American Economic Review 2020, 110(1): 298–336
https://doi.org/10.1257/aer.20151079

Optimal Taxation with Behavioral Agents[†]

By Emmanuel Farhi and Xavier Gabaix*

This paper develops a theory of optimal taxation with behavioral agents. We use a general framework that encompasses a wide range of biases such as misperceptions and internalities. We revisit the three pillars of optimal taxation: Ramsey (linear commodity taxation to raise revenues and redistribute), Pigou (linear commodity taxation to correct externalities), and Mirrlees (nonlinear income taxation). We show how the canonical optimal tax formulas are modified and lead to novel economic insights. We also show how to incorporate nudges in the optimal taxation framework, and jointly characterize optimal taxes and nudges. (*JEL* D62, D91, H21)

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

Our results take the form of optimal tax formulas that generalize the canonical formulas derived by Diamond (1975), Sandmo (1975), and Saez (2001). Our formulas are expressed in terms of similar sufficient statistics, which include both classical ones (social marginal utilities of income and of public funds, compensated demand elasticities, marginal externalities, and equilibrium demands) and new behavioral ones (wedges that arise from agents' misoptimization).

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

The value of our framework is three-fold. First, it unifies existing results in one single framework and identifies the key concepts that permeate many specialized behavioral public finance problems. Second, it allows to show how the forces

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

[†]Go to https://doi.org/10.1257/aer.20151079 to visit the article page for additional materials and author disclosure statements.

arising in isolation interact. Third, it delivers concrete new insights on some of the cornerstone results of public economics (of course, these results require specific assumptions, which we make explicit as we derive them). Findings (i)–(iii) have to do with limited attention, and for those we model the perceived tax as the true tax times an attention parameter (between 0 and 1).

(i) The Ramsey inverse elasticity rule states that optimal taxes to raise revenues are inversely proportional to the elasticity of demand. We show that when agents have limited attention to the tax, the Ramsey inverse elasticity rule is modified: optimal taxes increase and scale with the inverse of the square of the attention.

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

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

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

(v) A fundamental result of the Mirrlees nonlinear income tax model is that optimal marginal tax rates are weakly positive. We show that if the poor do not fully recognize the future benefits of work, perhaps because of myopia or hyperbolic discounting, then it is optimal to introduce negative marginal tax rates for low incomes. In addition, if the top marginal tax rate is particularly salient and contaminates perceptions of other marginal tax rates, then it should be lower than prescribed in the traditional analysis.

Relation to the Literature.—We rely on recent progress in behavioral public finance and basic behavioral modeling. We build on earlier behavioral public finance theory.[1] Chetty (2009) and Chetty, Looney, and Kroft (2009) analyze tax incidence

[1] Numerous studies now document inattention to prices, e.g., Abaluck and Gruber (2011); Allcott and Taubinsky (2015); Anagol and Kim (2012); Brown, Hossain, and Morgan (2010); DellaVigna (2009); and Gabaix (2019).

and welfare with misperceiving agents; however, they do not analyze optimal taxation in this context. An emphasis of previous work is on the correction of "internalities," i.e., misoptimization because of self-control or limited foresight, which can lead to optimal "sin taxes" on cigarettes or fats (Gruber and Kőszegi 2001, O'Donoghue and Rabin 2006).

Mullainathan, Schwartzstein, and Congdon (2012) offers an overview of behavioral public finance. In particular, they derive optimality conditions for linear taxes, in a framework with a binary action and a single good. Baicker, Mullainathan, and Schwartzstein (2015) further develops those ideas in the context of health care. Allcott, Mullainathan, and Taubinsky (2014) analyzes 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) studies a Mirrlees framework when agents 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 versus experienced utility model. Our behavioral Mirrlees framework is general enough to encompass, at a formal level, these models as well as many others relying on alternative behavioral biases.

We also take advantage of recent advances in behavioral modeling. We use a general framework that reflects previous analyses, including misperceptions and internalities. When modeling consumer demand with inattention to prices, we rely on part of the framework in Gabaix (2014), which builds on the burgeoning literature on inattention (Bordalo, Gennaioli, and Shleifer 2013; Caplin and Dean 2015; Chetty, Looney, and Kroft 2009; Gabaix 2019; Gabaix and Laibson 2006; Khaw, Li, and Woodford 2017; Kőszegi and Szeidl 2013; Schwartzstein 2014; Sims 2003). The agent in this framework misperceives prices while respecting the budget constraint in a way that gives a tractable behavioral version of basic objects of consumer theory, e.g., the Slutsky matrix and Roy's identity. Second, we also use the "decision utility" paradigm, in which the agent maximizes the wrong utility function. We unify those two strands in a general, agnostic framework that can be particularized to various situations.

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

I. 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 such biases. Armed with these results, we then analyze the problem of optimal linear commodity taxation without externalities (Ramsey) and with externalities

(Pigou). We also propose a model of nudges 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 particularizations of the general framework.

A. *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 $\mathbf{c}(\mathbf{q}, w)$ where $\mathbf{q}$ is the price vector and w is the budget of the consumer. Both $\mathbf{c}(\mathbf{q}, w)$ and $\mathbf{q}$ are of dimension $n \times 1$, where n is the number of commodities. The demand function incorporates all the behavioral biases to which the agent might be subject (internalities, misperceptions, etc.). The only restriction that we impose on this demand function is that it exhausts the agent's budget so that $\mathbf{q} \cdot \mathbf{c}(\mathbf{q}, w) = w$. We evaluate the welfare of this agent according to a utility function $u(\mathbf{c})$, which represents the agent's true or "experienced" utility. The resulting indirect utility function given by $v(\mathbf{q}, w) = u\big(\mathbf{c}(\mathbf{q}, w)\big)$. Crucially, the demand function $\mathbf{c}(\mathbf{q}, w)$ is not assumed to result from the maximization of the utility function $u(\mathbf{c})$ subject to the budget constraint $\mathbf{q} \cdot \mathbf{c} = w$.

A central object is the behavioral wedge, defined by

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

of dimension $n \times 1$. It is the difference between the price and marginal utility vectors (expressed in a money metric, as captured by $v_w(\mathbf{q}, w)$).[2] The wedge $\boldsymbol{\tau}^b(\mathbf{q}, w)$, which equals 0 in the rational agent model, encodes the welfare effects of a marginal reduction in the consumption of different goods, expressed in a money metric.

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 $v_{q_j}(\mathbf{q}, w) dq_j$, where $v_{q_j}(\mathbf{q}, w)$ is given by the following behavioral version of Roy's identity:[3]

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

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

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

The term $\boldsymbol{\tau}^b(\mathbf{q}, w) \cdot \mathbf{S}_j^C(\mathbf{q}, w)$ in equation (2) is a new term that arises with behavioral agents. The intuition is the following: a change dq_j in the price of

[2] The behavioral wedge is independent of the particular cardinalization chosen for experienced utility.

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

good j changes welfare by $v_{q_j}(\mathbf{q},w)dq_j = u_{\mathbf{c}}\big(\mathbf{c}(\mathbf{q},w)\big)\,\mathbf{c}_{q_j}(\mathbf{q},w)dq_j$, a change which can be decomposed into an income effect $-u_{\mathbf{c}}\big(\mathbf{c}(\mathbf{q},w)\big)\,\mathbf{c}_w(\mathbf{q},w)\,c_j(\mathbf{q},w)dq_j = -v_w(\mathbf{q},w)\,c_j(\mathbf{q},w)dq_j$ and a substitution effect $u_{\mathbf{c}}\big(\mathbf{c}(\mathbf{q},w)\big)\cdot\mathbf{S}_j^C(\mathbf{q},w)dq_j$. 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. With behavioral biases, income-compensated price changes lead to welfare changes in proportion to $\boldsymbol{\tau}^b$.

To make this more concrete, imagine that the agent is a net buyer of good j ($c_j > 0$) and that we are considering an increase dq_j in the price of good j. If the agent were rational, his welfare in monetary unit would be reduced by the usual term $-c_j dq_j < 0$. Now suppose that the agent is subject to biases such that the wedge is positive for good j ($\tau_j^b > 0$) and that the other goods have zero wedges ($\tau_i^b = 0$ for $i \neq j$). In addition, assume that the usual own-elasticity sign holds ($S_{jj}^C < 0$). In this case, the usual term $-c_j dq_j < 0$ overestimates the welfare loss for the agent because he was overconsuming good j to begin with.

To put some numbers on this effect, we use an example from Gruber and Kőszegi (2004), and consider a smoker who consumes $c_j = 1$ pack of cigarettes a day. Suppose the price of a pack of cigarettes increases by 1 dollar,[4] and daily consumption goes down by $-S_{jj}^C = 0.14$ packs. The smoker overconsumes cigarettes: the corresponding internality is $\tau_j^b = 10.5$ dollars per pack.[5] Then the behavioral Roy's identity says that his utility is *improved* by $-1 + 10.5 \times 0.14 = 0.47$ dollars a day rather than reduced, because increasing the price helps the agent curb his excessive smoking.

