Given $B_{\boldsymbol{c}}(\boldsymbol{q}^s,\boldsymbol{c})\cdot\boldsymbol{S}^H_j=0$*, we have:*

$$D_j=-\left(B_{\boldsymbol{c}}(\boldsymbol{q},\boldsymbol{c})-B_{\boldsymbol{c}}(\boldsymbol{q}^s,\boldsymbol{c})\right)\cdot\boldsymbol{S}^H_j=-B_{\boldsymbol{c}}(\boldsymbol{q},\boldsymbol{c})\cdot\boldsymbol{S}^H_j, \tag{44}$$

so that

$$D_j=-\overline{\boldsymbol{\tau}}^b\cdot\boldsymbol{S}^H_j \text{ with } \overline{\boldsymbol{\tau}}^b=B_{\boldsymbol{c}}(\boldsymbol{q},\boldsymbol{c})-B_{\boldsymbol{c}}(\boldsymbol{q}^s,\boldsymbol{c}). \tag{45}$$

This implies that in welfare formulas we can take $\boldsymbol{\tau}^b=B_{\boldsymbol{c}}(\boldsymbol{q},\boldsymbol{c})-B_{\boldsymbol{c}}(\boldsymbol{q}^s,\boldsymbol{c})$ rather than the more cumbersome $\boldsymbol{\tau}^b=B_{\boldsymbol{c}}(\boldsymbol{c},\boldsymbol{q})-\frac{B_{\boldsymbol{c}}(\boldsymbol{c},\boldsymbol{q}^s)}{B_{\boldsymbol{c}}(\boldsymbol{c},\boldsymbol{q}^s)\cdot\boldsymbol{c}_w}$.

Proof. Given $\boldsymbol{c}(\boldsymbol{q},w)=\boldsymbol{h}^r(\boldsymbol{q}^s(\boldsymbol{q},w),v(\boldsymbol{q},w))$, we have $\boldsymbol{c}_w=\boldsymbol{h}^r_u v_w+\boldsymbol{h}^r_{\boldsymbol{q}^s}\boldsymbol{q}^s_w$. Then,

$$\begin{aligned}
\boldsymbol{S}^H_j&=\boldsymbol{c}_{q_j}(\boldsymbol{q},w)-\boldsymbol{c}_w(\boldsymbol{q},w)\frac{v_{q_j}(\boldsymbol{q},w)}{v_w(\boldsymbol{q},w)}=\boldsymbol{h}^r_{\boldsymbol{q}^s}\boldsymbol{q}^s_{q_j}(\boldsymbol{q},w)+\boldsymbol{h}^r_u v_{q_j}-\boldsymbol{c}_w\frac{v_{q_j}}{v_w}\\
&=\boldsymbol{h}^r_{\boldsymbol{q}^s}\boldsymbol{q}^s_{q_j}(\boldsymbol{q},w)+\boldsymbol{h}^r_u v_{q_j}-\left(\boldsymbol{h}^r_u v_w+\boldsymbol{h}^r_{\boldsymbol{q}^s}\boldsymbol{q}^s_w(\boldsymbol{q},w)\right)\frac{v_{q_j}}{v_w}\text{ as }\boldsymbol{c}_w=\boldsymbol{h}^r_u v_w+\boldsymbol{h}^r_{\boldsymbol{q}^s}\boldsymbol{q}^s_w\\
&=\boldsymbol{S}^r\left(\boldsymbol{q}^s_{q_j}(\boldsymbol{q},w)-\boldsymbol{q}^s_w(\boldsymbol{q},w)\frac{v_{q_j}}{v_w}\right).
\end{aligned}$$

