Optimal Taxation with Behavioral Agents*

Emmanuel Farhi and Xavier Gabaix

August 22, 2019

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, internalities and mental accounting. 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. We explore the Diamond-Mirrlees productive efficiency result and the Atkinson-Stiglitz uniform commodity taxation proposition, and find that they are more likely to fail with behavioral agents. (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, internalities, or mental accounting), 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.

*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, C. Wang, and for helpful comments we thank the editor and referees, seminar participants at Berkeley, BEAM, BRIC, Brown, BU, Chicago, Columbia, Harvard, IIES, NBER, NYU, Princeton, PSE, Stanford, the UCL conference on behavioral theory, Yale, 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.

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 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.

The generality of our framework allows us to unify existing results in behavioral public finance as well as to derive new insights. A non-exhaustive list includes: a modified Ramsey inverse elasticity rule (for a given elasticity, inattention increases the optimal tax, essentially quadratically); a modified optimal Pigouvian tax rule (for a given externality, inattention increases the optimal tax, essentially linearly); a behavioral role for quantity regulation (heterogeneity in attention favors quantity regulation over price regulation); the attractiveness of targeted nudges (which respects freedom of choice for rational agents and limit the tax burden of the poor); a mental-account justification for vouchers (vouchers increase spending on food even if they are infra-marginal); a modification of the principle of targeting (in the traditional model, it is optimal to tax the externality-generating good, but not to subsidize substitute goods; in the behavioral model, it is actually optimal to subsidize substitute goods); in the Mirrleesian optimal nonlinear income tax, marginal income tax rates can be negative even with only an intensive labor margin, something that is not possible with rational agents; 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. Conversely, if the wealthy overperceive the productivity of effort, top marginal rates are higher than the traditional analysis. Of course, these results require specific assumptions, which we make explicit as we derive them.

We also revisit two classical results regarding supply elasticities and production efficiency. The first classical result states that optimal tax formulas do not depend directly on supply elasticities if there is a full set of commodity taxes. The second classical result, due to Diamond and Mirrlees (1971), states that under some technical conditions, production efficiency holds at the optimum (so that, for example, intermediate goods should not be taxed) if there is a complete set of commodity taxes and if there are constant returns to scale, or if profits are fully taxed. We show that both results can fail when agents are behavioral because agents might misperceive taxes. Roughly, a more stringent condition is required, namely, a full set of commodity taxes that agents perceive like prices (in addition perhaps to other commodity taxes which could be perceived differently from prices). Finally, we show that the celebrated uniform commodity taxation result of Atkinson and Stiglitz (1976) requires more stringent conditions when agents are behavioral.

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, Looney and Kroft (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)).

In a pioneering paper, Mullainathan, Schwartzstein and Congdon (2012) offer a rich 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, Mullainathan and Schwartzstein (2015) further develop those ideas in the context of health care. Allcott, Mullainathan and Taubinsky (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 Allcott, Lockwood and Taubinsky (2019) 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. We rely on the sparse agent of Gabaix (2014) for many illustrations, which builds on the burgeoning literature on inattention (Bordalo, Gennaioli and Shleifer (2013), Caplin and Dean (2015), Chetty, Looney and Kroft (2009), Gabaix and Laibson (2006), Kőszegi and Szeidl (2013), Schwartzstein (2014), Sims (2003), Woodford (2012)). This agent misperceives prices in a way that can be endogenized to economize on attention (hence the name "sparse") and respects 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 make some progress on the modelling of nudges and mental accounts.

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.

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

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. Section 5 revisits Diamond and Mirrlees (1971) and Atkinson and Stiglitz (1976). 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 Sections 3.1-3.6 are relatively self-contained and use small amount of formalism. They also contains 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, mental accounting, 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 $\boldsymbol{\tau}^b(\boldsymbol{q}, w)$ turns out to be an important sufficient statistic for behavioral biases: it encodes the welfare effects of a marginal reduction in the consumption of the 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 $\boldsymbol{S}_j^C(\boldsymbol{q}, w)$ is the "income-compensated" Slutsky matrix 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, and is equal to 0 with traditional rational 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.

As an example, consider the case of a smoker, who smokes $c_j = 1$ pack of cigarettes 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 identity says that if the smoker is rational, his utility is reduced by exactly 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 by a dollar equivalent of $\tau_j^b = 10.5$ dollars per pack. Then assuming that the behavioral wedges are zero for all goods but cigarettes ($\tau_i^b = 0$ for $i \neq j$), the behavioral Roy 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

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

[3] We refer the reader to Appendix 7 for the detailed derivations.

the price of cigarettes because it helps the agent curb his excessive smoking.[4]

We now present three useful concrete instantiations of the general formalism: decision vs. experienced utility, misperceptions, and mental accounts.

Decision vs. Experienced Utility Model We start with the decision vs. experienced utility model, in which 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 $u^s(\boldsymbol{c})$.

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.[5] 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 $\boldsymbol{q} \cdot \boldsymbol{c} = \boldsymbol{w}$.[6] 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

[4] Jonathan Gruber and Botond Kőszegi (2004) estimate that the total future health costs of a pack of cigarettes is $h = 35$ dollars. 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. Jonathan Gruber and Botond Kőszegi (2004) report a demand elasticity of below-median-income smokers of $\psi = 0.7$. With $q_j = 5$ dollars per pack and $c_j = 1$ pack a day, the Slutsky term is $S_{jj}^C = -\frac{\psi c_j}{q_j} = -0.14$ packs per dollar per day.

[5] 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.

[6] The problem has a solution under the usual Inada conditions. If there are several such λ, we take the lowest one, which is also the utility-maximizing one. This is the formulation advocated out in Gabaix (2014), who discusses it extensively.

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 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 $\boldsymbol{M} = \boldsymbol{q}_{\boldsymbol{q}}^s$ is the matrix of marginal misperceptions. It shows how a change in the price q_j of good j creates a change $M_{kj}(\boldsymbol{q}, w) = \frac{\partial q_k^s(\boldsymbol{q}, w)}{\partial q_j}$ in the perceived price q_k^s 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.[7] The term $\boldsymbol{c}_w(\boldsymbol{q}, w) \left(\boldsymbol{\tau}^b(\boldsymbol{q}, w)\right)'$ is a correction for wealth effects.

Mental Accounts There is no agreed-upon model of mental accounting. Here we propose a simple formalism which we think can be useful to capture some important dimensions of mental accounts. In Section 3.6, we flesh out a concrete application of this model in the context of food vouchers.

The primitives are an experienced utility function u, a partition of the set of commodities into K subsets or accounts indexed by $k = 1, ..., K$, mental accounting functions $\omega^k(\boldsymbol{q}, w)$, and an extended demand function $\boldsymbol{c}(\boldsymbol{q}, \boldsymbol{\omega})$, where $\boldsymbol{\omega} = \left(\omega^1, ..., \omega^K\right)$. The mental accounting functions $\omega^k(\boldsymbol{q}, w)$ indicates how much money is devoted to account k, and must satisfy $\sum_k \omega^k(\boldsymbol{q}, w) = w$. We denote by $\boldsymbol{C}^k$ the vector of commodities associated with account k and we write $\boldsymbol{c} = \left(\boldsymbol{C}^1, .., \boldsymbol{C}^K\right)$. The extended demand function must satisfy $\boldsymbol{q}^k \cdot \boldsymbol{C}^k(\boldsymbol{q}, \boldsymbol{\omega}) = \omega^k(\boldsymbol{q}, w)$. The demand function $\boldsymbol{c}(\boldsymbol{q}, w)$ is simply defined by $\boldsymbol{c}(\boldsymbol{q}, w) = \boldsymbol{c}(\boldsymbol{q}, \boldsymbol{\omega}(\boldsymbol{q}, w))$. We denote the extended indirect utility function by $v(\boldsymbol{q}, \boldsymbol{\omega}) = u(\boldsymbol{c}(\boldsymbol{q}, \boldsymbol{\omega}))$. The indirect utility function is $v(\boldsymbol{q}, w) = v(\boldsymbol{q}, \boldsymbol{\omega}(\boldsymbol{q}, w))$.

The expression for the behavioral wedges is particularly enlightening in the case where mental

[7] 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.

accounting is the only behavioral bias so that demand is rational subject to mental accounts:[8]

$$\tau_i^b = q_i \left(1 - \frac{v_{\omega^{k(i)}}(\boldsymbol{q}, \boldsymbol{\omega})}{v_w(\boldsymbol{q}, w)}\right),$$

where $k(i)$ denotes the mental account to which good i belongs. Intuitively, the behavioral wedge for good i is positive if it belongs to a mental account $k(i)$ on which the agent overspends $v_{\omega^{k(i)}}(\boldsymbol{q}, \boldsymbol{\omega}) < v_w(\boldsymbol{q}, w)$.

The Slutsky matrix is

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

With these specific particularizations in mind, we are now ready to study the basic taxation problems using the general behavioral model.

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 $\boldsymbol{p}$. We relax this assumption in Section 5.1 where we consider the case of imperfectly elastic supply with endogenous producer prices $\boldsymbol{p}$.

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: $\tau_0 = 0$.[9] 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.

The planning problem is[10]

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

where[11]

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

[8] Rational demand subject to mental accounts corresponds to $\boldsymbol{c}^r(\boldsymbol{q}, \boldsymbol{\omega}) = \arg\max_{\boldsymbol{c}} u(\boldsymbol{c})$ s.t. $\boldsymbol{q}^k \cdot \boldsymbol{C}^k = \omega^k$ for $k = 1, ..., K$. The traditional model with frictionless mental accounts can be recovered as a special case by specifying $\omega^{k,r}(\boldsymbol{q}, w) = \boldsymbol{q}^k \cdot \boldsymbol{C}^{k,r}(\boldsymbol{q}, w)$, where $\boldsymbol{c}^r(\boldsymbol{q}, w) = \left(\boldsymbol{C}^{1,r}(\boldsymbol{q}, w), .., \boldsymbol{C}^{K,r}(\boldsymbol{q}, w)\right)$ is the demand function of a rational agent.

[9] 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.

[10] 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.

[11] The analysis is identical if we allow for endowments $\boldsymbol{e}^h$, using the objective function

$$L(\boldsymbol{\tau}) = W\left(\left(v^h\left(\boldsymbol{p} + \boldsymbol{\tau}, w + \boldsymbol{p} \cdot \boldsymbol{e}^h\right)\right)_{h=1...H}\right) + \lambda \sum_h \left[\boldsymbol{\tau} \cdot \boldsymbol{c}^h\left(\boldsymbol{p} + \boldsymbol{\tau}, w + \boldsymbol{p} \cdot \boldsymbol{e}^h\right) - w\right].$$

Following Diamond (1975), we define $\gamma^h = \beta^h + \lambda \boldsymbol{\tau} \cdot \boldsymbol{c}_w^h$ to be the social marginal utility of income for agent h where $\beta^h = W_{v^h} v_w^h$ is the social marginal welfare weight. The difference $\lambda \boldsymbol{\tau} \cdot \boldsymbol{c}_w^h$ between γ^h and β^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

$$\widetilde{\boldsymbol{\tau}}^{b,h} = \frac{\beta^h}{\lambda} \boldsymbol{\tau}^{b,h}. \tag{6}$$

We now characterize the optimal tax system.[12]

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

$$\frac{\partial L(\boldsymbol{\tau})}{\partial \tau_i} = 0 \quad \textit{with} \quad \frac{\partial L(\boldsymbol{\tau})}{\partial \tau_i} = \sum_h [\left(\lambda - \gamma^h\right) c_i^h + \lambda (\boldsymbol{\tau} - \widetilde{\boldsymbol{\tau}}^{b,h}) \cdot \boldsymbol{S}_i^{C,h}]. \tag{7}$$

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 $\sum_h (\lambda - \gamma^h) c_i^h d\tau_i$. If there were no changes in behavior, then the government would collect additional revenues $c_i^h d\tau_i$ from agent h, which are valued by the government as $(\lambda - \gamma^h) c_i^h d\tau_i$. Indeed, transferring one dollar from agent h to the government creates a net welfare change of $\lambda - \gamma^h$, where λ is the value of public funds and γ^h is the social marginal utility of income for agent h (which includes the associated income effect on tax revenues).