A Concrete Model with Misperception of Prices and Utility.—We now present a concrete instantiation of the general formalism, in which agents misperceive prices and maximize the "wrong" utility. There are three primitives: an "experienced" utility function $u(\mathbf{c})$, a perceived "decision" utility $u^s(\mathbf{c})$, and a price perception function indicating the price $\mathbf{q}^s(\mathbf{q},w)$ perceived by the agent, as a function of the true price $\mathbf{q}$ and his income w (superscripts s indicate subjective perceptions).

The agent maximizes a perceived utility function $u^s(\mathbf{c})$ given perceived prices $\mathbf{q}^s(\mathbf{q},w)$, but ultimately experiences "true" utility $u(\mathbf{c})$. Given true prices $\mathbf{q}$, perceived prices $\mathbf{q}^s$, and budget w, the demand $\mathbf{c}^s(\mathbf{q},\mathbf{q}^s,w)$ is the consumption vector $\mathbf{c}$ satisfying $u_{\mathbf{c}}^s(\mathbf{c}) = \lambda^s\mathbf{q}^s$ for the value of $\lambda^s > 0$ such that the true budget

[4] The "dollar" is for ease of interpretation, as strictly speaking it only holds as a first-order approximation. The reader may prefer to think of a "cent."

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

constraint holds $\mathbf{q} \cdot \mathbf{c} = w$.[6] The primitive demand function $\mathbf{c}(\mathbf{q}, w)$ of the general model is then given by

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

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 $u^s_{ci}/u^s_{cj} = q^s_i/q^s_j$. The adjustment factor λ^s ensures that the budget constraint holds, despite the fact that agents misperceive prices. True and perceived prices, as well as perceived decision utility all influence both choices and welfare. By contrast, the experienced utility function only influences welfare but not choices.

The behavioral wedge is then given by

$$\boldsymbol{\tau}^b(\mathbf{q}, w) = \underbrace{\frac{u^s_c\big(\mathbf{c}(\mathbf{q}, w)\big)}{v^s_w(\mathbf{q}, w)} - \frac{u_c\big(\mathbf{c}(\mathbf{q}, w)\big)}{v_w(\mathbf{q}, w)}}_{\text{misperception of preferences}} + \underbrace{\mathbf{q} - \frac{\mathbf{q}^s(\mathbf{q}, w)}{\mathbf{q}^s(\mathbf{q}, w) \cdot \mathbf{c}_w(\mathbf{q}, w)}}_{\text{misperception of prices}}. \tag{3}$$

The first term is the simply the gap between decision and experienced marginal utilities. The second term is the discrepancy between true prices and (renormalized) perceived prices. Intuitively, if a good entails a negative internality, or if its price is underperceived, then the agent overconsumes it at the margin, and the corresponding behavioral wedge is positive.

To make things concrete, let there be two goods and quasilinear utility $u(c_0, c_1) = c_0 + U(c_1)$ and $u^s(c_0, c_1) = c_0 + U^s(c_1)$. Good 0 is the untaxed numéraire, the pre-tax price of good 1 is p_1, the post-tax price of good 1 is $q_1 = p_1 + \tau_1$ where τ_1 is the tax. The tax is not fully salient so that the perceived tax is $m_1 \tau_1$, where $m_1 \in [0, 1]$ is the attention to the tax, and the perceived price is $q^s_1 = p_1 + m_1 \tau_1$. In this case the behavioral wedges are $\tau^b_0 = 0$ and $\tau^b_1 = U^{s\prime}(c_1) - U'(c_1) + (1 - m_1)\tau_1$.

To derive the Slutsky matrix, we start by defining the Hicksian matrix of marginal perceptions $\mathbf{M}^H(\mathbf{q}, w)$, of dimension $n \times n$ and with elements $M^H_{ij}(\mathbf{q}, w) = \partial q^s_i(\mathbf{q}, w)/\partial q_j - \big(\partial q^s_i(\mathbf{q}, w)/\partial w\big)\big(v^s_{q_j}/v^s_w\big)$, each of which is the marginal impact of a change in true price q_j on the perceived price q^s_i. Next, we define $\mathbf{S}^r(\mathbf{q}, w)$ to be the Slutsky matrix of an agent with utility $u^s(\mathbf{c})$ who faces prices $\mathbf{q}^s(\mathbf{q}, w)$ and achieves utility $v^s(\mathbf{q}, w)$. The Slutsky matrix is given by

$$\mathbf{S}^C(\mathbf{q}, w) = \Big(\mathbf{I} - \mathbf{c}_w(\mathbf{q}, w)\big(\boldsymbol{\tau}^b(\mathbf{q}, w)\big)'\Big)\mathbf{S}^r(\mathbf{q}, w)\,\mathbf{M}^H(\mathbf{q}, w). \tag{4}$$

[6] This is the formulation advocated for in Gabaix (2014), which discusses it extensively and uses it to derive a behavioral version of classical consumer and equilibrium theory. The fixed-point problem for λ^s has a solution under the usual Inada conditions. If there are several such λ^s, we take the lowest one, which is also the utility-maximizing one. Note that λ^s is a function of $\mathbf{q}$, $\mathbf{q}^s$, and w.

Note that the Slutsky matrix is influenced by both true and perceived prices. By contrast, it is only affected by the perceived decision utility function and not by true experienced utility.

In the rest of the paper, we will consider only the case where $\mathbf{q}_w^s = \mathbf{0}$, so that $\mathbf{M}^H = \mathbf{M}$, where $\mathbf{M} = \mathbf{q}_{\mathbf{q}}^s$ is the matrix of marginal misperceptions. It shows how a change in the price q_j of good j creates a change $M_{kj}(\mathbf{q}, w) = \partial q_k^s(\mathbf{q}, w)/\partial q_j$ in the perceived price q_k^s of a generic good k. The term $\mathbf{S}^r(\mathbf{q}, w)$ encodes how this change in the perceived price changes the demand for goods. The term $\mathbf{c}_w(\mathbf{q}, w)\left(\boldsymbol{\tau}^b(\mathbf{q}, w)\right)'$ is a correction for wealth effects.

B. *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 IA. 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 $\mathbf{p}$.[7]

The government sets a tax vector $\boldsymbol{\tau}$, so that the vector of after-tax prices is $\mathbf{q} = \mathbf{p} + \boldsymbol{\tau}$. Good 0 is constrained to be untaxed: $\tau_0 = 0$.[8] We introduce a social welfare function $W(v^1, \ldots, v^H)$ and a marginal value of public funds λ. We omit the dependence of all functions on $(\mathbf{q}, w)$, unless an ambiguity arises.

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

$$L(\boldsymbol{\tau}) = W\left(\left(v^h(\mathbf{p} + \boldsymbol{\tau}, w^h)\right)_{h=1,\ldots,H}\right) + \lambda \sum_h \boldsymbol{\tau} \cdot \mathbf{c}^h(\mathbf{p} + \boldsymbol{\tau}, w^h), \tag{5}$$

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

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

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

We now characterize the optimal tax system.[9]

[7] The traditional justification for this assumption is the result, established by Diamond and Mirrlees (1971), that with a full set of commodity taxes, optimal tax formulas are independent of production elasticities. In Farhi and Gabaix (2019) we show that this result extends to environments with behavioral agents under stronger assumptions. We also show how to generalize our optimal tax formulas when these assumptions are not verified.

[8] Leisure for instance cannot be taxed. This assumption rules out the replication of lump sum taxes via uniform ad valorem taxes on all goods, which also entail no distortions since they do not change relative prices.

[9] 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.

PROPOSITION 1 (Behavioral Ramsey Formula): *If commodity i can be taxed, then at an interior optimum*

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

An intuition can be given along the following lines. The impact of a marginal increase in τ_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, $\sum_h \left(\lambda - \gamma^h\right) c_i^h d\tau_i$. If there were no changes in behavior besides income effects, then the government would reduce the utility of agent h by $\beta^h c_i^h d\tau_i$ and collect additional revenues $\left(1 - \boldsymbol{\tau} \cdot \mathbf{c}_w^h\right) c_i^h d\tau_i$, valued at $\lambda\left(1 - \boldsymbol{\tau} \cdot \mathbf{c}_w^h\right) c_i^h d\tau_i$, and leading to a total effect on the government objective of $\left(\lambda - \gamma^h\right) c_i^h d\tau_i$.

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

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

All in all, adding behavioral agents introduces the following differences. First, it modifies behavioral responses, which endogenously changes the values of β^h, γ^h, and $\mathbf{S}_i^{C,h}$.[10] Second, it leads to the new term $-\lambda \tilde{\boldsymbol{\tau}}^{b,h} \cdot \mathbf{S}_i^{C,h}$.