Next, observe that the demand satisfies $u_{\boldsymbol{c}}(\boldsymbol{q},w)=\Lambda B_{\boldsymbol{c}}(\boldsymbol{q}^s,\boldsymbol{c})$ for some Lagrange multiplier Λ, and that $B_{\boldsymbol{c}}(\boldsymbol{q}^s,\boldsymbol{c})\boldsymbol{S}^r=\boldsymbol{0}$ for a rational agent (see equation (41) applied to that agent). So, $B_{\boldsymbol{c}}(\boldsymbol{q}^s,\boldsymbol{c})\boldsymbol{S}^H=\boldsymbol{0}$. Next,

$$\begin{aligned}
D_j(\boldsymbol{q},w)&=-\boldsymbol{\tau}^b\cdot\boldsymbol{S}^H_j=-\left(B_{\boldsymbol{c}}-\frac{u_{\boldsymbol{c}}}{v_w}\right)\boldsymbol{S}^r\left(\boldsymbol{q}^s_{q_j}(\boldsymbol{q},w)-\boldsymbol{q}^s_w(\boldsymbol{q},w)\frac{v_{q_j}}{v_w}\right)\\
&=-\left(B_{\boldsymbol{c}}-\frac{\Lambda B_{\boldsymbol{c}}(\boldsymbol{q}^s,\boldsymbol{c})}{v_w(\boldsymbol{q},w)}\right)\boldsymbol{S}^r\left(\boldsymbol{q}^s_{q_j}(\boldsymbol{q},w)-\boldsymbol{q}^s_w(\boldsymbol{q},w)\frac{v_{q_j}}{v_w}\right)\\
&=-B_{\boldsymbol{c}}\boldsymbol{S}^r\left(\boldsymbol{q}^s_{q_j}(\boldsymbol{q},w)-\boldsymbol{q}^s_w(\boldsymbol{q},w)\frac{v_{q_j}}{v_w}\right)\\
&=-\left(B_{\boldsymbol{c}}-B_{\boldsymbol{c}}(\boldsymbol{q}^s,\boldsymbol{c})\right)\cdot\boldsymbol{S}^r\left(\boldsymbol{q}^s_{q_j}(\boldsymbol{q},w)-\boldsymbol{q}^s_w(\boldsymbol{q},w)\frac{v_{q_j}}{v_w}\right).
\end{aligned}$$

Given (34), $\frac{u_{\boldsymbol{c}}(\boldsymbol{c}(\boldsymbol{q},w))}{v_w(v,w)}=\frac{u_{\boldsymbol{c}}}{u_{\boldsymbol{c}}\cdot\boldsymbol{c}_w}=\frac{B_{\boldsymbol{c}}(\boldsymbol{c},\boldsymbol{q}^s)}{B_{\boldsymbol{c}}(\boldsymbol{c},\boldsymbol{q}^s)\cdot\boldsymbol{c}_w}$.

Finally, (5) comes from (40):[73]

$$\boldsymbol{S}_j^C = \boldsymbol{S}_j^H - \boldsymbol{c}_w \left(\boldsymbol{\tau}^b \cdot \boldsymbol{S}_j^H\right) = \left(I - \boldsymbol{c}_w \left(\boldsymbol{\tau}^b\right)'\right) \boldsymbol{S}_j^H.$$

□

8 Additional Proofs

Proof of Proposition 2.1 We have

$$\frac{\partial L}{\partial \tau_i} = \sum_h \left[W_{v^h} v_w^h \frac{v_{q_i}^h}{v_w^h} + \lambda c_i^h + \lambda \boldsymbol{\tau} \cdot \boldsymbol{c}_{q_i}^h \right].$$

Using the definition of $\beta^h = W_{v^h} v_w^h$, the behavioral versions of Roy's identity (2), and the Slustky relation, we can rewrite this as

$$\frac{\partial L}{\partial \tau_i} = \sum_h [\beta^h \left(-c_i^h - \boldsymbol{\tau}^{b,h} \cdot \boldsymbol{S}_i^{C,h}\right) + \lambda c_i^h + \lambda \boldsymbol{\tau} \cdot (-\boldsymbol{c}_w^h c_i^h + \boldsymbol{S}_i^{C,h})].$$

We then use the definition of the social marginal utility of income $\gamma^h = \beta^h + \lambda \boldsymbol{\tau} \cdot \boldsymbol{c}_w^h$ to get

$$\frac{\partial L}{\partial \tau_i} = \sum_h [\left(\lambda - \gamma^h\right) c_i^h + \left[\lambda \boldsymbol{\tau} - \beta^h \boldsymbol{\tau}^{b,h}\right] \cdot \boldsymbol{S}_i^{C,h}].$$

The result follows using the renormalization (7) of the behavioral wedge.