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

Finally, let us analyze the misoptimization effect $-\sum_h \lambda \widetilde{\boldsymbol{\tau}}^{b,h} \cdot \boldsymbol{S}_i^{C,h} d\tau_i$. This effect is linked to the substitution effect. If agent h were rational, then the change in behavior captured by the substitution effect would have no first-order effects on his utility. This is a consequence of the envelope theorem. When agent h is behavioral, this logic fails, and the change in behavior associated with the substitution effect has first-order effects $-\beta^h \boldsymbol{\tau}^{b,h} \cdot \boldsymbol{S}_i^{C,h} d\tau_i = -\lambda \widetilde{\boldsymbol{\tau}}^{b,h} \cdot \boldsymbol{S}_i^{C,h} d\tau_i$ on his utility.

All in all, adding behavioral agents introduces the following differences. First, introducing behavioral agents modifies the social welfare weights, income effect, and substitution effect, leading to different values for β^h, γ^h, and a different Slutsky matrix $\boldsymbol{S}_i^{C,h}$. Second, there is a new effect (the

[12] Suppose that there is uncertainty, possibly heterogeneous beliefs, several dates for consumption, and complete markets. Then, our formula (7) 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.

misoptimization effect) leading to a new term $-\lambda\widetilde{\boldsymbol{\tau}}^{b,h}\cdot\boldsymbol{S}_i^{C,h}$.

One way to think about the optimal tax formulas (7) is as a linear system of equations indexed by i in the optimal taxes τ_j for the different commodities

$$\frac{-\sum_{j,h}\boldsymbol{S}_{ji}^{C,h}\tau_j}{c_i}=1-\frac{\bar{\gamma}}{\lambda}-cov(\frac{\gamma^h}{\lambda},\frac{Hc_i^h}{c_i})-\frac{\sum_{j,h}\widetilde{\boldsymbol{\tau}}_j^{b,h}\boldsymbol{S}_{ji}^{C,h}}{c_i},$$

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. Of course the coefficients in this linear system of equations and the right-hand-side terms are endogenous and depend on taxes τ_j. Nevertheless, at the optimum, one can in principle solve out the linear system to express the taxes τ_j as a function of these coefficients and the forcing terms (valued at optimal taxes). 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. The second right-hand-side term $\frac{-\sum_{j,h}\widetilde{\tau}_j^{b,h}\boldsymbol{S}_{ji}^{C,h}}{c_i}$ captures the corrective objective of taxation to address the effects of misoptimization.[13]

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 $\xi=\xi\left(\left(\boldsymbol{c}^h\right)_{h=1...H}\right)$ is a one-dimensional externality (for simplicity) that depends on the consumption vectors of all agents and is therefore endogenous to the tax system.[14] 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

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

and $\xi=\xi\left(\left(\boldsymbol{c}^h\left(\boldsymbol{p}+\boldsymbol{\tau},w,\xi\right)\right)_{h=1...H}\right)$. We call $\Xi=\frac{\sum_h\left[\beta^h\frac{v_\xi^h}{v_w^h}+\lambda\boldsymbol{\tau}\cdot\boldsymbol{c}_\xi^h\right]}{1-\sum_h\xi_{\boldsymbol{c}^h}\boldsymbol{c}_\xi^h}$ 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 $\lambda\boldsymbol{\tau}\cdot\boldsymbol{c}_\xi^h$ in the numerator) 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

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

[13] 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"). This amounts to assuming that the government can adjust w. 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 w yielding $\bar{\gamma}=\lambda$.

[14] 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 $u^h\left(\boldsymbol{c}^h,\xi\right)=u^h\left(\boldsymbol{c}^h\right)-\xi$.

It represents the 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 $\gamma^{\xi,h} = \gamma^h + \Xi \xi_{\boldsymbol{c}^h} \boldsymbol{c}_w^h = \beta^h + \lambda \left(\boldsymbol{\tau} - \boldsymbol{\tau}^{\xi,h} \right) \cdot \boldsymbol{c}_w^h$.[15] The next proposition generalizes Proposition 2.1.

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

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

Formally, misoptimization and externality wedges $(\widetilde{\boldsymbol{\tau}}^{b,h}, \boldsymbol{\tau}^{\xi,h})$ enter symmetrically in the optimal tax formula. In some particular cases, behavioral biases can be alternatively modeled as externalities (for example, this is the case for a decision vs. experienced utility model with a representative agent). But this is not true in general. For example, misperceptions of prices typically give rise to non-symmetric Slutsky matrices $\boldsymbol{S}_i^{C,h}$ 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 401ks 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 $u^h(\boldsymbol{c}, \chi)$.[16]

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

[15] 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 $\xi_{\boldsymbol{c}^h} \boldsymbol{c}_w^h$, with a welfare impact $\Xi \xi_{\boldsymbol{c}^h} \boldsymbol{c}_w^h$.

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

perceived price of good to $q_j^{s,h} = q_j^{s,h,*} + \chi\eta^h$ if $j = i$ and $q_j^{s,h} = q_j^{s,h,*}$ otherwise, 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 ($\chi < 0$). 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 $u^h(\boldsymbol{c}, \chi) = u^h(\boldsymbol{c}) - \iota^h \chi c_i$.[17]

Proposition 2.3 (Optimal nudge formula) *Optimal nudges satisfy*

$$\frac{\partial L(\boldsymbol{\tau}, \chi)}{\partial \chi} = 0, \quad \textit{with} \quad \frac{\partial L}{\partial \chi}(\boldsymbol{\tau}, \chi) = \sum_h [\lambda(\boldsymbol{\tau} - \boldsymbol{\tau}^{\xi,h} - \widetilde{\boldsymbol{\tau}}^{b,h}) \cdot \boldsymbol{c}_\chi^h + \beta^h \frac{u_\chi^h}{v_w^h}]. \tag{9}$$

The optimality conditions for taxes $\frac{\partial L(\boldsymbol{\tau}, \chi)}{\partial \tau_i} = 0$ *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.[18]

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.[19] This approach, which is common but not uncontroversial, departs from

[17] 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.

[18] 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).

[19] Arguably, agents' mistakes can be persistent. For example, Slemrod (2006) argues that Americans overestimate on average the odds their inheritance 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)).

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's 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. Lewis (1970) puts it eloquently:

> Of all tyrannies, a tyranny sincerely exercised for the good of its victims may be the most oppressive. It would be better to live under robber barons than under omnipotent moral busybodies. The robber baron's cruelty may sometimes sleep, his cupidity may at some point be satiated; but those who torment us for our own good will torment us without end for they do so with the approval of their own conscience.

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.

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*b*) and references therein).

"Lucas Critique" A difficulty confronting all behavioral policy approaches is a form of 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 3.7). 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 $\tilde{\boldsymbol{\tau}}^{b,h}$.[20]

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. 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.

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.

Indeed, 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 concrete a 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.[21] 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, Looney and Kroft (2009), Allcott, Mullainathan and

[20] Sometimes a given bias can me modeled using two distinct particularizations (decisions vs. experienced utility, misperceptions, and mental accounting). 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, Schwartzstein and Congdon (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).

[21] 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 (2014), and Feldman, Katuscak and Kawano (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 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 $\boldsymbol{\tau}^{b,h} = \frac{u_c^s}{v_w^s} - \frac{u_c}{v_w}$.[22]

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, Allcott, Lockwood and Taubinsky (2019) 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 Case with Quasilinear Utility

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 quasilinear 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 $\xi = \xi(\left(\boldsymbol{C}^h\right)_{h=1...H})$ 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

[22] 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.

price is $\boldsymbol{q} = \boldsymbol{p} + \boldsymbol{\tau}$, he perceives prices to be

$$\boldsymbol{q}^{s,h} = \boldsymbol{p} + \boldsymbol{M}^h \boldsymbol{\tau}, \tag{10}$$

where $\boldsymbol{M}^h$ is a constant matrix of marginal perceptions (which in practice will be diagonal $\boldsymbol{M}^h = diag\left(m_i^h\right)_{i=1...n}$). The corresponding perception function is $\boldsymbol{q}^{s,h}(\boldsymbol{q}) = \boldsymbol{p} + \boldsymbol{M}^h(\boldsymbol{q} - \boldsymbol{p})$.[23]

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 $\boldsymbol{S}^{C,h}(\boldsymbol{q}, w) = \boldsymbol{S}^{r,h}\left(\boldsymbol{q}^{s,h}(\boldsymbol{q})\right) \cdot \boldsymbol{M}^h$, 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 $\boldsymbol{\tau}^{X,h} = \frac{\beta^h}{\lambda}\boldsymbol{\tau}^{I,h} + \boldsymbol{\tau}^{\xi,h}$.[24] 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*

$$\boldsymbol{\tau} = -[\sum_h \boldsymbol{M}^{h\prime}\boldsymbol{S}^{r,h}(I - \left(I - \boldsymbol{M}^h\right)\frac{\gamma^h}{\lambda})]^{-1}\sum_h[(1 - \frac{\gamma^h}{\lambda})\boldsymbol{C}^h + \boldsymbol{M}^{h\prime}\boldsymbol{S}^{r,h}\boldsymbol{\tau}^{X,h}]. \tag{11}$$

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.

This formula yields closed forms with explicit comparative statics 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 (11).

3 Examples

In this section, we analyze different applications of the general model in order to extract concrete insights from the optimal tax formulations of the previous section.

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

[24] 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.

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 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.

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 (11).[25] However, we find it useful to also provide a short self-contained rendition. The Ramsey planning problem is

$$\max_{\{\tau_i\}} \gamma \sum_{i=1}^{n} [\frac{[c_i(\tau_i)]^{1-1/\psi_i} - 1}{1 - 1/\psi_i} - (p_i + \tau_i) c_i(\tau_i)] + \lambda \sum_{i=1}^{n} \tau_i c_i(\tau_i), \tag{12}$$

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*

$$\frac{\tau_i}{p_i} = \frac{\Lambda}{\psi_i m_i^2} \cdot \frac{1}{1 + \Lambda \left(\frac{1 - m_i - 1/\psi_i}{m_i} \right)}. \tag{13}$$

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. In their seminal contribution, Mullainathan, Schwartzstein and Congdon (2012) discuss verbally 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 understand better the two terms on the right-hand-side of (13), it is useful to consider the limit of small taxes, which obtains when Λ itself is small: optimal taxes are then given by the first term $(\frac{\Lambda}{\psi_i m_i^2})$ up to the first order in Λ (the second term only introduces second order corrections in Λ).