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

$$\frac{-\sum_{j,h} S_{ji}^{C,h} \tau_j}{c_i} = \underbrace{1 - \frac{\bar{\gamma}}{\lambda}}_{\text{raising revenues}} - \underbrace{\operatorname{cov}\left(\frac{\gamma^h}{\lambda}, \frac{H c_i^h}{c_i}\right)}_{\text{redistributing}} - \underbrace{\frac{\sum_{j,h} \tilde{\tau}_j^{b,h} S_{ji}^{C,h}}{c_i}}_{\text{correcting biases}}, \tag{8}$$

where $c_i = \sum_h c_i^h$ is total consumption of good i and $\bar{\gamma} = \left(1/H\right)\sum_h \gamma^h$ is the average social marginal utility of income. The term on the left-hand-side encodes the extent to which the consumption of good i is discouraged by the overall tax system. The first and second right-hand-side terms capture respectively the revenue raising and redistributive objectives of taxation: at the optimum, good i is more discouraged if the need for government revenues is large and if agents with low social marginal utility of income consume relatively more of good i. The last right-hand-side term captures the corrective objective of taxation: at the

[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. Behavioral biases then also modify the value of λ. See online Appendix Section VIII.B.1 for an example.

optimum, good i is more discouraged if good i and complements to i have large behavioral wedges.[11]

We can view this as a linear system of equations in the τ_js indexed by i with *endogenous* coefficients given by $\sum_{j,h} S_{ji}^{C,h}/c_i$ and *endogenous* forcing terms given by the right-hand-side of (8). Solving this system allows us to express the taxes as functions of these endogenous objects which we refer to as *sufficient statistics* since they mediate the dependence of optimal taxes on primitives of the model and of observables at the optimum. Since these observables in turn depend on taxes, one can view this mapping as a nonlinear fixed-point equation.

To be clear, the sufficient statistics must be computed at the optimum. In certain parametric models, these objects will be constant, leading to a closed-form solution for taxes. Indeed, this will be the case for many of the examples explored in Section II, which require specific functional forms (e.g., isoelastic, quasilinear), in which elasticities or key derivatives are independent of the tax. In general, however, these sufficient statistics are not constant. It would then be incorrect to use estimates obtained away from the optimum to infer optimal taxes. Instead, they can be used to test for optimality of an observed tax system, and in case of suboptimality, to determine the direction of welfare-improving marginal tax reforms. Alternatively, meta-analyses of empirical estimates of these sufficient statistics can be used to determine a plausible range for optimal taxes.[12]

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

C. *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 $u^h(\mathbf{c}^h, \xi)$, where $\xi = \xi\big((\mathbf{c}^h)_{h=1,\ldots,H}\big)$ is a one-dimensional externality (for simplicity) that depends on the consumption vectors of all agents and is therefore endogenous to the tax system. All individual functions encoding the behavior and welfare of agents now depend on the externality ξ.

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

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

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

[12] This requires the assumption that the meta-analysis be of cases in which taxes are not too far from the optimum, in the sense defined by the welfare criterion (5).

and $\xi = \xi\Big(\big(\mathbf{c}^h\big(\mathbf{p}+\boldsymbol{\tau}, w^h, \xi\big)\big)_{h=1,\ldots,H}\Big)$. We let the social marginal value of the externality be

$$\Xi = \frac{\sum_h \left[\beta^h \frac{v_\xi^h}{v_w^h} + \lambda \boldsymbol{\tau}\cdot \mathbf{c}_\xi^h\right]}{1 - \sum_h \xi_{\mathbf{c}^h}\cdot \mathbf{c}_\xi^h}.$$

This concept includes all the indirect effects of the externality on consumption and the associated effects on tax revenues (the term $\sum_h \lambda \boldsymbol{\tau}\cdot \mathbf{c}_\xi^h$ in the numerator) as well as the associated multiple round effects on the externality (the "multiplier" term encapsulated in the denominator). With this convention, Ξ is negative for a bad externality, like pollution. We call $\boldsymbol{\tau}^{\xi,h}$ the (agent-specific) Pigouvian wedge

$$\boldsymbol{\tau}^{\xi,h} = -\frac{\Xi \xi_{\mathbf{c}^h}}{\lambda}, \tag{9}$$

which represents the social dollar value of the externality created by one more unit of consumption by agent h. We finally define the externality-augmented social marginal utility of income $\gamma^{\xi,h} = \gamma^h + \Xi \xi_{\mathbf{c}^h}\cdot \mathbf{c}_w^h = \beta^h + \lambda\big(\boldsymbol{\tau} - \boldsymbol{\tau}^{\xi,h}\big)\cdot \mathbf{c}_w^h$.[13] The next proposition generalizes Proposition 1.

PROPOSITION 2 (Behavioral Pigou Formula): *If commodity i can be taxed, then at an interior optimum*

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

Formally, misoptimization and externality wedges ($\tilde{\boldsymbol{\tau}}^{b,h}$, $\boldsymbol{\tau}^{\xi,h}$) enter symmetrically in the optimal tax formula. We will see later that in some particular cases, behavioral biases can be alternatively modeled as externalities. But this is not true in general for at least two reasons: our model allows for behavior which could not be rationalized via an externality (for example, nonsymmetric Slutsky matrices $\mathbf{S}^{C,h}$), and allows for internalities mediated via prices which cannot be captured in the standard externalities treatment (for example, misperception of taxes creates internalities $\mathbf{q} - \mathbf{q}^s$).

D. *Optimal Nudges*

We turn our attention to another type of instrument with no counterpart in the traditional theory: nudges (Thaler and Sunstein 2008). The concept of nudge captures many different forms of interventions ranging from shocking pictures (for example, the picture of a cancerous lung on a pack of cigarettes), to default options (for example, in 401(k) retirement savings accounts). The goal of this section is to make an attempt at proposing a simple model that will allow us to jointly characterize optimal nudges and optimal taxes (see also Section IID for an application).

[13] As one dollar is given to the agent, his direct social utility increases by γ^h, but the extra dollar changes consumption by $\mathbf{c}_w^h$, and, hence, the total externality by $\xi_{\mathbf{c}^h}\cdot \mathbf{c}_w^h$, with a welfare impact $\Xi \xi_{\mathbf{c}^h}\cdot \mathbf{c}_w^h$.

While our model is stylized and makes particular parametric assumptions on how nudges alter marginal and absolute utilities, we think this framework is useful in order to capture the key economics of nudges within a theory of optimal taxation.

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. We acknowledge that this approach to nudges is conceptually standard, and is similar to how one might model information or advertising à la Becker and Murphy (1993). It also involves shortcuts vis-à-vis the "deep psychology" of nudges. However, it allows us to provide a unified treatment of optimal nudges, optimal taxes, and of their interactions.

The demand function $\mathbf{c}^h(\mathbf{q}, w, \chi)$ satisfies the budget constraint $\mathbf{q} \cdot \mathbf{c}^h(\mathbf{q}, w, \chi) = w$, where χ is the nudge vector. In general, a nudge may also affect the agents' utility $u^h(\mathbf{c}, \chi)$.[14] The nudge changes the perceived utility to $u^{s,h}(\mathbf{c}, \chi) = u^{s,h,*}(\mathbf{c}) - \chi \eta^h c_i$, where $u_c^{s,h,*}$ is the marginal decision utility in the absence of nudges, and $\eta^h \geq 0$ captures the nudgeability of the agent so that $\eta^h = 0$ corresponds to a non-nudgeable agent.[15] This captures that the nudge alters the perceived marginal utility of good i. 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 $u^h(\mathbf{c}, \chi) = u^h(\mathbf{c}) - \iota^h \chi c_i$.

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

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

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

PROPOSITION 3 (Optimal Nudge Formula): *At an interior optimum, nudges satisfy*

$$\frac{\partial L(\boldsymbol{\tau}, \chi)}{\partial \chi} = \sum_h \left[\lambda\big(\boldsymbol{\tau} - \boldsymbol{\tau}^{\xi,h} - \tilde{\boldsymbol{\tau}}^{b,h}\big) \cdot \mathbf{c}_\chi^h + \beta^h \frac{u_\chi^h}{v_w^h}\right] = 0. \tag{11}$$

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

Equation (11) has four terms corresponding to the potentially conflicting goals of nudges. The first term, $\lambda \boldsymbol{\tau} \cdot \mathbf{c}_\chi^h$, captures the fact that the changes in behavior induced by nudges directly change tax revenues. The second term, $-\lambda \boldsymbol{\tau}^{\xi,h} \cdot \mathbf{c}_\chi^h$, captures the fact that the changes in behavior induced by nudges affect welfare and tax revenues through their effect on externalities. The third

[14] Glaeser (2006), Loewenstein and O'Donoghue (2006), and Galle (2014) discuss the idea that nudges have a psychic cost.

[15] Here η^h depends on the cardinalization of utility. To get a unit-less parameter, we could write $\eta^h = \tilde{\eta}^h v_w(\mathbf{q}^*, w^*) q_i^*$, where stars denote a reference point. Then $\tilde{\eta}^h$ is unit-less, hence potentially portable across situations.

term, $\lambda\tilde{\boldsymbol{\tau}}^{b,h}\cdot\mathbf{c}_{\chi}^{h}$, captures the fact that the changes in behavior induced by nudges affect welfare because agents misoptimize. The fourth term, $\beta^{h}\left(u_{\chi}^{h}/v_{w}^{h}\right)$, captures the potential direct effects of nudges on utility.[16]

E. *A Simple Specification*

We lay out a particularization of the concrete model with misperception of prices and preferences of Section IA. This case will prove useful to construct many of our examples in Section II.