Proof of Proposition 2.3

$$\frac{d\xi}{d\chi} = \sum_h \xi_{\boldsymbol{c}^h} \left[\boldsymbol{c}_\chi^h + \boldsymbol{c}_\xi^h \frac{d\xi}{d\chi} \right],$$

so $\frac{d\xi}{d\chi} = \frac{\sum_h \xi_{\boldsymbol{c}^h} \cdot \boldsymbol{c}_\chi^h}{1 - \sum_h \xi_{\boldsymbol{c}^h} \cdot \boldsymbol{c}_\xi^h}$. Thus the additional term in $\frac{\partial L}{\partial \chi}$ arising due to externality is

$$\frac{d\xi}{d\chi} \left\{ \sum_h W_{v^h} v_w^h \frac{v_\xi^h}{v_w^h} + \lambda \sum_h \boldsymbol{\tau} \cdot \boldsymbol{c}_\xi^h \left(\boldsymbol{q}, w^h, \xi, \chi\right) \right\} = \Xi \sum_h \xi_{\boldsymbol{c}^h} \cdot \boldsymbol{c}_\chi^h$$

[73] Another useful relation is that $u_{\boldsymbol{c}} \boldsymbol{S}^H = \boldsymbol{0}$ in the (static) misperception model (this is because $u_{\boldsymbol{c}} = \Lambda B_{\boldsymbol{c}} \left(\boldsymbol{c}, \boldsymbol{q}^s\right)$ for some scalar Λ, and $B_{\boldsymbol{c}} \left(\boldsymbol{c}, \boldsymbol{q}^s\right) \boldsymbol{S}^H = 0$ from equation (41)). This is not true in the decision vs. experienced utility model.

We use the fact that $\boldsymbol{q}\cdot\boldsymbol{c}\left(\boldsymbol{q},w,\chi\right)=w$ implies $\boldsymbol{q}\cdot\boldsymbol{c}_{\chi}=0$:

$$\begin{aligned}\frac{\partial L}{\partial\chi}&=\sum_{h}\left\{W_{v^{h}}v_{w}^{h}\frac{u_{\boldsymbol{c}}^{h}}{v_{w}^{h}}\boldsymbol{c}_{\chi}^{h}+W_{v^{h}}v_{w}^{h}\frac{u_{\chi}^{h}}{v_{w}^{h}}+\lambda\boldsymbol{\tau}\cdot\boldsymbol{c}_{\chi}^{h}+\Xi\xi_{\boldsymbol{c}^{h}}\cdot\boldsymbol{c}_{\chi}^{h}\right\}\\&=\sum_{h}\left\{\left[W_{v^{h}}v_{w}^{h}\frac{u_{\boldsymbol{c}}^{h}}{v_{w}^{h}}+\lambda(\boldsymbol{\tau}-\boldsymbol{\tau}^{\xi,h})\right]\boldsymbol{c}_{\chi}^{h}+W_{v^{h}}v_{w}^{h}\frac{u_{\chi}^{h}}{v_{w}^{h}}\right\}\\&=\sum_{h}\left\{\left[\beta^{h}\left(\frac{u_{\boldsymbol{c}}^{h}}{v_{w}^{h}}-\boldsymbol{q}+\boldsymbol{q}\right)+\lambda(\boldsymbol{\tau}-\boldsymbol{\tau}^{\xi,h})\right]\boldsymbol{c}_{\chi}^{h}+\beta^{h}\frac{u_{\chi}^{h}}{v_{w}^{h}}\right\}\\&=\sum_{h}\left\{[-\lambda\widetilde{\boldsymbol{\tau}}^{b,h}+\lambda(\boldsymbol{\tau}-\boldsymbol{\tau}^{\xi,h})]\boldsymbol{c}_{\chi}^{h}+\beta^{h}\frac{u_{\chi}^{h}}{v_{w}^{h}}\right\}.\end{aligned}$$