We find it instructive to provide a self-contained derivation for the limit of small taxes. We can derive a second order approximation of the objective function of the government $\mathcal{L}(\boldsymbol{\tau}) - \mathcal{L}(0) =$

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

$L(\boldsymbol{\tau}) + o\left(\|\boldsymbol{\tau}\|^2\right) + O\left(\Lambda \|\boldsymbol{\tau}\|^2\right)$, with

$$L(\boldsymbol{\tau}) = \frac{-1}{2} \sum_{i=1}^{n} \left(\frac{\tau_i^s}{p_i}\right)^2 \psi_i y_i + \Lambda \sum_{i=1}^{n} \frac{\tau_i}{p_i} y_i, \tag{14}$$

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, Looney and Kroft (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 $L'(\boldsymbol{\tau}) = 0$.

In the limit of small taxes, the traditional Ramsey inverse elasticity rule prescribes that the optimal tax should be $\frac{\tau_i^R}{p_i} = \frac{\Lambda}{\psi_i}$. With inattention, optimal taxes are higher at

$$\frac{\tau_i}{p_i} = \frac{\Lambda}{m_i^2 \psi_i}. \tag{15}$$

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.[26] 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 $\frac{\tau_i^s}{p_i} = \frac{\Lambda}{m_i \psi_i}$.[27]

The upshot of this analysis is that optimal taxes τ_i increase relatively fast with inattention m_i. Formally, in the limit of small taxes, taxes increase quadratically with inattention, so that partial attention m_i leads to a multiplication of the traditional tax by $\frac{1}{m_i^2}$.

Heterogeneity in Attention We now turn our attention to the case where perceptions of taxes are heterogeneous.[28]

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

[26] 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).

[27] 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 $\frac{\tau_i}{p_i} = \frac{\tau_i^s / p_i}{m_i} = \frac{\Lambda}{m_i^2 \psi_i}$.

[28] 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),Dmitry Taubinsky and Alex Rees-Jones (2017).

particular case of formula (11). 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.[29] 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[30]

$$\frac{\tau_i}{p_i} = \frac{\Lambda}{\psi_i \mathbb{E}\left[m_i^{h\,2}\right]} = \frac{\Lambda}{\psi_i \left(\mathbb{E}\left[m_i^h\right]^2 + var\left[m_i^h\right]\right)}, \tag{16}$$

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 demand to the tax), an increase 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.

Before turning to a quantitative illustration, we briefly flesh out two important variants.

Default Taxes It is sometimes important to introduce a distinction between the misperception of marginal tax changes and the misperception of the average level of taxes. To capture this possibility, we assume that perceived taxes are given by $\tau_i^s = m_i \tau_i + (1 - m_i)\, \tau_i^d$, where τ_i^d is a default tax. This change introduces a new additive term $-\frac{\tau_i^d}{p_i}\frac{(1-m_i)(1-\Lambda)}{m_i+(1-m_i)\Lambda}$ in the optimal tax formula (13) to correct for this new form of misperception.[31]

To take a concrete example, suppose that we start from an equilibrium where taxes are optimal and default taxes are equal to true taxes. Imagine that there is a reduction in the need for public funds Λ, but that default taxes τ_i^d remain high at the pre-change level. Then lowering taxes induces agents to over-perceive the average level taxes, and creates a force for the government to lower taxes even further to correct this new bias.

A Costlier Budget-Adjustment Rule 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.

We now explore a variant which increases their costs. We assume that the budget adjustments

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

[30] This can be directly seen by maximizing the second order approximation of the objective function of the government

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

[31] For a treatment with heterogeneous agents, see the online appendix (Section 9.1.2).

are concentrated on a "shock absorber" good with a sharply decreasing marginal utility. This increases the distortionary costs of non-salient taxes and reduces optimal taxes in a way that we characterize precisely below.

The general procedure is laid out in Section 9.5.3 of the online appendix.[32] Here we only present a simple particular case. Utility is separable, $u(\mathbf{c}) = \sum_{i=0}^{n} u_i(c_i)$ with $u_0'(c_0) = 1$, $u_i'(c_i) = c_i^{-1/\psi_i}$ for $i = 1, ..., n-1$ and the "shock absorber" good n has constant marginal utility of $u_n'(c) = 1 - \nu < 1$ if $c_n \geq 1$ and $1 + \mu > 1$ if $c_n < 1$.[33] We call $\mu > 0$ the marginal distortionary cost of budget adjustment. Goods 0 and n cost \$1, and they are untaxed.

The agent chooses his consumption of goods $c_0, ..., c_{n-1}$ based on the perceived prices $q_i^s = 1 + m_i\tau_i$ and the rest of his money is spent on the last good. Specifically, the demands are a follows. For goods $i = 1, ..., n-1$, $c_i = (q_i^s)^{-\psi_i}$ (as the consumer solves $u'(c_i) = q_i^s$). The demand for good 0 is $c_0 = w - \sum_{i=1}^{n-1} (q_i^s)^{1-\psi_i} - 1$, as the consumer plans to consume $c_i = (q_i^s)^{-\psi_i}$ for all good $i = 1...n-1$, and 1 of good n. Once goods 0 through $n-1$ have been purchased, the remaining disposable income for good n is $c_n = w - \sum_{i=0}^{n-1} q_i c_i$.

Then (as derived in the online appendix), the optimal tax on good $i < n$ is as in (13), replacing Λ by $\Lambda_i = \frac{\Lambda - (1-\Lambda)(1-m_i)\mu}{1-(1-\Lambda)(1-m_i)\mu}$. A direct consequence is that the optimal tax τ_i is lower than in the baseline case and is decreasing in μ, particularly for less salient taxes with a small m_i. Indeed, the measure of the social marginal cost of public funds Λ_i is decreasing in the marginal distortionary cost of budget adjustment μ (recall $\Lambda < 1$), coincides with its baseline value of Λ when $\mu = 0$, and is lower than Λ for all $\mu > 0$. Furthermore μ enters the formula through the $\mu(1 - m_i)$ so that these effects are particularly pronounced when attention m_i is low.

Quantitative Illustration To gauge the real-world importance of these effects, we calibrate the model, 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 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)[34] $\mathbb{E}\left[m^h\right] = 0.25$ and $var\left(m^h\right) = 0.13$, so that heterogeneity is very large, $\frac{var\left(m^h\right)}{\mathbb{E}\left[m^h\right]^2} = \frac{0.13}{0.25^2} = 2.1$.[35] In our calibration, we take $\psi = 1$ (as in the Cobb-Douglas case, which is often a good benchmark for the elasticity

[32] This generalizes one of the two adjustment rules studied by Chetty, Looney and Kroft (2009) in the context of a two-good model with separable utility.

[33] The level of ν is unimportant provided it is between 0 and 1.

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

[35] The estimate of mean attention is broadly consistent with the results of Chetty (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).

between broad categories of goods) and $\Lambda = 1.25\%$, which is consistent with the baseline tax in their setup, at $\tau = 7.3\%$.[36] 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 multiplied by 2.8.[37]

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 (11). However, to help build intuition, we start with an elementary and self-contained analysis of the basic Pigou problem. We then use formula (11) 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

$$\max_c U(c) - pc. \tag{17}$$

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 $c_0 + U(c) - \xi_* c$. This would capture the idea that good 1 is tempting and has extra unperceived negative effects $\xi_* c$. The analysis is identical in both cases.

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

$$U(c) - (p + \xi_*) c. \tag{18}$$

To attempt to correct the externality/internality, the government can set a tax τ. Consider first

[36] We use a two-point distribution with rational and behavioral agents to match the mean and dispersion of attention.

[37] The numbers we report in the main text use formula (7) 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 $\frac{\mathbb{E}\left[m^h\right]^2 + var\left(m^h\right)}{\mathbb{E}[m^h]^2} \simeq 3$).

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

$$\max_c U(c) - (p+m\tau)c. \tag{19}$$

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 result.

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

Suppose for concreteness that a good has a negative externality of \$1. With rational agents, it should be taxed by exactly \$1. This is the "dollar-for-dollar" principle of traditional Pigouvian taxation. Accounting for misperception leads to a relaxation of this principle. Indeed, suppose that agents perceive only half of the tax. Then, the good should be taxed by \$2, so that agents perceive a tax of \$1.

It may be contrasted with the modified optimal Ramsey tax (Proposition 3.1), for which in $\frac{\tau_i}{p_i} = \frac{\Lambda}{\psi_i}\frac{1}{m_i^2}$ 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 that Pigouvian taxes should be higher when they are not fully salient is also discussed in Mullainathan, Schwartzstein and Congdon (2012) and could be formalized using their framework.

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.[38]

Heterogeneity in Attention, Externality, Internality We now explore this issue more thoroughly. We now 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, $U^{s,h}\left(c^h\right) - U^h\left(c^h\right) = \xi^h c^h$,

[38] If the budget adjustment is concentrated on a "shock absorber" good with a sharply decreasing marginal utility (as near the end of Section 3.1), then we obtain another force making Pigouvian taxes more distortionary, resulting in lower optimal Pigouvian taxes. This is developed in Section 9.5.3 of the online appendix.

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:

$$U^h(c) = \frac{a^h c - \frac{1}{2}c^2}{\Psi}, \tag{20}$$

which implies a demand function $c^h(q^s) = a^h - \Psi q^s$.[39] 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.

The first best cannot be implemented unless all agents have the same ideal Pigouvian tax, $\frac{\xi^h}{m^h}$. A direct application of formula (11) yields the optimal Pigouvian tax:[40]

$$\tau^* = \frac{\mathbb{E}\left[\xi^h m^h\right]}{\mathbb{E}\left[m^{h2}\right]} = \frac{\mathbb{E}\left[\xi^h\right]\mathbb{E}\left[m^h\right] + cov\left(\xi^h, m^h\right)}{\mathbb{E}\left[m^h\right]^2 + var\left[m^h\right]}. \tag{21}$$

As in the Ramsey case, an increase 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, Knittel and Taubinsky (2015) for a study where subsidies to weatherization is hampered by the fact that people who benefit the most pay the least attention.

Inattention and Tax vs. Quantity Regulation 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*

$$\frac{1}{2\Psi} var\left(c^{h*}\right) < \Psi \frac{\mathbb{E}\left[\xi^{h2}\right]\mathbb{E}\left[m^{h2}\right] - \left(\mathbb{E}\left[\xi^h m^h\right]\right)^2}{2\mathbb{E}\left[m^{h2}\right]}. \tag{22}$$

[39] 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.

[40] This is a direct application of formula (11), with one non-quasilinear good, $\boldsymbol{M}^h = m^h$, $\boldsymbol{S}^{r,h} = -\Psi$, $\boldsymbol{\tau}^{X,h} = \xi^h$.

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 the traditional case with full attention ($m^h = 1$). Then, the right-hand side of (22) 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. Moreover, a higher demand elasticity (high Ψ) favors quantity restrictions, because agents suffer less from a given deviation from their optimal quantity and more from a given price distortion, an effect reminiscent of Weitzman (1974).