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

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

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

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

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

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

where $\mathbf{M}^{h}$ is a constant matrix of marginal perceptions. The corresponding perception function is $\mathbf{q}^{s,h}(\mathbf{q})=\mathbf{p}+\mathbf{M}^{h}(\mathbf{q}-\mathbf{p})$.[17]

Consumption of goods $1,\ldots,n$ is

$$\mathbf{C}^{h}(\mathbf{q},w)=\mathbf{C}^{s,h}\left(\mathbf{q}^{s,h}(\mathbf{q})\right)=\underset{\mathbf{C}}{\operatorname{argmax}}\,U^{s,h}(\mathbf{C})-\mathbf{q}^{s,h}(\mathbf{q})\cdot\mathbf{C}$$

and consumption of good 0 is $c_0(\mathbf{q},w)=w-\mathbf{q}\cdot\mathbf{C}^{h}(\mathbf{q},w)$.

The behavioral wedge is $\boldsymbol{\tau}^{b,h}=\boldsymbol{\tau}^{I,h}+\left(\mathbf{I}-\mathbf{M}^{h}\right)\boldsymbol{\tau}$, where $\boldsymbol{\tau}^{I,h}=U_{\mathbf{C}}^{s,h}(\mathbf{C})-U_{\mathbf{C}}^{h}(\mathbf{C})$.

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

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

We refer to $\boldsymbol{\tau}^{I,h}$ as the internality wedge. The externality wedge is

$$\tau^{\xi,h} = \frac{\beta^h}{\lambda}\xi.$$

Finally, the Slutsky matrix is

$$\mathbf{S}^{C,h}(\mathbf{q},w) = \mathbf{S}^{r,h}\big(\mathbf{q}^{s,h}(\mathbf{q})\big)\mathbf{M}^h, \quad \text{where} \quad \mathbf{S}^{r,h}\big(\mathbf{q}^{s,h}\big) = \frac{\partial \mathbf{C}^{s,h}\big(\mathbf{q}^{s,h}\big)}{\partial \mathbf{q}^{s,h}}.$$

The matrix $\mathbf{S}^{r,h}\big(\mathbf{q}^{s,h}\big)$ is the Slutsky matrix for an agent who correctly perceives prices.

We assume a utilitarian welfare function with exogenous Pareto weights b^h. Since utility is quasilinear, we have $\gamma^{\xi,h} = \gamma^h = \beta^h = b^h$.

Closed-form solutions can be obtained only in special cases, e.g., when utility is isoelastic or quadratic, as we shall see in Section II. Closed-form solutions can also be obtained as an approximation in the limit of small taxes often emphasized in public finance, and to which we now turn.[18]

Optimal taxes can then be derived in terms of fundamentals, as a first-order approximation:

$$\boldsymbol{\tau} \simeq \underbrace{-\Big[\sum_h \mathbf{M}^{h\prime}\mathbf{S}^{r,h}(\mathbf{p},w)\,\mathbf{M}^h\Big]^{-1}\sum_h\Big[\Big(1-\frac{b^h}{\lambda}\Big)\mathbf{C}^h(\mathbf{p},w)\Big]}_{\text{raising revenues and redistributing}} \tag{13}$$

$$+\underbrace{\Big[\sum_h \mathbf{M}^{h\prime}\mathbf{S}^{r,h}(\mathbf{p},w)\,\mathbf{M}^h\Big]^{-1}\sum_h\Big[\mathbf{M}^{h\prime}\mathbf{S}^{r,h}(\mathbf{p},w)\big(\boldsymbol{\tau}^{I,h}(\mathbf{p},w)+\boldsymbol{\tau}^{\xi,h}(\mathbf{p},w)\big)\Big]}_{\text{correcting internalities and externalities}}.$$

The objects on the right-hand side are evaluated at the zero-tax equilibrium, and can therefore be taken to be primitives (independent of taxes). This expression can be broken down in the different motives for taxation: the revenue-raising and redistributive motives (the first term on the right-hand side), and the internality-externality corrective motives (the second term on the right-hand side). In the Appendix, we derive a similar formulation for the optimal tax without assuming quasilinear utility (see equation (41)).

Equation (13) delivers some explicit comparative statics. In response to a change $d\boldsymbol{\tau}^{I,h}(\mathbf{p},w)$ in the internalities, we have the following average comparative static result (up to the third order in η):[19]

$$d\boldsymbol{\tau}'\sum_h \mathbf{M}^{h\prime}\mathbf{S}^{r,h}(\mathbf{p},w)d\boldsymbol{\tau}^{I,h}(\mathbf{p},w) \leq 0. \tag{14}$$

[18] To consider the limit of small taxes, we assume that $b^h-\lambda$ is small, and that $\boldsymbol{\tau}^{I,h}$ and $\boldsymbol{\tau}^{\xi,h}$ are small when taxes are equal to 0 (and hence that they remain small for small taxes). To be formal, we introduce a small disturbance vector $\eta = \big(\{b^h-\lambda\},\{\boldsymbol{\tau}^{I,h}\},\{\boldsymbol{\tau}^{\xi,h}\}\big)$ and we compute a Taylor expansion in η around 0. Equation (13) holds up to the second order in η.

[19] Indeed (13) gives $d\boldsymbol{\tau} = Q^{-1}d\mathbf{x}$ with $d\mathbf{x} = \sum_h \mathbf{M}^{h\prime}\mathbf{S}^{r,h}(\mathbf{p},w)d\boldsymbol{\tau}^{I,h}(\mathbf{p},w)$ and $Q = \sum \mathbf{M}^{h\prime}\mathbf{S}^{r,h}(\mathbf{p},w)\mathbf{M}^h$ is a negative definite matrix, as rational Slutsky matrices are negative semi-definite. So, $d\boldsymbol{\tau}'d\mathbf{x} = d\boldsymbol{\tau}'Qd\boldsymbol{\tau} \leq 0$, i.e., (14).

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

We can also get some explicit comparative statics with respect to the attention matrix $\mathbf{M}^h$. For example, both the revenue-raising and redistributive term and the internality-externality correcting terms increase (in absolute value) when taxes are made less salient via a proportional reduction of all the elements of all the attention matrices.

Increases in the heterogeneity of attention unrelated to the heterogeneity of internalities also tend to lower both components of the optimal tax (in absolute value). For instance, this result obtains when each agent is duplicated into two otherwise-identical agents, with respective attention $\mathbf{M}^h + \Delta\mathbf{M}^h$ and $\mathbf{M}^h - \Delta\mathbf{M}^h$, and all matrices $\mathbf{M}^h, \Delta\mathbf{M}^h$, and $\mathbf{S}^{r,h}$ are diagonal. The intuition is that higher heterogeneity in attention introduces an extra cost of taxation in the form of misallocation across consumers who do not all perceive the same post-tax price.

F. *Discussion*

Measurement.—Operationalizing our optimal tax formulas requires taking a stand on the relevant sufficient statistics, which are also required in the rational agent model, except for the behavioral wedges $\tilde{\boldsymbol{\tau}}^{b,h}$. In principle, they can be estimated with rich enough data on observed choices.[20]

As pointed out in Section IB, there are several ways to use estimates of these sufficient statistics. In general, estimates for a given tax system can be used to test for optimality of this tax system or to identify the direction of welfare-improving local tax reforms. With extra assumptions about functional forms, these sufficient statistics reflect constant deep parameters (e.g., demand elasticities for isoelastic utility functions) and then local estimates can be used to compute globally optimal tax system, as we shall see in Section II.

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 dependent on contextual factors. 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 one of estimating marginal externalities $\boldsymbol{\tau}^{\xi,h}$ to calibrate corrective Pigouvian taxes in the traditional model with no behavioral biases. 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.

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

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

(i) *Comparing Choices in Clear versus 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. Indeed, 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. Differences in choices in environments where behavioral biases are present would then allow to measure the marginal internalities $\boldsymbol{\tau}^{b,h} = \mathbf{q} - \left(u^h_{\mathbf{c}}/v^h_w\right)$.

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 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 IIA.

Another example is when agents may not fully understand the utility consequences of their choices, which can be captured with misperceived utility. For instance, Allcott and Taubinsky (2015) studies 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 $\boldsymbol{\tau}^{b,h} = \left(u^s_c/v^s_w\right) - \left(u_c/v_w\right)$.

(ii) *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.

(iii) *Structural Models.*—Finally, it is sometimes possible to use a calibrated structural model. For example, Allcott, Lockwood, and Taubinsky (2019) combines an assessment of the health consequences of soda consumption with a hyperbolic discounting model (Laibson 1997) to estimate the associated internality. See Section IID for a more detailed explanation.

Paternalism.—In our model, agents make mistakes that the government can identify, which is difficult in practice. This approach departs from the revealed preferences welfare paradigm and has elements of paternalism (Bernheim and Rangel 2009). There are several objections to this approach. Governments may not

understand agents' motives and constraints well enough; may also not be benevolent or fully optimizing; and may face political economy constraints. While we acknowledge these objections, they are beyond the scope of this paper.

Information-Based Biases.—Despite our model's generality, 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 2006b and references therein).

Lucas Critique.—A difficulty confronting all behavioral policy approaches is a form of the Lucas critique: how do the underlying biases change with policy? The empirical evidence is limited, but we try to bring it to bear when we discuss the endogeneity of attention to taxes (Section IIE). We hope that more empirical evidence on this will become available as the field of behavioral public finance develops.