Tax formula in the limit of small taxes. We can obtain a formula similar to (14) for the optimal tax, without assuming quasilinear utility (for simplicity, we assume no Pigouvian externality). We assume that for small taxes agent consume $\boldsymbol{c}^{h}\left(\boldsymbol{p}+\boldsymbol{\tau}\right)=\boldsymbol{c}^{r,h}\left(\boldsymbol{p}+\boldsymbol{\tau}\right)+\hat{\boldsymbol{c}}^{u,h}\left(\boldsymbol{p},w\right)+\hat{\boldsymbol{c}}^{M,h}\left(\boldsymbol{p},w\right)\boldsymbol{\tau}+O\left(\|\boldsymbol{\tau}\|^{2}\right)+O\left(\|\hat{\boldsymbol{c}}^{u,h}\left(\boldsymbol{p},w\right)\|^{2}\right)$. This formulation captures two forces. First, even if taxes are 0, consumers may misoptimize, as captured by the term $\hat{\boldsymbol{c}}^{u,h}\left(\boldsymbol{p},w\right)$, which we take to be small in our limit of small taxes. Second, they may mis-react to taxes, as captured by the term $\hat{\boldsymbol{c}}^{M,h}\left(\boldsymbol{p},w\right)\boldsymbol{\tau}$. This general formulation gives an attention $\boldsymbol{M}^{h}=\boldsymbol{I}+\left(\boldsymbol{c}_{\boldsymbol{p}}^{r,h}\right)^{-1}\hat{\boldsymbol{c}}^{M,h}$. Then, (as detailed in Section 11.1 of the online appendix), the optimal tax is, up to the second order in $\tilde{\eta}$:

$$\boldsymbol{\tau}=-\left[\sum_{h}\left(\boldsymbol{S}^{r,h}+\hat{\boldsymbol{c}}^{M,h}\right)'\left(\boldsymbol{I}-\Omega^{h}\hat{\boldsymbol{c}}^{M,h}\right)+\frac{v_{ww}}{v_{w}}\boldsymbol{c}^{h}\boldsymbol{c}^{h\prime}\right]_{>0}^{-1}\left[\sum_{h}\left(1-\frac{b^{h}}{\lambda}\right)\boldsymbol{c}^{h}-\left(\boldsymbol{S}^{r,h}+\hat{\boldsymbol{c}}^{M,h}\right)'\Omega^{h}\hat{\boldsymbol{c}}^{u,h}\right]_{>0},\tag{46}$$

where $\Omega^{h}=-\frac{u_{\boldsymbol{cc}}^{h}(\boldsymbol{p},w)}{v_{w}^{h}(\boldsymbol{p},w)}$, $\boldsymbol{S}^{r,h}=\boldsymbol{c}_{\boldsymbol{p}}^{h}+\boldsymbol{c}_{w}^{h}\boldsymbol{c}^{h\prime}$, $\tilde{\eta}=\sum_{h}\left|b^{h}-\lambda\right|+\|\hat{\boldsymbol{c}}^{u,h}\left(\boldsymbol{p},w\right)\|$. All the variables are evaluated at $(\boldsymbol{p},w)$ and subscript >0 indicates the selection of the $(N-1)\times(N-1)$ sub-matrix corresponding to all goods except good 0.

The numerator of (46) features: $\left(1-\frac{b^{h}}{\lambda}\right)\boldsymbol{c}^{h}$, which is the revenue-raising / redistributive motive; $\hat{\boldsymbol{c}}^{u,h}$, which captures the consumption mistakes made by the agents before any taxes; and $\left(\boldsymbol{S}^{r,h}+\hat{\boldsymbol{c}}^{M,h}\right)'\Omega^{h}$, which captures the Slutsky matrix of the agent, corrected by their misperception to taxes $\hat{\boldsymbol{c}}^{M,h}$. The denominator is a matrix version of the inverse elasticity, adjusted for income effects. This is the expression that shows up in more user-friendly terms throughout Section 3 and in (14).