With homogeneous inattention $m^h = m < 1$, whether taxes or quantity restrictions are superior remains completely unchanged. With heterogeneous attention, however, the tradeoff is modified in important ways. For example, imagine first that there is no heterogeneity in externality/internalities. If attention were homogeneous, then taxes would dominate. Heterogeneity of attention $var\left[m^h\right]$ then renders taxes less attractive because they introduce misallocation across consumers but do not affect the effectiveness of quantity restrictions.[41] Second, consider the case where externalities/internalities are also heterogeneous. Then 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 optimal criterion (22) 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. For example, compared to this naive reasoning, formula (22) puts less weight on less attentive agents. The main reason is that these agents are also less affected by any given tax.

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.[42] 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.

[41] 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 potentency (the maximization of the right-hand of (22) size of 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, $\Psi \frac{\mathbb{E}\left[\xi^{h\,2}\right]}{2}$.

[42] 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 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 in Section 2.7. In the externality case, we simply assume that $\xi(\left(\boldsymbol{C}^h\right)_{h=1...H}) = \xi_* \frac{1}{H} \sum_h C_1^h$. 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 $\boldsymbol{M}^h = diag(m_1^h, m_2^h)$. Good 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, 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 misperceptions* m_1^h *and* m_2^h *of 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).*

Proposition 3.4 shows that if people have heterogeneous attention to a fuel tax, then solar panels should be subsidized (Allcott, Mullainathan and Taubinsky (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. It should therefore be supplemented with a subsidy on substitute goods and a tax on complement goods. A fuel tax should therefore 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 why the principle of targeting might fail in the

presence of behavioral biases: heterogeneous perceptions of corrective taxes.

3.4 Internalities and Redistribution

Suppose that the poor consume "too many" sugary sodas. This brings up a difficult policy trade-off. On the one hand, taxing sugary sodas corrects the poor's internality. On the other hand, taxing sugary sodas redistributes away from the poor. These were the arguments regarding a recent proposal in New York City. In independent work, Allcott, Lockwood and Taubinsky (2019) examine a related problem, in the context of a Mirrleesian income tax.[43]

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 sodas". We also assume that good 1 is separable, $U^{s,h^*}(\boldsymbol{C}) = U_1^{s,h^*}(c_1) + U_2^{s,h^*}(\boldsymbol{C}_2)$, 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 (11) 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*

$$\tau_1 = \frac{\frac{\gamma^{h^*}}{\lambda}\xi^{h^*} + \left(1 - \frac{\gamma^{h^*}}{\lambda}\right)\frac{p_1}{\psi_1}}{1 + \left(\frac{\gamma^{h^*}}{\lambda} - 1\right)\frac{1}{\psi_1}}. \tag{23}$$

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 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 redistributed consequences.

[43] 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.

Empirical 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 Allcott, Lockwood and Taubinsky (2019): drawing from the medical literature, they find that the "quality adjusted life year" cost of a can of soda is C =12 minutes. Translating this into dollars (using a value of a life year of about \$50,000) gives a dollar cost of soda equal to $C^{\$} = \$1.15$. They next assume a hyperbolic β - δ model with short-run discount factor of $\beta = 0.7$, which translates into an internality $\xi^{h^*} = (1-\beta)\, C^{\$} = \$0.35$.

We use our formula (23). 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 tax is given by the traditional Pigouvian formula $\tau_1 = \xi^{h^*} = \$0.35$, 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 : 0.2, 1, and 2. The optimal tax is then respectively $-\$0.56$, $-\$0.02$ and \$0.22.

3.5 Is it Better to Tax or to Nudge?

In the environment of Section 3.4, 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.[44]

Proposition 3.6 (Optimal nudge vs. tax) *Consider an optimal tax or an optimal nudge in the example of Section 3.4. 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.*

Formula (9) shows that the optimal nudge is given by $\chi = \frac{\xi^{h^*}}{\eta}$, where η is the nudgeability of these agents.[45] 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 (23). 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

[44] Galle (2013) provides a nuanced discussion of nudges vs. taxes.

[45] Section 9.1.3 of the online appendix discusses optimal nudges to correct externalities/internalities with heterogeneous nudgeability.

tax, this guarantees that the optimal nudge does better than the tax. That an optimal nudge does better than the optimal tax 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 motives of the government, a tax helps achieve both motives while a nudge only addresses the latter.

3.6 Mental Accounts

We now study mental accounts (Thaler (1985), Hastings and Shapiro (2013)). First, we present a simple model of optimal vouchers in the presence of mental accounting. Second, we derive some general results on optimal taxation within and across mental accounts.

Vouchers and Mental Accounts Governments often provide assistance in form of vouchers earmarked for a specific category of goods. This is surprising from the point of view of traditional public finance. We show that mental accounting offers a simple way to understand this form of intervention.

We imagine that there are two goods, food (good 1) and non-food (good 2). We allow for two forms of biases. First, there is a difference between decision and experienced utility, which are given by

$$u^s(c_1, c_2) = \frac{c_1^{\alpha_1^s} c_2^{\alpha_2^s}}{\alpha_1^{\alpha_1} \alpha_2^{\alpha_2}}, \qquad u(c_1, c_2) = \frac{c_1^{\alpha_1} c_2^{\alpha_2}}{\alpha_1^{\alpha_1} \alpha_2^{\alpha_2}},$$

with $\alpha_1^s + \alpha_2^s = \alpha_1 + \alpha_2 = 1$. We assume that $\alpha_1^s < \alpha_1$ to capture the sometimes-held notion that the agent suffers from an internality that leads him to spend too little on "wholesome" food and too much on less wholesome non-food. Second, food is subject to mental accounting but non-food is not.

The government gives out general transfers t, and vouchers b which can only be spent on food. Overall income is given by $w = w^* + t + b$, where w^* is pre-tax income. The voucher influences the default expenditure on food according to $\omega_1^d = \alpha_1^s w + \beta b$: a greater $\beta \in [0, 1 - \alpha_1^s]$ indexes a greater degree of mental accounting. We normalize all prices to one.

Mental accounting changes the perceived budget constraint to $c_1 + c_2 + \kappa_1 \left| c_1 - \omega_1^d \right| = w$, but it does not change the true budget constraint $c_1 + c_2 = w$.[46] Here κ_1 parametrizes the degree of mental accounting: deviating from the default expenditure ω_1^d is psychologically costly to the agent.

We next describe the agent's behavior, which is formally analyzed in Section 9.3.5 of the online appendix. If κ_1 is large enough, which we assume from now on, then the agent spends according to the default: $c_1 = \omega_1^d = \alpha_1^s (w^* + t + b) + \beta b$, and so the marginal propensity to consume food (MPCF) out of the voucher is larger than out of a general transfer ($\alpha_1^s + \beta$ vs. α_1^s). This is true even if $c_1 > b$, in sharp contrast to the no-mental accounting case ($\kappa_1 = 0$) where infra-marginal

[46] Formally the consumer solves: $\max_{c_1, c_2} u^s(c_1, c_2) - \lambda \left(c_1 + c_2 + \kappa_1 \left| c_1 - \omega_1^d \right| - w \right)$, where λ is tuned to enforce the budget constraint $c_1 + c_2 = w$.

vouchers are equivalent to general transfers and carry the same marginal propensity to consume (α_1^s).

For a given total amount of revenue $T = t + b$ transferred to the agent, it is always possible to completely correct the agent's internality and to ensure that his spending on food is first best at $\alpha_1 w$ by setting the voucher to income ratio at $\frac{b}{w} = \frac{\alpha_1 - \alpha_1^s}{\beta}$. This by itself provides a rationale for the use of vouchers: because of mental accounting, vouchers on food tilt the spending towards food, which is desirable to the extent that agents under-spend on food.

We now turn to the question of how vouchers affect redistribution. To simplify, we focus on one class of agent, the poor. Call $\boldsymbol{c}(t, b)$ the agent's consumption bundle given transfer t and voucher b. The government solves

$$\max_{t,b} \frac{[u(\boldsymbol{c}(t,b))]^{1-\sigma}}{1-\sigma} - \lambda(t+b),$$

where σ parametrizes the intensity of the preference for redistribution.

We start by imagining that the government does not have access to vouchers, so that b is constrained to be 0. Then, the indirect utility given post-tax income $w = w^* + T$ is $v(w) = Aw$ with $A = \left(\frac{\alpha_1^s}{\alpha_1}\right)^{\alpha_1} \left(\frac{\alpha_2^s}{\alpha_2}\right)^{\alpha_2} < 1$: the fact that the agent under-spends on food lowers both the average and marginal utility of income. Then, the optimal transfer solves $\max_T \frac{[A(w^*+T)]^{1-\sigma}}{1-\sigma} - \lambda T$ yielding $T = A^{\frac{1-\sigma}{\sigma}} \lambda^{\frac{1}{\sigma}} - w^*$.

We now re-introduce the voucher. The government sets the voucher as above, to reach the first best for a given total transfer $T = t + b$. Then, the agent's indirect utility becomes $v(w) = A'w$, where $A' = 1 > A$: the voucher increases both the average and the marginal utility of income. The optimal transfer is $T = (A')^{\frac{1-\sigma}{\sigma}} \lambda^{\frac{1}{\sigma}} - w^*$.

Proposition 3.7 (Internalities, vouchers and redistribution) *Optimal vouchers improve welfare above and beyond general cash transfers. If the preference for redistribution is weak ($\sigma < 1$), vouchers lead to higher overall transfers. Conversely, if the preference for redistribution is strong ($\sigma > 1$), vouchers lead to lower overall transfers.*

Vouchers have two opposing effects on the overall level of redistribution through the social marginal utility of income, which changes from $\gamma = (Aw)^{-\sigma} A$ without vouchers to $\gamma = (A'w)^{-\sigma} A'$ with vouchers. By increasing the average utility of income, vouchers reduce the average weight of the agent in social welfare from $(Aw)^{-\sigma}$ to $(A'w)^{-\sigma}$, but they also increase the marginal utility of income from A to A'. The resulting effect on the social marginal utility of income γ depends on the relative strength of these two effects.[47] When $\sigma < 1$, the latter effect dominates and vouchers lead to higher γ. The opposite occurs when $\sigma > 1$. When $\sigma = 1$, the two effects exactly cancel out.

[47] The NBER working paper version of Kaplow (2015) discusses a similar idea, in the context of a model with myopic agents.

Quantitative Illustrations To illustrate our analysis of vouchers and mental accounts, we draw from Hastings and Shapiro (2018), who analyze empirically the effect of food vouchers ("food stamps" aka "Supplemental Nutrition Assistance Program" or SNAP). They find an MPCF out of vouchers of 0.5, which is higher than the MPCF of general transfers of 0.1. Our model can capture these facts, interpreting good 1 as food and good 2 as a composite good capturing all other goods. We take $\alpha_1^s = 0.1$, $\beta = 0.4$, and the specific value of κ_1 does not matter for the calibration, provided that it is high enough.[48] Given that overall spending on SNAP eligible items is around \$500, existing vouchers of around \$200 are infra-marginal. This leads to a transfer to income ratio $\frac{b}{w}$ of around 0.1. Independent of attitudes towards redistribution, the optimal voucher to income ratio is $\frac{b}{w} = \frac{\alpha_1 - \alpha_1^s}{\beta}$. To rationalize the current level of vouchers therefore requires a relatively modest bias of $\alpha_1 - \alpha_1^s = 0.1 \times 0.4 = 0.04$, i.e. a spending share on food that is 4% too low.