II. Examples

A. *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. 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 $\Lambda = 1 - (\gamma/\lambda)$ so that a higher Λ corresponds to a higher relative benefit of raising revenues. Utility is $c_0 + \sum_{i=1}^{n}\left(c_i^{1-1/\psi_i} - 1\right)/(1 - 1/\psi_i)$. The only bias is that the agent perceives the tax τ_i as $\tau_i^s = m_i\tau_i$, where $m_i \in (0, 1]$ captures the attention to the tax.

The Ramsey planning problem is thus

$$\max_{\{\tau_i\}} \gamma \sum_{i=1}^{n}\left[\frac{\left[c_i(\tau_i)\right]^{1-1/\psi_i} - 1}{1 - 1/\psi_i} - (p_i + \tau_i)\, c_i(\tau_i)\right] + \lambda \sum_{i=1}^{n}\tau_i c_i(\tau_i), \tag{15}$$

where $c_i(\tau_i) = (p_i + m_i\tau_i)^{-\psi_i}$ is the demand of the consumer perceiving the price to be $p_i + m_i\tau_i$. The optimal tax formula can be derived either by specializing the general Ramsey formula (7) or by directly taking first-order conditions in (15).

PROPOSITION 4 (Modified Ramsey Inverse Elasticity Rule): *The optimal tax on good i is*

$$\frac{\tau_i}{p_i} = \frac{\Lambda}{m_i^2\psi_i} \cdot \frac{1}{1 + \Lambda\left(\frac{1 - m_i - 1/\psi_i}{m_i}\right)}. \tag{16}$$

When $m_i = 1$ 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 $m_i < 1$ the tax is higher. Mullainathan, Schwartzstein, and

Congdon (2012) discusses intuitively that taxes should be higher when they are underperceived, but does not derive a formal mathematical behavioral counterpart to the Ramsey inverse elasticity rule.

To gain intuition for equation (16), we consider the limit of small taxes (i.e., the small Λ limit). Up to the first order in Λ, optimal taxes are then given by the first term in equation (16). Thus, whereas the traditional Ramsey rule prescribes that $\tau_i^R/p_i = \Lambda/\psi_i$, with inattention optimal taxes are higher and equal to

$$\frac{\tau_i}{p_i} = \frac{\Lambda}{m_i^2 \psi_i}. \tag{17}$$

Loosely speaking, this is because inattention makes agents' demand less price-elastic. Given $m_i \leq 1$, the effective elasticity of the demand for good i is $m_i \psi_i$, rather than the parametric elasticity ψ_i.[21] However, a naïve application of the Ramsey rule would lead to the erroneous conclusion that $\tau_i/p_i = \Lambda/(m_i \psi_i)$ rather than $\tau_i/p_i = \Lambda/(m_i^2 \psi_i)$. The fact that τ_i should rise by even more than stated by the "naïve" formula arises because it is the *perceived* tax, and not the true tax, that should be inversely proportional to the effective demand elasticity: $\tau_i^s/p_i = \Lambda/(m_i \psi_i)$.[22]

Heterogeneity in Attention.—We now turn our attention to the case where perceptions of taxes are heterogeneous.[23]

We suppose that type h has attention m_i^h to the tax on good i. With isoelastic utility, no closed-form solution for the optimal tax is available, and so we directly place ourselves in the limit of small taxes to derive analytical insights. 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 an application of (13):[24]

$$\frac{\tau_i}{p_i} = \frac{\Lambda}{\psi_i E\left[\left(m_i^h\right)^2\right]} = \frac{\Lambda}{\psi_i\left(\left(E\left[m_i^h\right]\right)^2 + \mathrm{var}\left[m_i^h\right]\right)}, \tag{18}$$

where here and elsewhere E and var denote respectively the average and the variance computed over the different types h of agents. As we saw at the end of Section IE, controlling for average attention $E\left[m_i^h\right]$ (which determines the effective elasticity of total demand to the tax), an increase in the heterogeneity of attention $\mathrm{var}\left[m_i^h\right]$ reduces the optimal tax because it increases misallocation across consumers.

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

[22] Section IIA extends the analysis to an endogenous social cost of public funds.

[23] 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) and Taubinsky and Rees-Jones (2018).

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

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

Quantitative Illustration.—To gauge the real-world importance of these effects, we calibrate the behavioral Ramsey formula (7) with heterogeneity in misperceptions, based on the findings of Taubinsky and Rees-Jones (2018) for sales taxes. Sales taxes are not included in the tag price. To elicit their salience, Taubinsky and Rees-Jones design an online experiment and elicit the maximum tag price that agents would be willing to pay when there are no taxes or when there are standard taxes corresponding to their city of residence (in the latter case, they are not reminded what the tax rate is). In our notation, the ratio of these two prices is $1 + (\tau/p)m^h$, where p is the maximum tax price when there are no taxes (we focus on a given good, and suppress the index i). This allows the estimation of tax salience m^h.

Taubinsky and Rees-Jones (2018) finds (in their standard tax treatment) $E[m^h] = 0.25$ and $\text{var}(m^h) = 0.13$, so that heterogeneity is very large, $\text{var}(m^h)/(E[m^h])^2 = 0.13/0.25^2 = 2.1$. In our calibration, we take $\psi = 1$ (as in the Cobb-Douglas case, which is often a good benchmark for the elasticity between broad categories of goods), a two-point distribution with rational and behavioral agents to match the mean and dispersion of attention, and $\Lambda = 1.25$ percent, which is consistent with the baseline tax in their setup, at $\tau = 7.3$ percent (see online Appendix Section VIII.A.1 for details). 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.[25]

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.

B. *Basic Pigou Problem: Externalities, Internalities, and Inattention*

Dollar for Dollar Principle: A Behavioral Version.—We continue to assume a quasilinear utility function. We assume that there is only one taxed good, $n = 1$. The decision and experienced utilities of the representative agent coincide and are given by $u(c_0, c, \xi) = c_0 + U(c) - \xi$ where the negative externality that depends on the aggregate consumption of good 1 (think for example of second-hand smoke) is $\xi = \xi_* c$. Alternatively, this setup could represent an internality with decision utility $c_0 + U(c)$ and experienced utility $c_0 + U(c) - \xi_* c$.

To focus on the corrective role of taxes, we assume that $\Lambda = 0$ and that the government can rebate tax revenues lump sum to consumers. As before, we suppose that the agent perceives a fraction m of the tax. The optimal Pigouvian corrective tax (10) required to ensure that agents correctly internalize the externality/internality

[25] 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 $(E[m^h])^2 + \text{var}(m^h) \simeq 0.2$). If heterogeneity disappeared (but keeping mean attention constant), the optimal tax would be multiplied by $\left((\mathbb{E}[m^h])^2 + \text{var}(m^h)\right)/(\mathbb{E}[m^h])^2 \simeq 3$.

is $\tau = \xi_*/m$.[26] A dollar of externality must be corrected with $1/m$ dollars of tax. We record this simple modification of the "dollar for dollar" principle of traditional Pigouvian taxation, which assumes $m = 1$ and yields $\tau = \xi_*$.[27]

PROPOSITION 5 (Modified Pigou Formula): *In the basic Pigou problem with misperceptions, the optimal Pigouvian corrective tax is modified by inattention according to* $\tau = \xi_*/m$.

It is interesting to contrast this result with the modified optimal Ramsey tax (Proposition 4), for which $\tau_i/p_i = (\Lambda/\psi_i)(1/m_i^2)$ in the limit of small taxes. Partial attention m_i leads to a multiplication of the traditional tax by $1/m_i$ in the Pigou case and by $1/m_i^2$ in the Ramsey case.

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

If different consumers have heterogeneous perceptions, then Proposition 5 suggests that no uniform tax can perfectly correct all of them. Hence, heterogeneity in attention prevents the implementation of the first best.[28]

Heterogeneity.—We now explore this issue more thoroughly. We assume that there are several consumers, all with the same welfare weight $\gamma^h = \beta^h = \lambda$. Agent h maximizes $u^h(c_0^h, c^h) = c_0^h + U^h(c^h)$. The associated externality/internality is $\xi^h c^h$. To be more precise, in the internality case, $U^{s,h}(c^h) - U^h(c^h) = \xi^h c^h$, and in the externality case, the externality is $\xi = (1/H)\sum_h \xi^h c^h$. Agent h pays an attention m^h to the tax so that perceived taxes are $\tau_h^s = m^h\tau$. We specify (or approximate) utility to be quadratic, $U^h(c) = (a^h c - (1/2)c^2)/\Psi$, which implies a demand function $c^h(q^s) = a^h - \Psi q^s$.

With heterogeneous externality or attention, to reach the first best, we would need a person-specific Pigouvian tax, ξ^h/m^h. However, under our maintained assumption of a single uniform tax, the first best cannot be implemented except in the knife-edge case where ξ^h/m^h is the same across agents.