Optimal Taxation Within and Across Rigid Mental Accounts We now generalize the model and state two propositions characterizing optimal taxation within (Proposition 3.8) and across (Proposition 3.9) rigid mental accounts.

We assume that the only friction is mental accounting and we consider a representative agent. We use the following model of mental accounts. The agent perceives the budget constraint to be $\sum_k \boldsymbol{q}^k \cdot \boldsymbol{C}^k + \kappa^k |\boldsymbol{q}^k \cdot \boldsymbol{C}^k - \omega_k^d(\boldsymbol{q}, w)| \leq w$, where $\omega_k^d(\boldsymbol{q}, w)$ is an exogenous default mental accounting function, but the true budget constraint remains $\sum_k \boldsymbol{q}^k \cdot \boldsymbol{C}^k \leq w$. The idea is that there are frictions on mental accounting so that the consumer faces a psychic cost given by $\kappa^k |\boldsymbol{q}^k \cdot \boldsymbol{C}^k - \omega_k^d(\boldsymbol{q}, w)|$ when the expenditure on account k is different from the default expenditure. Let $\boldsymbol{C}^k(\boldsymbol{q}, w)$ be the corresponding demand functions. The mental accounting functions are then given by $\omega^k(\boldsymbol{q}, w) = \boldsymbol{q}^k \cdot \boldsymbol{C}^k(\boldsymbol{q}, w)$. The extended demand function is given by $\boldsymbol{c}(\boldsymbol{q}, \boldsymbol{\omega}) = \boldsymbol{c}^r(\boldsymbol{q}, \boldsymbol{\omega})$, where the latter is $\boldsymbol{c}^r(\boldsymbol{q}, \boldsymbol{\omega}) = \arg\max_{\boldsymbol{c}} u(\boldsymbol{c})$ s.t. $\boldsymbol{q}^k \cdot \boldsymbol{C}^k = \omega^k$ for $k = 1, ..., K$.

In the interest of space, we focus in the main text on the case of rigid mental accounts, by which we mean the following: a mental account k is *rigid* when the amount $\omega^k(\boldsymbol{q}, w)$ allocated to account k is independent of $\boldsymbol{q}$.[49] This will arise when the default $\omega_k^d(\boldsymbol{q}, w)$ is independent of $\boldsymbol{q}$ and when κ^k is large enough so that $\omega^k(\boldsymbol{q}, w) = \omega_k^d(\boldsymbol{q}, w)$.

Proposition 3.8 (Uniform commodity taxation within a rigid mental account) *Suppose that there is just one type of agent, that mental account k is rigid, that the only taxation motive is to raise revenues, and that all commodities in this account can be taxed. Then, all commodities associated with mental account k should be taxed at the same rate.*

The intuition is that it is efficient to tax all commodities associated in a rigid mental account

[48] Matching both an MPCF out of general transfers of 0.1 and a budget share of food of 0.2 as measured by Hastings and Shapiro (2018) would require going beyond Cobb-Douglas preferences. To keep the model simple we refrain from doing that, at the cost of slightly deteriorating its fit.

[49] The online appendix (Section 9.3) develops other applications in the more general case with flexible accounts where $\omega^k(\boldsymbol{q}, w)$ depends on $\boldsymbol{q}$ and w.

at the same rate in order to avoid distorting the relative consumption of two commodities within the account.

We now turn to the structure of optimal taxes across mental accounts. We consider the basic Ramsey and Pigou setups with no misperceptions ($m_i = 1$ for all i) but with rigid mental accounts instead. We make the further simplification that there is one commodity per mental account. Consumption is therefore given by $c_i = \frac{\omega^i}{q_i} = \frac{\omega^i}{p_i+\tau_i}$. We assume that before taxes, the optimal amount ω^i is allocated to good i, so that $U^{i\prime}(\omega^i) = p_i$, and that the rigid mental account ω^i does not adjust after the introduction of taxes. The following Proposition gives the optimal taxes, in the case of isoelastic utility $u_i'(c_i) = c_i^{-1/\psi_i}$ for good i.

Proposition 3.9 (Ramsey and Pigou formulas with rigid mental accounts) *Suppose that agents use a rigid mental account for good i. In the basic Ramsey problem, the optimal ad-valorem tax is*

$$\frac{\tau_i}{p_i} = \lambda^{\psi_i} - 1, \tag{24}$$

while in the basic Pigou problem, it is

$$\frac{\tau_i}{p_i} = \left(1 + \frac{\xi}{p_i}\right)^{\psi_i} - 1. \tag{25}$$

To get further intuition, it is useful to consider the limit of small taxes ($\lambda_i = \frac{1}{1-\Lambda}$with Λ small). The formula for the Ramsey problem becomes $\frac{\tau_i}{p_i} = \Lambda\psi_i$, which is in stark contrast with the traditional Ramsey case where $\frac{\tau_i}{p_i} = \frac{\Lambda}{\psi_i}$, as well as the misperception case where $\frac{\tau_i}{p_i} = \frac{\Lambda}{m_i^2\psi_i}$. With rigid mental accounts, a low (rational) elasticity ψ_i leads to low taxes, not to high taxes, as in the basic Ramsey. The intuition is as follows. If a good is very "necessary", rational demand is very inelastic and ψ_i is low. But with a rigid mental accounts, a tax τ_i leads to a consumption $c_i = \frac{\omega^i}{p_i+\tau_i}$. So, a high tax leads to a high distortion. Hence, when (rational) demand is very inelastic, the tax should be low.

Likewise, the modified Pigou formula $\frac{\tau_i}{p_i} = \xi_i\psi_i$ now features the rational elasticity of demand ψ_i. This is in contrast to the traditional case, where $\tau_i = \xi_i$, and to the case with misperception m_i where $\tau_i = \frac{\xi_i}{m_i}$ (Proposition 3.2).

3.7 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 $\boldsymbol{m}(\boldsymbol{q}, w) = \arg\max_{\boldsymbol{m}} u^s(\boldsymbol{C}(\boldsymbol{q}, w, \boldsymbol{m}), \boldsymbol{m})$, 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 $c(q) = q^{-\psi}$.[50] 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 (13), 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.[51] We have the following.

Proposition 3.10 *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 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 $\alpha = \frac{\ln 2}{\ln 3} \simeq 0.6$.[52] 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, Looney and Kroft (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.[53]

[50] 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.

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

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

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

[53] 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

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.[54] 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.

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.[55] In the basic Ramsey model and in the limit of small taxes, optimal welfare is given by $\frac{H}{2}\sum_i \frac{\Lambda^2}{\psi_i}\frac{1}{\mathbb{E}\left[m_i^h\right]^2[1+\frac{var\left[m_i^h\right]}{\mathbb{E}\left[m_i^h\right]^2}]}y_i$ up to an additive constant (see Footnote 30). 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 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.[56]

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

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.

[54]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.

[55]One can 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.

[56]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 (7), 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 2012 and Section 9.5.2 of the online appendix).

in the online appendix (Section 10). We focus on the intensive margin of labor supply, and refer the reader to the online appendix (Section 10.3.1) 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 $u_z \leq 0$.[57] 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.[58] If the agent works too much—perhaps because he underperceives taxes (see Feldman, Katuscak and Kawano (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 $\widetilde{\tau}^b(z) = g(z)\tau^b(z)$.

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.[59] 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.

[57] 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). We discuss the impact of relaxing this assumption in Sections 5.1 and 10.3.2.

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

[59] 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, Gennaioli and Shleifer (2013), Bushong, Rabin and Schwartzstein (2017)).

Proposition 4.1 *Optimal taxes satisfy the following formulas (for all z^*)*[60]

$$\frac{T'(z^*) - \widetilde{\tau}^b(z^*)}{1 - T'(z^*)} = \frac{1}{\zeta^c(z^*)} \frac{1 - H(z^*)}{z^* h^*(z^*)} \int_{z^*}^{\infty} (1 - \gamma(z)) \frac{h(z)}{1 - H(z^*)} dz - \int_0^{\infty} \frac{\zeta^c_{Q_{z^*}}(z)}{\zeta^c(z^*)} \frac{T'(z) - \tilde{\tau}^b(z)}{1 - T'(z)} \frac{z h^*(z)}{z^* h^*(z^*)} dz. \tag{26}$$

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 (26) 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 (26)), compared to the Saez optimal tax formula. Second, suppose that $\tilde{\tau}^b(z) < 0$ (perhaps because the agent 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.

We now put this formula to use to uncover a number of concrete insights in different behavioral settings.

The Optimal Top Marginal Tax Rate We apply (26) to derive a formula for the marginal tax rate at very high incomes. To be concrete, we specialize the general model and consider a case in which the only behavioral bias is that agents are influenced by tax rates on incomes different from theirs. We assume that the perceived marginal tax rate is

$$T'^{,s}(z) = mT'(z) + (1 - m) \left[\int_0^{\infty} T'(az) \psi(a) da + b(z) T(0) \right], \tag{27}$$

[60] This formula can also be expressed as a modification of the Saez (2001) formula. The modified Saez formula (see equation (92) in the online appendix) uses the concept of the social marginal welfare weight $g(z)$ rather than the social marginal utility of income $\gamma(z)$.

with $\int \psi(a)\,da = 1$ and $\lim_{z\to\infty} b(z) = 0$. This means that the subjectively perceived marginal tax rate $T'^{,s}(z)$ is a weighted average with respective weights m and $1-m$ of: (i) the true marginal tax rate $T'(z)$; and (ii) a sum of the average of the marginal tax rates $T'(az)$ at different incomes, with weights $\psi(a)$, and of the intercept $T(0)$, with a vanishing weight.[61]

We will obtain a general formula that we will apply to two polar cases capturing two different directions of misperceptions. In the first case, we take $\psi(a) = 0$ for $a < 1$ and $b(z) = 0$, so that agents are only influenced by incomes higher than theirs. One motivation is that people might be overconfident about their probability of achieving high incomes, as they are optimistic about mobility in general (as in Bénabou and Tirole (2006*a*); Alesina, Stantcheva and Teso (2018)). Another might be that the top rates are very salient.[62] In the second case, we take $\psi(a) = 1_{a\leq 1}$ and $b(z) = \frac{1}{z}$. Then, we recover the schmeduling case of Liebman and Zeckhauser (2004) and Rees-Jones and Taubinsky (2019), in which one's perceived marginal tax rate is a weighted average of one's true marginal tax rate (with weight m) and of one's average tax rate (with weight $1-m$).[63]

We proceed like Saez (2001) and assume that for very large incomes the various elasticities converge. We denote by $\bar{\zeta}^{c,r}$ the rational elasticity of labor supply (positive), $\bar{\eta}^r$ the rational labor income elasticity (negative if leisure is a normal good), and $\bar{g}$ the social welfare weight—all being asymptotic for large incomes.[64] The earnings distribution is asymptotically Pareto with exponent π (i.e. when z is large, $1 - H(z) \propto z^{-\pi}$).

Proposition 4.2 (Optimal tax rate for top incomes) *The optimal marginal rate $\overline{\tau}$ for top incomes is*

$$\overline{\tau} = \frac{1-\bar{g}}{1-\bar{g}+\bar{\eta}^r + \bar{\zeta}^{c,r}\pi\left(m+(1-m)A\right)}, \tag{28}$$

where $1-m$ and $A = \int_0^\infty a^{\pi-1}\psi(a)\,da$ index the degree of misperception of taxes (as in equation (27)). Hence when agents are more behavioral (i.e. when m is lower), then the optimal top marginal tax rate is: (i) lower when agents are overinfluenced by higher incomes so that $A > 1$ (e.g. because of overconfidence); (ii) higher when agents are overinfluenced by lower incomes so that $A < 1$ (e.g. because of schmeduling). With rational agents ($m = 1$) we recover the rational Saez (2001) formula.

The proof (detailed in the online appendix) is a direct application of the optimal tax formula (26), using the fact that $\zeta^c(z) = m\zeta^{c,r}(z)$, that $\zeta^c_{Q_{z^*}}(z) = (1-m)\frac{\psi(z^*/z)}{z}\zeta^{c,r}(z)$, that $\bar{\eta} = \bar{\eta}^r$, and that $\tilde{\tau}^b$ tends to 0 for high incomes.

[61] As before when dealing with misperceived prices, the behavioral first-order condition of an agent with wage n earning z in equilibrium is: $n(1-T'^{,s}(z))\,u_c(c,L) + u_L(c,L) = 0$ with $(c,L) = \left(z - T(z), \frac{z}{n}\right)$.

[62] Concretely, think of the recent case of France where increasing the top rate to 75 percent might have created an adverse general climate with the perception that even earners below the top income would pay higher taxes. Relatedly, people overestimate the probability that they will be subjected to the estate tax (Slemrod (2006)).

[63] Indeed, $\int_0^\infty T'(az)\,\psi(a)\,da + \frac{T(0)}{z} = \frac{T(z)}{z}$ is the average tax rate.

[64] These asymptotic elasticities are well defined for popular utility functions of the form $U(c,L) = \bar{U}\left(\frac{c^{1-\gamma}-1}{1-\gamma} - \kappa L^{1+1/\psi}\right)$ for which we get $\bar{\eta}^r = -\gamma\psi$ and $\bar{\zeta}^{c,r} = \psi$.

As a numerical example, we use the Saez calibration with $\bar{\zeta}^c = 0.2$, $\bar{g} = \bar{\eta}^r = 0$ and $\pi = 2$. Then, in the rational case ($m = 1$), we recover the Saez optimal tax rate $\bar{\tau} = 0.71$. For the case where agents are over-influenced by higher incomes, we use $\psi(a) = \xi a^{-\xi-1} 1_{a \geq 1}$ with $\xi = 1.5$, so that the very rich matter more than their empirical frequency (since $\xi < \pi$), perhaps because they are more frequently talked about in the media. We are not aware of attempts at estimating the behavioral parameters m and ξ, and so we explore different values of m. If $m = 0.6$, then $\bar{\tau} = 0.58$; if $m = 0.4$, then $\bar{\tau} = 0.53$. For the "schmeduling" case, if we use the value of $m = 0.6$ estimated by Rees-Jones and Taubinsky (2019), then $\bar{\tau} = 0.76$.

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, Friedman and Saez (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.[65]

5 Revisiting Diamond-Mirrlees and Atkinson-Stiglitz

To complete our tour of behavioral version of classic taxation theory, we now revisit two classical public finance results: the Diamond and Mirrlees (1971) production efficiency result and the asso-

[65] 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 10.3.1) for a behavioral treatment of the Saez (2002) extensive margin of labor supply model.

ciated result that supply elasticities do not enter in optimal tax formulas, as well as the Atkinson and Stiglitz (1972) uniform commodity taxation result.

5.1 Diamond-Mirrlees (1971)

Supply Elasticities: Optimal Taxes with Efficient Production So far, we have assumed a perfectly elastic production function (constant production prices $\boldsymbol{p}$). In traditional, non-behavioral models, this is without loss of generality. Indeed, with a complete set of commodity taxes, optimal tax formulas depend only on production prices but not on production elasticities. In behavioral models, this result must be qualified. This section therefore generalizes the model to imperfectly elastic production function (non-constant prices $\boldsymbol{p}$).

In behavioral models, prices $\boldsymbol{p}$ and taxes $\boldsymbol{\tau}$ might affect behavior differently. We introduce a distinction between taxes $\boldsymbol{\tau^p}$, that affect behavior like prices, and taxes, $\boldsymbol{\tau}^c$ that affect behavior different from prices. For example, $\boldsymbol{\tau^p}$ could represent taxes that are included in listed prices $\boldsymbol{p}+\boldsymbol{\tau^p}$ (either because they are levied on producers or because they are levied on consumers but the listed prices are inclusive of the tax) and taxes $\boldsymbol{\tau}^c$ that are not included in listed prices. An agent's demand function can then be written as $\boldsymbol{c}^h\left(\boldsymbol{p}+\boldsymbol{\tau^p},\boldsymbol{\tau}^c,w\right)$. This distinction will prove to be important for the generalization of our results to imperfectly elastic production functions.

We denote the associated indirect utility function by $v^h\left(\boldsymbol{p}+\boldsymbol{\tau^p},\boldsymbol{\tau}^c,w\right)$ and the Slutsky matrices corresponding to $\boldsymbol{\tau}^p$ (or $\boldsymbol{p}$) and $\boldsymbol{\tau}^c$ by $\boldsymbol{S}_i^{C,\boldsymbol{p},h}$ and $\boldsymbol{S}_i^{C,c,h}$, respectively. We allow for the possibility (but we do not impose it) that these Slutsky matrices do not coincide.

We assume that the government must finance a vector of government consumption $\boldsymbol{g}$ and that profits are fully taxed—we allow for decreasing returns to scale and nonzero profits. The production set is expressed as $\{\boldsymbol{y}$ s.t. $F\left(\boldsymbol{y}\right)\leq 0\}$. Perfect competition imposes that $F\left(\boldsymbol{y}\right)=0$ and $\boldsymbol{p}=F'\left(\boldsymbol{y}\right)$, where $\boldsymbol{y}$ is the equilibrium production. Market clearing requires that $\boldsymbol{g}+\sum_h \boldsymbol{c}^h\left(\boldsymbol{p}+\boldsymbol{\tau}^p,\boldsymbol{\tau}^c,w\right)=\boldsymbol{y}$. We denote by $\bar{\boldsymbol{\tau}}=\boldsymbol{\tau}^c+\boldsymbol{\tau}^p$ the sum of the tax vectors.

We can write the planning problem as $\max_{\boldsymbol{p},\boldsymbol{\tau}^p,\boldsymbol{\tau}^c} W\left(\left(v^h\left(\boldsymbol{p}+\boldsymbol{\tau}^p,\boldsymbol{\tau}^c,w\right)\right)_{h=1\ldots H}\right)$ subject to the resource constraint $F\left(\boldsymbol{g}+\sum_h \boldsymbol{c}^h\left(\boldsymbol{p}+\boldsymbol{\tau}^p,\boldsymbol{\tau}^c,w\right)\right)=0$, and the competitive pricing condition $\boldsymbol{p}=F'\left(\boldsymbol{g}+\sum_h \boldsymbol{c}^h\left(\boldsymbol{p}+\boldsymbol{\tau}^p,\boldsymbol{\tau}^c,w\right)\right)$.

The competitive pricing equation is a fixed point in $\boldsymbol{p}$. We denote the solution by $\boldsymbol{p}\left(\boldsymbol{\tau}^p,\boldsymbol{\tau}^c,w\right)$. The derivatives of this function $\boldsymbol{p}$ encapsulate the incidence of taxes depending on the demand and supply elasticities. We define the price derivative matrix by $\varepsilon_{ij}^{\kappa}=\frac{\partial p_i}{\partial \tau_j^{\kappa}}$, the derivative of the prices p_i of commodity i with respect to the tax τ_j^{κ} with $\kappa\in\{p,c\}$. We also define the supply elasticity matrix ε_S by $\varepsilon_{S,ij}=\frac{p_j}{y_i}\left(F''^{-1}\right)_{ij}$ and the demand elasticities ε_D^{κ} by $\varepsilon_{D,ij}^{\kappa}=-\frac{1}{y_i}\sum_h p_j \boldsymbol{c}_{i,\tau_j^{\kappa}}^h$. Finally we define the matrix $diag\left(\boldsymbol{p}\right)$ as the diagonal matrix with i-th element given by p_i. Then, by applying the implicit function theorem to the competitive pricing condition, we obtain after some manipulations that the matrix of price derivatives ε^{κ} is given by $\varepsilon^{\kappa}=-diag(\boldsymbol{p})\left(\varepsilon_S+\varepsilon_D^p\right)^{-1}\varepsilon_D^{\kappa}diag\left(\boldsymbol{p}\right)^{-1}$ so that the ε^{κ} reflects both demand and supply elasticities. This formula is the behavioral extension of the

standard incidence calculations determining how the burden of taxes is shared between consumers and producers. Compared with the traditional model without behavioral biases, the difference is that ε_D^κ depends on the salience of the tax instrument κ. Incidence ε^κ therefore depends on salience (and more generally on how taxes are perceived). This conceptual point already appears in Chetty, Looney and Kroft (2009). Our incidence formula only generalizes it to many goods and arbitrary preferences.

We replace $\boldsymbol{p}$ in the objective function and the resources constraint, and we put a multiplier λ on the resource constraint. We form the Lagrangian

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

The optimal tax formulas can be written as $L_{\tau_i^\kappa} = 0$ for $\kappa \in \{p, c\}$ if tax τ_i^κ is available.

Proposition 5.1 (Impact of production elasticities) *With an imperfectly elastic production function, the optimal tax formula* (7) *can be written as*

$$0 = \sum_h \left[\left(\lambda - \gamma^h\right) c_i^h + \lambda(\bar{\boldsymbol{\tau}} - \widetilde{\boldsymbol{\tau}}^{b,h}) \cdot \boldsymbol{S}_i^{C,\kappa,h}\right] + \sum_h \sum_j \left[\left(\lambda - \gamma^h\right) c_j^h + \lambda(\bar{\boldsymbol{\tau}} - \widetilde{\boldsymbol{\tau}}^{b,h}) \cdot \boldsymbol{S}_j^{C,p,h}\right] \varepsilon_{ji}^\kappa, \quad (29)$$

which depends on production elasticities and does not coincide with those of Sections 2 or 3.7. However, if there is a full set of commodity taxes $\boldsymbol{\tau}^p$, the second term in equation (29) *becomes 0 and taxes are independent of production elasticities and coincide with those of Section 2 if taxes are restricted to be of the $\boldsymbol{\tau}^p$ type, or with those of Section 3.7 if taxes can be both of the $\boldsymbol{\tau}^p$ type and the $\boldsymbol{\tau}^c$ type.*

Therefore, with behavioral agents, the principle from traditional models that supply elasticities do not enter in optimal tax formulas as long as there is a full set of commodity taxes extends if taxes are understood to be of the $\boldsymbol{\tau}^p$ form. The difference is that even with a full set of commodity taxes of the $\boldsymbol{\tau}^c$ type (which would be enough to guarantee that optimal tax formulas do not depend on supply elasticities in the traditional model), optimal tax formulas do depend on supply elasticities if there is only a restricted set of commodity taxes of the $\boldsymbol{\tau}^p$ form.