A direct application of the general behavioral Pigou formula (10) yields the optimal Pigouvian tax:

$$\tau^* = \frac{E[\xi^h m^h]}{E[(m^h)^2]} = \frac{E[\xi^h]E[m^h] + \mathrm{cov}(\xi^h, m^h)}{(E[m^h])^2 + \mathrm{var}[m^h]}. \tag{19}$$

[26] The derivation is as follows, in the externality case. We drop the h as there is just one type of agent. From (3),

$$\tilde{\tau}^b = \tau^b = q - q^s = (p + \tau) - (p + m\tau) = (1 - m)\tau,$$

while (9) gives $\tau^\xi = \xi_*$. Finally, (10) gives $\tau^\xi = \tau - \tau^b = m\tau$, i.e., $\tau = \xi_*/m$.

[27] 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.

[28] If the budget adjustment is concentrated on a "shock absorber" good with a sharply decreasing marginal utility, then we obtain another force making Pigouvian taxes more distortionary, resulting in lower optimal Pigouvian taxes. This is developed in online Appendix Section VIII.B.3.

As observed at the end of Section IE and in the Ramsey case, an increase in the heterogeneity of inattention $\text{var}(m^h)$ reduces the optimal tax. In addition, there is something new and specific to the Pigouvian setup. 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 $\text{cov}(\xi^h, m^h)$. See Allcott, Knittel, and Taubinsky (2015) for a study where subsidies to weatherization are hampered by the fact that people who benefit the most pay the least attention.

Inattention and Tax versus Quantity Regulation.—We continue to use the assumptions of heterogeneous consumers with quasilinear utilities and utilitarian government with no motives of raising revenues or redistributing. The fact that the first best is generally not achievable in the presence of heterogeneity opens up a potential role for quantity regulations. Suppose the government imposes a uniform quantity restriction, mandating $c^h = c^*$. A simple calculation reveals that the optimal quantity restriction is given by the intuitive formula $c^* = E[c^{h*}]$, where $c^{h*} = \text{argmax}_{c^h} U^h(c^h) - (p + \xi^h) c^h$ the quantity consumed by agent h at the first best.

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 6 (Pigouvian Tax versus 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*

$$\frac{1}{2\Psi}\text{var}(c^{h*}) < \Psi \frac{E[(\xi^h)^2]E[(m^h)^2] - (E[\xi^h m^h])^2}{2E[(m^h)^2]}, \tag{20}$$

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

Consider first the case with homogeneous attention ($m^h = m$). Then, the right-hand side of (20) is $\Psi(\text{var}(\xi^h)/2)$. Quantity restrictions tend to dominate taxes if heterogeneity in externalities/internalities is high compared to the heterogeneity in preferences. 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. This generalizes the results in Weitzman (1974) which provided a treatment in the case with full (and hence homogeneous) attention.

Let us turn to the case with homogeneous externalities ($\xi^h = \xi$). Then, the right-hand side of (20) is $\Psi \xi^2 \left(\text{var}(m^h)/2E[(m^h)^2]\right)$. Whether quantity restrictions dominate taxes is determined by similar principles as in the previous case, with heterogeneity now in attention instead of externalities. Heterogeneity of attention renders taxes less attractive because they introduce misallocation across consumers but do not affect the effectiveness of quantity restrictions, and this difference in

effectiveness is magnified by high elasticities of substitution. Note however a difference: with homogeneous attention the common level of attention is irrelevant, whereas with homogeneous externalities the common level of the externality is relevant and higher levels favor quantity restrictions.

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

Quantitative Illustration.—To get a sense of magnitudes, we use again the empirical findings of Taubinsky and Rees-Jones (2018) regarding the mean and dispersion of attention ($E\left[m^h\right] = 0.25$ and $\text{var}\left(m^h\right) = 0.13$). We consider the case where the internality/externality ξ is the same across agents. We saw that the optimal Pigouvian tax is $\tau^* = \xi\left(E\left[m^h\right]/\left(\left(E\left[m^h\right]\right)^2 + \text{var}\left(m^h\right)\right)\right)$. 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 $\left(\left(E\left[m^h\right]\right)^2 + \text{var}\left(m^h\right)\right)/\left(E\left[m^h\right]\right)^2 = 3$. As in the Ramsey case, the effects of attention and its heterogeneity on optimal taxes are important.

C. *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). 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 IE. We assume that $\gamma^h = \beta^h = \lambda$, so there is no revenue-raising motive and no redistribution motive. We also assume that agents are identical except for their attention to taxes.

We consider the case with $n = 2$ taxed goods (in addition to the untaxed good 0), where the consumption of good 1 features an internality/externality so that $\boldsymbol{\tau}^X = (\xi_*, 0)$ with $\xi_* > 0$. This can be generated as follows in the specialization of the general model developed in Section IE. In the externality case, we simply assume that $\xi\left(\left(\mathbf{C}^h\right)_{h=1,\ldots,H}\right) = \xi_*(1/H)\sum_h C_1^h$. In the internality case, we assume that $U^h(\mathbf{C}) = U^{s,h}(\mathbf{C}) - \xi_* C_1^h$. 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 $\mathbf{M}^h = \text{diag}\left(m_1^h, m_2^h\right)$. Goods 1 and 2 are substitutes (respectively complements) if at all points $S_{12}^r(\mathbf{q}, w) > 0$ (respectively < 0).

PROPOSITION 7 (Modified Principle of Targeting): *Suppose that the consumption of good 1 (but not good 2) entails a negative internality/externality. If agents perceive taxes correctly ($m^h = 1$ for all h), then good 1 should be taxed, but good 2 should be left untaxed: the classical principle of targeting holds. If agents' misperceptions*

of the tax on good 1 are heterogeneous ($\text{var}(m_1^h) > 0$), *and if the price of good 2 is homogeneously perceived or if the misperceptions* m_1^h *and* m_2^h *of the taxes on the two goods are not too correlated* (*i.e.*, *if* $E\left[m_2^h - \left(E\left[m_1^h m_2^h\right]/E\left[\left(m_1^h\right)^2\right]\right)m_1^h\right] > 0$), *then good 2 should be subsidized* (*respectively taxed*) *if and only if goods 1 and 2 are substitutes* (*respectively complements*).[29]

Proposition 7 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 binary consumption). The reason is that the tax on good 1 is an imperfect instrument in the presence of attention heterogeneity. Heterogeneity is key for this result. Indeed, if attention to good 1 were uniform at $m_1 > 0$, the first best could be attained by taxing good 1 with tax $\tau_1 = \xi_*/m_1$. This ceases to be true only in the knife-edge case where $m_1 = 0$.

A similar logic applies in the traditional model with no behavioral biases, but then only if the externality is heterogeneous across agents (Green and Sheshinski 1976). Our result offers an additional reason for why the principle of targeting might fail in the presence of behavioral biases: heterogeneous perceptions of corrective taxes. We believe that this new rationale is important because it applies even with homogeneous externalities, which is arguably the relevant case for one of the most pressing externalities: global warming due to the release of greenhouse gases in the atmosphere, where the externality is mediated by the total quantity of emissions, independently of the identity of their emitter.

D. *Correcting Internalities via Taxes or Nudges with Distributive Concerns*

Suppose that the poor consume "too much" sugary soda. This brings up a difficult policy trade-off. On the one hand, taxing sugary soda corrects this internality. On the other hand, taxing sugary soda redistributes away from the poor. These were the arguments regarding a recent proposal in New York City. In independent work, Allcott, Lockwood, and Taubinsky (2019) examines a related problem, in the context of a Mirrleesian income tax.[30]

To gain insights on how to balance these two conflicting objectives, we continue to use the specialization of the general model developed in Section IE. For simplicity, we assume that good 1 is solely consumed by a class of agents h^* but not by other agents $h \neq h^*$. As a concrete example, h^* could stand for "poor" and good 1 for "sugary soda." We also assume that utility is separable in good 1, $U^{s,h^*}(\mathbf{C}) = U_1^{s,h^*}(c_1) + U_2^{s,h^*}(\mathbf{C}_2)$, where $\mathbf{C}_2 = (c_i)_{i\geq 2}$ and $U^{s,h}(\mathbf{C}) = U_2^{s,h}(\mathbf{C}_2)$ for $h \neq h^*$. We assume that experienced utility for good 1 is $U_1^{h^*}(c_1) = \left(c_1^{1-1/\psi_1} - 1\right)/(1 - 1/\psi_1)$

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

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

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.

PROPOSITION 8 (*Taxation with Both Redistributive and Corrective Motives*): *Suppose that good 1 is consumed only by agent h^*, and entails an internality (captured by the behavioral/internality wedge $\tau_1^{I,h^*} = \xi^{h^*}$). Then the optimal tax on good 1 is*

$$\tau_1 = \frac{\frac{\gamma^{h^*}}{\lambda}\xi^{h^*} + \left(1 - \frac{\gamma^{h^*}}{\lambda}\right)\frac{p_1}{\psi_1}}{1 + \left(\frac{\gamma^{h^*}}{\lambda} - 1\right)\frac{1}{\psi_1}}. \tag{21}$$

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 $\left(\gamma^{h^*}/\lambda\right)\xi^{h^*}$ corresponds to the internality-corrective motive of taxes and is unambiguously positive. The second term $\left(1 - \left(\gamma^{h^*}/\lambda\right)\right)(p_1/\psi_1)$ corresponds to the redistributive objective of taxes, and is negative if the government wants to redistribute toward the agent (i.e., if $\gamma^{h^*}/\lambda > 1$). 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 (γ^{h^*}/λ close to 1), soda should be taxed. If the redistribution motive is large ($\gamma^{h^*}/\lambda \to \infty$) soda should taxed if and only if $\xi^{h^*} > p/\psi_1$, i.e., if the internality correction motive is large enough or if the demand elasticity is large enough. The former is intuitive, the latter arises because if demand is very elastic, then a given tax increase leads to a larger reduction in consumption and hence to a larger reduction in the amount of fiscal revenues extracted from the agents, thereby mitigating the associated adverse redistributive consequences.