Productive Inefficiency at the Optimum A canonical result in public finance Diamond and Mirrlees (1971) shows that there is production efficiency at the optimum if there is a complete set of commodity taxes, along with either constant returns or fully taxed profits. We show that this result can fail even when the planner has a full set of commodity taxes of the $\boldsymbol{\tau}^c$ type (which would be enough to guaranty production efficiency in the traditional model), as long as there is not a full set of commodity taxes of the $\boldsymbol{\tau}^p$ type.

We start by considering the case where there is a full set of commodity taxes of the $\boldsymbol{\tau}^p$ type and show that production efficiency holds under some extra conditions. We denote by $\boldsymbol{q} = \boldsymbol{p} + \boldsymbol{\tau}^p$. We fol-

low Diamond and Mirrlees (1971) and establish that production efficiency holds by assuming that the planner can control production, showing that the planner chooses an optimum on the frontier of the production possibility set. The corresponding planning problem is $\max_{\boldsymbol{q},\boldsymbol{\tau}^c} W\left(\left(v^h\left(\boldsymbol{q},\boldsymbol{\tau}^c,w\right)\right)_{h=1...H}\right)$ subject to the resource constraint $F\left(\boldsymbol{g}+\sum_h \boldsymbol{c}^h\left(\boldsymbol{q},\boldsymbol{\tau}^c,w\right)\right)\leq 0$.

Proposition 5.2 *With a full set of commodity taxes $\boldsymbol{\tau}^p$, production efficiency holds if either: (i) there are lump sum taxes and for all $\boldsymbol{q},\boldsymbol{\tau}^c$ and w, $v_w^h\left(\boldsymbol{q},\boldsymbol{\tau}^c,w\right)\geq 0$ for all h with a strict inequality for some h ; or (ii) for all $\boldsymbol{q}$, $\boldsymbol{\tau}^c$and w, there exists a good i with $v_{q_i}^h\left(\boldsymbol{q},\boldsymbol{\tau}^c,w\right)\leq 0$ for all h with a strict inequality for some h.*

The proof is almost identical to the original proof of Diamond and Mirrlees (1971). Note however that the conditions $v_w^h\left(\boldsymbol{q},\boldsymbol{\tau}^c,w\right)>0$ or $v_{\boldsymbol{q}}^h\left(\boldsymbol{q},\boldsymbol{\tau}^c,w\right)<0$ can more easily be violated in the behavioral model than in the traditional model without behavioral biases. Indeed, when agents misoptimize, it is entirely possible that the marginal utility of income be negative $v_w^h\left(\boldsymbol{q},\boldsymbol{\tau}^c,w\right)<0$. Loosely speaking, this happens if mistakes get worse as income increases. Similarly, it is entirely possible that $v_{q_i}^h\left(\boldsymbol{q},\boldsymbol{\tau}^c,w\right)>0$ for all i, since Roy's identity does not hold ($\frac{v_{q_i}}{v_w}\neq -c_i$). Failures of production efficiency could then arise even with a full set of commodity taxes $\boldsymbol{\tau}^p$. In the interest of space, we do not explore these conditions any further.

5.2 Atkinson-Stiglitz (1976): Direct vs. Indirect Taxation

Atkinson and Stiglitz (1976) show that under some separability conditions, indirect commodity taxation is superfluous in the presence of a flexible direct nonlinear income tax. With behavioral agents, we show that this condition is no longer sufficient and that in general, differential indirect commodity taxation is part of the optimum.

Formally, the setup extends that of Section 4.1. There are n_g taxable goods $c_1,...,c_{n_g}$ and one non-taxable good, leisure. For simplicity, we consider the special case where behavioral bias can be captured by a decision vs. experienced utility framework. Experienced utility is $u^n\left(c,z\right)$ where $c=V(c_1,...,c_{n_g})$ is a scalar consumption aggregator, and z is pre-tax income. Decision utility is $u^{s,n}\left(V^s\left(c_1,...,c_{n_g}\right),z\right)$, differing from experienced utility in two ways: $u^{s,n}$ vs. u^n and V^s vs. V.

Proposition 5.3 *(Direct vs. Indirect Taxation). Consider the decision vs. experienced utility model outlined above, and assume that the available tax instruments are linear commodity taxes on taxable goods $\tau_1,...,\tau_{n_g}$ and a nonlinear income tax $T\left(z\right)$. Then, if $V^s=V$, the optimum can be implemented with zero commodity taxation, but, if $V^s\neq V$, then this is not true in general.*[66]

The interpretation is plain: if people smoke too many cigarettes, then it is optimal to tax cigarettes more than the other goods, even if there is a nonlinear income tax. We could generalize

[66]It is well-understood that uniform ad valorem (percentage) commodity taxation is equivalent to zero commodity taxation with a rescaled non-linear income tax. The Atkinson-Stiglitz result can therefore alternatively be interpreted as a uniform commodity taxation result, and our behavioral extension as a non-uniform commodity taxation result.

the model to allow for the consumption aggregator $V^{s,n}$ in decision utility to depend on the agent type n.[67]

5.3 Conclusion

We have generalized the main results of the traditional theory of optimal taxation to allow for large class of behavioral biases. Natural extensions 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 plan to develop these issues in future work.

References

n.d..

Abaluck, Jason, and Jonathan Gruber. 2011. "Heterogeneity in Choice Inconsistencies Among the Elderly: Evidence from Prescription Drug Plan Choice." *The American Economic Review: Papers and Proceedings*, 101(3): 377–381.

Alesina, Alberto, Stefanie Stantcheva, and Edoardo Teso. 2018. "Intergenerational mobility and preferences for redistribution." *The American Economic Review*, 108(2): 521–54.

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

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

Allcott, Hunt, Benjamin B Lockwood, and Dmitry Taubinsky. 2019. "Regressive Sin Taxes, with an Application to the Optimal Soda Tax." *Quarterly Journal of Economics*, 134(3): 1557–1626.

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

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

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

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

[67] See Christian Moser and Pedro Olea de Souza e Silva (2019) for an analysis in the context of retirement savings with present bias.

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

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

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

Bénabou, Roland, and Jean Tirole. 2006*a*. "Belief in a just world and redistributive politics." *The Quarterly journal of economics*, 121(2): 699–746.

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

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

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

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

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

Bushong, Benjamin, Matthew Rabin, and Joshua Schwartzstein. 2017. "A Model of Relative Thinking." http://www.hbs.edu/faculty/Pages/download.aspx?name=RelativeThinking.pdf.

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

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

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

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

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

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

Chetty, Raj, John N. Friedman, and Emmanuel Saez. 2013. "Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings." *The American*

Economic Review, 103(7): 2683–2721.

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

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

Dávila, Eduardo. 2017. "Optimal Financial Transaction Taxes." http://www.eduardodavila.com/research/davila_optimal_ftt.pdf.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hansen, Lars Peter, and Thomas J Sargent. 2007. "Recursive Robust Estimation and Control without Commitment." *Journal of Economic Theory*, 136(1): 1–27.

Hastings, Justine, and Jesse M Shapiro. 2018. "How are SNAP benefits spent? Evidence from a retail panel." *The American Economic Review*, 108(12): 3493–3540.

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

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

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

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

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

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

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

Lewis, C.S. 1970. *God in the Dock: Essays on Theology and Ethics.* Eerdmans.

Liebman, Jeffrey B., and Richard J. Zeckhauser. 2004. "Schmeduling." https://scholar.harvard.edu/files/jeffreyliebman/files/Schmeduling_WorkingPaper.pdf.

Lockwood, Benjamin. 2017. "Optimal Income Taxation with Present Bias." https://benlockwood.com/papers/Lockwood_IncomeTaxationWithPresentBias.pdf.

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

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

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

Moser, Christian, and Pedro Olea de Souza e Silva. 2019. "Optimal Paternalistic Savings Policies." *Columbia Business School Research Paper*, , (17-51).

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

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

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

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

Rees-Jones, Alex, and Dmitry Taubinsky. 2019. "Measuring "Schmeduling"." *Review of Economic Studies*.

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

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

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

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

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

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

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

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

Taubinsky, Dmitry, and Alex Rees-Jones. 2017. "Attention variation and welfare: theory and evidence from a tax salience experiment." *Review of Economic Studies*, 85(4): 2462–2496.

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

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

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

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

Woodford, Michael. 2012. "Inattentive Valuation and Reference-Dependent Choice." *Working Paper*.

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 after-tax perceived 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
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 problem: $\max_{\boldsymbol{c}} u(\boldsymbol{c})$ s.t. $B(\boldsymbol{c}, \boldsymbol{p}) \leq w$. When the budget constraint is linear, $B(\boldsymbol{c}, \boldsymbol{p}) = \boldsymbol{p} \cdot \boldsymbol{c}$, so that $B_{p_j} = c_j, B_{c_j} = p_j$.

The agent, whose utility is $u(\boldsymbol{c})$, may not completely maximize. Instead, his policy is described by $\boldsymbol{c}(\boldsymbol{p}, w)$, which exhausts his budget $B(\boldsymbol{c}(\boldsymbol{p}, w), \boldsymbol{p}) = w$. 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 $v(\boldsymbol{p}, w) = u(\boldsymbol{c}(\boldsymbol{p}, w))$ and the expenditure function as $e(\boldsymbol{p}, \widehat{u}) = \min_{\boldsymbol{c}} B(\boldsymbol{c}, \boldsymbol{p})$ s.t. $u(\boldsymbol{c}) \geq \widehat{u}$. This implies $v(\boldsymbol{p}, e(\boldsymbol{p}, \widehat{u})) = \widehat{u}$ (with $\widehat{u}$ a real number). Differentiating with respect to p_j, this implies

$$\frac{v_{p_j}(\boldsymbol{p}, w)}{v_w(\boldsymbol{p}, w)} = -e_{p_j}. \tag{30}$$

We define the compensated-demand based Slutsky matrix as:

$$\boldsymbol{S}_j^C(\boldsymbol{p},w)=\boldsymbol{c}_{p_j}(\boldsymbol{p},w)+\boldsymbol{c}_w(\boldsymbol{p},w)\,B_{p_j}(\boldsymbol{c},\boldsymbol{p})_{|\boldsymbol{c}=\boldsymbol{c}(\boldsymbol{p},w)}. \tag{31}$$

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

The Slutsky matrices represent how the demand changes when prices change by a small amount, and the budget is compensated to make the previous basket available, or to make the previous utility available: $\boldsymbol{S}^C(\boldsymbol{p},w)=\partial_{\boldsymbol{x}}\boldsymbol{c}(\boldsymbol{p}+\boldsymbol{x},B(\boldsymbol{c}(\boldsymbol{p},w),\boldsymbol{p}+\boldsymbol{x}))_{|\boldsymbol{x}=0}$ and $\boldsymbol{S}^H(\boldsymbol{p},w)=\partial_{\boldsymbol{x}}\boldsymbol{c}(\boldsymbol{p}+\boldsymbol{x},e(\boldsymbol{p}+\boldsymbol{x},v(\boldsymbol{p},w)))_{|\boldsymbol{x}=0}$ i.e., using (30),

$$\boldsymbol{S}_j^H(\boldsymbol{p},w)=\boldsymbol{c}_{p_j}(\boldsymbol{p},w)-\boldsymbol{c}_w(\boldsymbol{p},w)\frac{v_{p_j}(\boldsymbol{p},w)}{v_w(\boldsymbol{p},w)}. \tag{32}$$