Is It Better to Tax or to Nudge?—In this environment there is a tension between the redistributive and corrective objectives of the government. Correcting for the internality of good 1 calls for a tax, but this tax redistributes revenues away from the agents of type h^* consuming the good. In this context, a nudge is attractive because it allows the government to correct the internality without increasing the tax bill of these agents. Indeed, formula (11) shows that the optimal nudge is given by $\chi = \xi^{h^*}/\eta$, where η is the nudgeability of these agents. It perfectly corrects the internality of the agent.

The following proposition formalizes this comparison of the optimality of nudges and taxes.

PROPOSITION 9 (Optimal Nudge versus Tax): *If $\gamma^{h^*}/\lambda > 1$ and $\xi^{h^*} > \left(1 - \left(\lambda/\gamma^{h}\right)\right)(p_1/\psi_1)$, then a nudge is better than a tax. If $\gamma^{h^*}/\lambda = 1$, a tax and a nudge are equally good and each achieve the first best. If $\gamma^{h^*}/\lambda < 1$, a tax is better than a nudge.*

The intuition for this proposition is as follows. Suppose $\gamma^{h^*}/\lambda > 1$ so that the government wants to redistribute toward agents of type h^*. If the internality is strong

enough so that $\xi^{h^*} > \left(1 - \left(\lambda/\gamma^{h^*}\right)\right)(p_1/\psi_1)$, then the optimal tax τ_1 is positive as shown by (21). A nudge can always be designed to achieve the same level of consumption of good 1. Compared to the optimal tax, this nudge leaves more income to agents of type h^*. This guarantees that the optimal nudge does better than the tax. In the case $\gamma^{h^*}/\lambda < 1$ there is no conflict between the redistributive and corrective goals of the government: a tax helps achieve both goals, while a nudge only addresses the latter.[31]

E. *Endogenous Attention and Salience*

We now allow for endogenous attention to taxes and analyze its impact on optimal taxes. For conciseness, we illustrate this in the basic Ramsey case of Section IIA with just one taxed good (good 1, whose index we drop, and whose pre-tax price is p). Then, optimal attention is

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

where $c(q) = q^{-\psi}$.[32] We only present the results in the "no attention in welfare" case, i.e., when the experienced utility is $c_0 + u(c)$ rather than $c_0 + u(c) - g(m)$.

The optimal tax formula with endogenous attention takes a form similar to formula (16), the only difference being that ψ must be replaced by $\psi\big(1 + \tau\big(m'(\tau)/m(\tau)\big)\big)$ to account for the increase in the elasticity of demand arising from endogenous attention.[33] We have the following.

PROPOSITION 10: *Consider two economies. The first economy features endogenous attention with "no attention cost in welfare," and an optimal tax rate τ^* such that $m\big(\tau^*\big)$ and $m'\big(\tau^*\big)$ are strictly positive. The second economy has exogenous attention fixed at $m\big(\tau^*\big)$. Then the optimal tax in the second economy is higher than in the first one.*

A partial intuition is that consumers' demand is 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. This intuition is only partial

[31] In a model with heterogeneity in misperceptions of taxes and heterogeneity in nudgeability, both corrective taxes and nudges become imperfect instruments which generate additional misallocation. Higher heterogeneity in nudgeability (respectively misperceptions) makes nudges (respectively taxes) less desirable (see online Appendix Section VIII.A.2). In general, it is preferable to use both in conjunction. This remains true if there is more heterogeneity in income and richer instruments for redistribution.

[32] 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's 2011 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 $m\big(E\big[\tau^2\big]^{1/2}\big)$), imagining the tax as drawn from the distribution of taxes. See Gabaix (2014) for discussion of this.

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

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

because inattention not only reduces the consumption elasticity but also introduces a behavioral wedge. We now show that this result has quantitative bite.

Quantitative Illustration.—We rely again on Taubinsky and Rees-Jones (2018). 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 $\tau\big(m'(\tau)/m(\tau)\big) = \alpha$ to the tax, and find an elasticity $\alpha = \ln 2/\ln 3 \simeq 0.6$. 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 $1 + \tau\big(m'(\tau)/m(\tau)\big) \simeq 1.6$.

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) presents 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.

We investigate this question in the context of two simple examples, the basic Ramsey and Pigou models developed in Sections IIA and IIB. 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.[34] 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 5.

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. In the basic Ramsey model and in the limit of small taxes, optimal welfare is given by $(H/2)\sum_i(\Lambda^2/\psi_i)\Big(y_i/\Big(\big(E\big[m_i^h\big]\big)^2\Big[1 + \mathrm{var}\big[m_i^h\big]/\big(E\big[m_i^h\big]\big)^2\Big]\Big)\Big)$, up to an additive constant (see footnote 24). It is therefore possible for a tax system with a lower average salience $\big(E\big[m_i^{h\prime}\big]\big)^2 < \big(E\big[m_i^h\big]\big)^2$ to be dominated if it is associated with enough of an increase in attention heterogeneity $\mathrm{var}\big[m_i^{h\prime}\big]/\big(E\big[m_i^{h\prime}\big]\big)^2 > \mathrm{var}\big[m_i^h\big]/\big(E\big[m_i^h\big]\big)^2$. The same reasoning holds for the Pigou case.

[34] 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 $\tau_i' = (m_i/m_i')\tau_i > \tau_i$. 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 $\big((m_i - m_i')/m_i'\big)\tau_i c_i$. The improvement in welfare $\big((m_i - m_i')/m_i'\big)\tau_i c_i(\lambda - \gamma) > 0$ constitutes a lower bound for the welfare gains from moving to a fully optimal less salient tax system.

III. Nonlinear Income Taxation: Mirrlees Problem

A. *Setup*

We next give a behavioral version of the celebrated Mirrlees (1971) income tax problem. To help the readers, we provide here the major building blocks and intuitions. Many details are spelled out in the online Appendix (Section IX).

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 $u_z \leq 0$.[35] 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 IB) 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" $\tau^b(z) = -\big(\big(1 - T'(z)\big)u_c(c,z) + u_z(c,z)\big)/v_w$, where v_w is the marginal utility of a dollar received lump sum.[36] 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 $\tilde{\tau}^b(z) = g(z)\tau^b(z)$.

Planning Problem.—The objective of the planner is to design the tax schedule $T(z)$ in order to maximize the following objective function: $\int_0^\infty W\big(v(n)\big)f(n)\,dn + \int_0^\infty \big(z(n) - c(n)\big)f(n)\,dn$, 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 $\zeta^c_{z^*}(z)$: 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, $\zeta^c_{z^*}(z) = 0$. But this is no longer true with behavioral agents.[37] 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 $\zeta^c_{z^*}(z) > 0$.

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 $h^*(z) = \big(q(z)/\big(1 - T'(z) + \zeta^c z T''(z)\big)\big)h(z)$.

[35] If the agent's pre-tax wage is n, L is his labor supply, and utility is $U(c,L)$, then $u^n(c,z) = U\big(c, z/n\big)$. Note that this assumes that the wage is constant (normalized to 1).

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

[37] 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).

B. *Optimal Income Tax Formula*

We next present the optimal income tax formula.

PROPOSITION 11: *Optimal taxes satisfy the following formulas* (*for all* z^*):

$$\frac{T'(z^*)-\tilde{\tau}^b(z^*)}{1-T'(z^*)} = \frac{1}{\zeta^c(z^*)}\frac{1-H(z^*)}{z^*h^*(z^*)}\int_{z^*}^{\infty}\big(1-\gamma(z)\big)\frac{h(z)}{1-H(z^*)}\,dz - \int_0^{\infty}\frac{\zeta^c_{z^*}(z)}{\zeta^c(z^*)}\frac{T'(z)-\tilde{\tau}^b(z)}{1-T'(z)}\frac{zh^*(z)}{z^*h^*(z^*)}\,dz. \tag{22}$$

The first term $\big(1/\zeta^c(z^*)\big)\big((1-H(z^*))/z^*h^*(z^*)\big)\int_{z^*}^{\infty}(1-\gamma(z))\big(h(z)/(1-H(z^*))\big)dz$ on the right-hand side of the optimal tax formula (22) is a simple reformulation of Saez's formula. The second term $-\big(1/z^*\big)\int_0^{\infty}\big(\zeta^c_{z^*}(z)/\zeta^c(z^*)\big) \times \big(T'(z)-\tilde{\tau}^b(z)\big)/\big(1-T'(z)\big)\,z\big(h^*(z)/h^*(z^*)\big)\,dz$ on the right-hand side is new and, together with the term $-\tilde{\tau}^b(z^*)/\big(1-T'(z^*)\big)$ on the left-hand side, captures misoptimization effects.[38]

The intuition is as follows. First, suppose that $\zeta^c_{z^*}(z) > 0$. 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 toward a lower tax rate (hence the minus sign in front of the last integral in (22)), 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 toward a lower tax rate.