In the traditional model, $\boldsymbol{S}^C=\boldsymbol{S}^H$, but we shall see that this won't be the case in general. [68]

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

$$B_{\boldsymbol{c}}\cdot\boldsymbol{c}_w=1,\qquad B_{\boldsymbol{c}}\cdot\boldsymbol{c}_{p_i}=-B_{p_i},\qquad u_{\boldsymbol{c}}\cdot\boldsymbol{c}_w=v_w. \tag{33}$$

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

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

$$\frac{v_{p_j}(\boldsymbol{p},w)}{v_w(\boldsymbol{p},w)}=-B_{p_j}(\boldsymbol{c}(\boldsymbol{p},w),\boldsymbol{p})+D_j(\boldsymbol{p},w), \tag{34}$$

where

$$D_j(\boldsymbol{p},w)=-\boldsymbol{\tau}^b(\boldsymbol{p},w)\cdot\boldsymbol{c}_{p_j}(\boldsymbol{p},w)=-\boldsymbol{\tau}^b\cdot\boldsymbol{S}_j^H=-\boldsymbol{\tau}^b\cdot\boldsymbol{S}_j^C, \tag{35}$$

and the behavioral wedge is defined to be

$$\boldsymbol{\tau}^b(\boldsymbol{p},w)=B_{\boldsymbol{c}}(\boldsymbol{c}(\boldsymbol{p},w),\boldsymbol{p})-\frac{u_{\boldsymbol{c}}(\boldsymbol{c}(\boldsymbol{p},w))}{v_w(\boldsymbol{p},w)}. \tag{36}$$

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

Proof: Relations (33) imply: $\boldsymbol{\tau}^{\boldsymbol{b}}\cdot\boldsymbol{c}_w=\left(B_{\boldsymbol{c}}-\frac{u_{\boldsymbol{c}}}{v_w}\right)\boldsymbol{c}_w=1-1=0$. Next, we differentiate

[68] See (n.d.) for a recent study of Slutsky matrices with behavioral models.

$v(\boldsymbol{p}, w) = u(\boldsymbol{c}(\boldsymbol{p}, w))$

$$\frac{v_{p_i}}{v_w} = \frac{u_{\boldsymbol{c}}\boldsymbol{c}_{p_i}}{v_w} = \frac{(u_{\boldsymbol{c}} - v_w B_{\boldsymbol{c}} + v_w B_{\boldsymbol{c}})\,\boldsymbol{c}_{p_i}}{v_w} = \frac{(u_{\boldsymbol{c}} - v_w B_{\boldsymbol{c}})\,\boldsymbol{c}_{p_i}}{v_w} - B_{p_i} \text{ as } B_{\boldsymbol{c}} \cdot \boldsymbol{c}_{p_i} = -B_{p_i} \text{ from (33)}$$

$$= -\boldsymbol{\tau}^b \cdot \boldsymbol{c}_{p_i} - B_{p_i}. \tag{37}$$

Next,

$$\begin{aligned} D_j = -\boldsymbol{\tau}^b \cdot \boldsymbol{c}_{p_j} &= -\boldsymbol{\tau}^b \cdot \left(\boldsymbol{S}_j^H + \boldsymbol{c}_w(\boldsymbol{p}, w) \frac{v_{p_j}(\boldsymbol{p}, w)}{v_w(\boldsymbol{p}, w)} \right) \text{ by (32)} \\ &= -\boldsymbol{\tau}^b \cdot \boldsymbol{S}_j^H \text{ as } \boldsymbol{\tau}^b \cdot \boldsymbol{c}_w = 0. \end{aligned} \tag{38}$$

Likewise, (31) gives, using again $\boldsymbol{\tau}^b \cdot \boldsymbol{c}_w = 0$: $D_j = -\boldsymbol{\tau}^b \cdot \boldsymbol{c}_{p_j} = -\boldsymbol{\tau}^b \cdot \left(\boldsymbol{S}_j^C - \boldsymbol{c}_w B_{p_j} \right) = -\boldsymbol{\tau}^b \cdot \boldsymbol{S}_j^C$. □

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

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

and

$$\boldsymbol{S}_j^C - \boldsymbol{S}_j^H = \boldsymbol{c}_w D_j = -\boldsymbol{c}_w \left(\boldsymbol{\tau}^b \cdot \boldsymbol{S}_j^H \right). \tag{39}$$

Proof.

$$\begin{aligned} \boldsymbol{c}_{p_j} &= \boldsymbol{c}_w \frac{v_{p_j}(\boldsymbol{p}, w)}{v_w(\boldsymbol{p}, w)} + \boldsymbol{S}_j^H \text{ by (32)} \\ &= \boldsymbol{c}_w \left(-B_{p_j} + D_j \right) + \boldsymbol{S}_j^H \text{ by Proposition 7.1.} \end{aligned}$$

Also, (31) gives: $\boldsymbol{c}_{p_j} = -\boldsymbol{c}_w B_{p_j} + \boldsymbol{S}_j^C$.□

Lemma 7.1 *We have*

$$B_{\boldsymbol{c}} \cdot \boldsymbol{S}_j^C = 0, \qquad B_{\boldsymbol{c}} \cdot \boldsymbol{S}_j^H = -D_j. \tag{40}$$

Proof Relations (33) imply $B_{\boldsymbol{c}} \cdot \boldsymbol{S}_j^C = B_{\boldsymbol{c}} \cdot \left(\boldsymbol{c}_{p_j} + \boldsymbol{c}_w B_{p_j} \right) = -B_{p_j} + B_{p_j} = 0$. 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-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 $\boldsymbol{c}(\boldsymbol{p}, w) = \arg\max_{\boldsymbol{c}} u^s(\boldsymbol{c})$ s.t. $B(\boldsymbol{p}, \boldsymbol{c}) \leq w$.

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

Proposition 7.3 *In the decision-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:*

$$\boldsymbol{\tau}^b = \frac{u^s_{\boldsymbol{c}}(\boldsymbol{c}(\boldsymbol{p}, w))}{v^s_w(\boldsymbol{p}, w)} - \frac{u_{\boldsymbol{c}}(\boldsymbol{c}(\boldsymbol{p}, w))}{v_w(\boldsymbol{p}, w)}.$$

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

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

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

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

$$\boldsymbol{S}_j^H(\boldsymbol{p}, w) = \boldsymbol{S}^r(\boldsymbol{p}, w) \cdot \boldsymbol{p}^s_{p_j}(\boldsymbol{p}, w), \tag{41}$$

i.e. $S^H_{ij} = \sum_k S^r_{ik} \frac{\partial p^s_k(\boldsymbol{p}, w)}{\partial p_j}$, where $\frac{\partial p^s_k(\boldsymbol{p}, w)}{\partial p_j}$ is the matrix of perception impacts. Also

$$\boldsymbol{\tau}^b = B_{\boldsymbol{c}}(\boldsymbol{c}, \boldsymbol{p}) - \frac{B_{\boldsymbol{c}}(\boldsymbol{c}, \boldsymbol{p}^s)}{B_{\boldsymbol{c}}(\boldsymbol{c}, \boldsymbol{p}^s) \cdot \boldsymbol{c}_w(\boldsymbol{p}, w)}. \tag{42}$$

Given $B_{\boldsymbol{c}}(\boldsymbol{p}^s, \boldsymbol{c}) \cdot \boldsymbol{S}_j^H = 0$, we have:

$$D_j = (B_{\boldsymbol{c}}(\boldsymbol{p}, \boldsymbol{c}) - B_{\boldsymbol{c}}(\boldsymbol{p}^s, \boldsymbol{c})) \cdot \boldsymbol{S}_j^H = B_{\boldsymbol{c}}(\boldsymbol{p}, \boldsymbol{c}) \cdot \boldsymbol{S}_j^H, \tag{43}$$

so that

$$D_j = \overline{\boldsymbol{\tau}}^b \cdot \boldsymbol{S}_j^H \text{ with } \overline{\boldsymbol{\tau}}^b = B_{\boldsymbol{c}}(\boldsymbol{p}, \boldsymbol{c}) - B_{\boldsymbol{c}}(\boldsymbol{p}^s, \boldsymbol{c}). \tag{44}$$

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

Proof Given $\boldsymbol{c}(\boldsymbol{p},w)=\boldsymbol{h}^r\left(\boldsymbol{p}^s\left(\boldsymbol{p},w\right),v\left(\boldsymbol{p},w\right)\right)$, we have $\boldsymbol{c}_w=\boldsymbol{h}^r_u v_w$. Then,

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

Next, observe that the demand satisfies $u_{\boldsymbol{c}}\left(\boldsymbol{p},w\right)=\Lambda B_{\boldsymbol{c}}\left(\boldsymbol{p}^s,\boldsymbol{c}\right)$ for some Lagrange multiplier Λ, and that $B_{\boldsymbol{c}}\left(\boldsymbol{p}^s,\boldsymbol{c}\right)\cdot\boldsymbol{S}^r=0$ for a rational agent (see equation (40) applied to that agent). So, $B_{\boldsymbol{c}}\left(\boldsymbol{p}^s,\boldsymbol{c}\right)\cdot\boldsymbol{S}^H=0$.

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

Finally, we have $\frac{u_{\boldsymbol{c}}}{v_w}=\Lambda B_{\boldsymbol{c}}\left(\boldsymbol{c},\boldsymbol{p}^s\right)$ for some scalar $\Lambda>0$. Given (33) $\frac{u_{\boldsymbol{c}}(\boldsymbol{c}(\boldsymbol{p},w))}{v_w(v,w)}=\frac{u_{\boldsymbol{c}}}{u_{\boldsymbol{c}}\cdot\boldsymbol{c}_w}=\frac{B_{\boldsymbol{c}}(\boldsymbol{c},\boldsymbol{p}^s)}{B_{\boldsymbol{c}}(\boldsymbol{c},\boldsymbol{p}^s)\cdot\boldsymbol{c}_w}$ (indeed, both are equal to $\frac{u_{\boldsymbol{c}}}{u_{\boldsymbol{c}}\cdot\boldsymbol{c}_w}$).□

Finally, (5) comes from (39):[69]

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

8 Additional Proofs

Proof of Proposition 2.1 We have

$$\frac{\partial L}{\partial\tau_i}=\sum_h[W_{v^h}v^h_w\frac{v^h_{q_i}}{v^h_w}+\lambda c^h_i+\lambda\boldsymbol{\tau}\cdot\boldsymbol{c}^h_{q_i}].$$

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

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

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

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

[69] Another useful relation is that $u_{\boldsymbol{c}}\cdot\boldsymbol{S}^H=0$ in the (static) misperception model (this is because $u_{\boldsymbol{c}}=\Lambda B_{\boldsymbol{c}}\left(\boldsymbol{c},\boldsymbol{p}^s\right)$ for some scalar Λ, and $B_{\boldsymbol{c}}\left(\boldsymbol{c},\boldsymbol{p}^s\right)\cdot\boldsymbol{S}^H=0$ from Proposition 7.3). This is not true in the decision-utility model.

The result follows using the renormalization (6) of the behavioral wedge.

Proof of Proposition 2.3 We use the fact that $\boldsymbol{q} \cdot \boldsymbol{c}(\boldsymbol{q}, w, \chi) = w$ implies $\boldsymbol{q} \cdot \boldsymbol{c}_\chi = 0$:

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