C. *Implications*

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 (22) 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

$$T'^{,s}(z) = mT'(z) + (1-m)\Big[\int_0^{\infty}T'(az)\psi(a)\,da + b(z)T(0)\Big], \tag{23}$$

[38] As usual, these objects are endogenous to the tax schedule and so the solution must be found as a fixed point of formula (22). This is perhaps particularly true of the behavioral cross-influence $\zeta^c_{z^*}(z)$. However, we will see in the next section that this term simplifies for a class of misperceptions.

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.[39]

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 2006a; Alesina, Stantcheva, and Teso 2018). Another might be that the top rates are very salient.[40] In the second case, we take $\psi(a) = \mathbf{1}_{a\leq 1}$ and $b(z) = 1/z$. Then, we recover the schmeduling case of Liebman and Zeckhauser (2004) and Rees-Jones and Taubinsky (forthcoming), 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 $1-m$).[41]

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.[42] The earnings distribution is asymptotically Pareto with exponent π (i.e., when z is large, $1 - H(z) \propto z^{-\pi}$).

PROPOSITION 12 (Optimal Tax Rate for Top Incomes): *The optimal marginal rate $\bar{\tau}$ for top incomes is*

$$\bar{\tau} = \frac{1-\bar{g}}{1-\bar{g}+\bar{\eta}^r+\bar{\zeta}^{c,r}\pi\big(m+(1-m)A\big)}, \tag{24}$$

where $1-m$ and $A = \int_0^\infty a^{\pi-1}\psi(a)\,da$ index the degree of misperception of taxes (as in equation (23)). 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.

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

[40] 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).

[41] Indeed, $\int_0^\infty T'(az)\psi(a)\,da + \big(T(0)/z\big) = T(z)/z$ is the average tax rate.

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

The proof (detailed in the online Appendix) is a direct application of the optimal tax formula (22), using the fact that $\zeta^c(z) = m\zeta^{c,r}(z)$, that $\zeta^c_{z^*}(z) = (1-m)\big(\psi(z^*/z)/z\big)\zeta^{c,r}(z)$, that $\bar{\eta} = \bar{\eta}^r$, and that $\tilde{\tau}^b$ tends to 0 for high incomes.

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 $\psi(a) = \xi a^{-\xi-1}\mathbf{1}_{a\geq 1}$ 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 (forthcoming), 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. Instead, this is possible with behavioral agents. Consider an example using the misperceived 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, $\tilde{\tau}^b(z) = -g(z)\phi'(z)((\theta-1)/\theta)$, $\gamma = g$, and $\zeta^c_{z^*} = 0$. When $\theta > 1$, we have $\tilde{\tau}^b(z^*) < 0$, 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 formalization of a behavioral rationale for the EITC (Earned Income Tax Credit) program. Indeed, this type of bias is the focus of the intuitive policy arguments that have been put forth for this program: helping individuals overcome a "culture of poverty," transmitted both within and across generations. In parallel and independent work, Gerritsen (2016) and Lockwood (2018) derive a modified Saez formula in the context of a misperceived utility model. Lockwood (2018) provides an empirical analysis documenting significant present-bias among EITC recipients, showing that a calibrated version of the model goes a long way toward rationalizing the negative marginal tax rates associated with the EITC program.[43]

[43] This differs from alternative rationales for negative marginal income tax rates that have been put forth in the traditional literature. For example, Saez (2002) and Choné and Laroque (2005) show that if the Mirrlees model is extended to allow for an extensive margin of labor supply with unobserved heterogeneous disutilities of work, then negative marginal income tax rates can arise at the optimum.

IV. Conclusion

We have generalized the main results of the traditional theory of optimal taxation to allow for a large class of behavioral biases. Our analysis revisits a number of classical results and encompasses the traditional arguments of Ramsey, Pigou, and Mirrlees.

In Farhi and Gabaix (2019) we extend our analysis to the production economy results of Diamond and Mirrlees (1971) and the uniform commodity taxation result of Atkinson and Stiglitz (1976). We also present a modest attempt at modeling mental accounts with an application to optimal vouchers, which has since been adopted by Hastings and Shapiro (2018).

One upshot of this paper is that numerous quantities can in principle have a big impact on optimal policy, but have scarcely or not yet been measured. Measuring those quantities presents an exciting research opportunity.

Appendix A. Notations

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

$\mathbf{c}$: consumption vector
h: index for household type h
L: government's objective function
$\mathbf{m}, \mathbf{M}$: attention vector, matrix
$\mathbf{p}$: pre-tax price
$\mathbf{q} = \mathbf{p} + \boldsymbol{\tau}$: after-tax price
$\mathbf{q}^s$: subjectively perceived after-tax price
$\mathbf{S}_j$: column of the Slutsky matrix when price j changes
$u(\mathbf{c})$: experienced utility
$u^s(\mathbf{c})$: subjectively perceived utility
$v(\mathbf{q}, w)$: experienced indirect utility
$v^s(\mathbf{q}, w)$: subjectively perceived indirect utility
w: personal income
W: social utility
γ^h (resp. $\gamma^{\xi,h}$): marginal social utility of income (resp. adjusted for externalities)
η^h: nudgeability of agents of type h
λ: weight on revenue raised in planner's objective
ψ_i: demand elasticity for good i
$\boldsymbol{\tau}$: tax
$\boldsymbol{\tau}^b$: behavioral wedge
$\boldsymbol{\tau}^s$: subjectively perceived tax
ξ: externality
χ: intensity of the nudge

Appendix B. Behavioral Consumer Price Theory

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

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

A. *Abstract General Framework*

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

$$\frac{v_{q_j}(\mathbf{q},w)}{v_w(\mathbf{q},w)} = -e_{q_j}. \tag{25}$$

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

$$\mathbf{S}^C_j(\mathbf{q},w) = \mathbf{c}_{q_j}(\mathbf{q},w) + \mathbf{c}_w(\mathbf{q},w)\,B_{q_j}(\mathbf{c},\mathbf{q})_{|\mathbf{c}=\mathbf{c}(\mathbf{q},w)}. \tag{26}$$

The Hicksian demand is $\mathbf{h}(\mathbf{q},\hat{u}) = \mathbf{c}\big(\mathbf{q},e\big(\mathbf{q},\hat{u}\big)\big)$, and the Hicksian-demand-based Slutsky matrix is defined as $\mathbf{S}^H_j(\mathbf{q},\hat{u}) = \mathbf{h}_{q_j}(\mathbf{q},\hat{u})$.

The Slutsky matrices represent how demand changes when prices change by a small amount, and the budget is compensated to make the previous basket or the previous utility level available: $\mathbf{S}^C(\mathbf{q},w) = \partial_{\mathbf{x}}\mathbf{c}\left(\mathbf{q}+\mathbf{x},B\big(\mathbf{c}(\mathbf{q},w),\mathbf{q}+\mathbf{x}\big)\right)_{|\mathbf{x}=0}$ and $\mathbf{S}^H(\mathbf{q},w) = \partial_{\mathbf{x}}\mathbf{c}\left(\mathbf{q}+\mathbf{x},e\big(\mathbf{q}+\mathbf{x},v(\mathbf{q},w)\big)\right)_{|\mathbf{x}=0}$, i.e., using (25),

$$\mathbf{S}^H_j(\mathbf{q},w) = \mathbf{c}_{q_j}(\mathbf{q},w) - \mathbf{c}_w(\mathbf{q},w)\frac{v_{q_j}(\mathbf{q},w)}{v_w(\mathbf{q},w)}. \tag{27}$$

In the traditional model, $\mathbf{S}^C = \mathbf{S}^H$, but we shall see that this won't be the case in general.[44]

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

$$B_{\mathbf{c}}\cdot\mathbf{c}_w = 1, \quad B_{\mathbf{c}}\cdot\mathbf{c}_{q_i} = -B_{q_i}, \quad u_{\mathbf{c}}\cdot\mathbf{c}_w = v_w. \tag{28}$$

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

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

PROPOSITION 13 (Behavioral Roy's Identity): *We have*

$$\frac{v_{q_j}(\mathbf{q},w)}{v_w(\mathbf{q},w)} = -B_{q_j}\big(\mathbf{c}(\mathbf{q},w),\mathbf{q}\big) + D_j(\mathbf{q},w), \tag{29}$$

where

$$D_j(\mathbf{q},w) = -\boldsymbol{\tau}^b(\mathbf{q},w)\cdot\mathbf{c}_{q_j}(\mathbf{q},w) = -\boldsymbol{\tau}^b\cdot\mathbf{S}_j^H = -\boldsymbol{\tau}^b\cdot\mathbf{S}_j^C, \tag{30}$$

and the behavioral wedge is defined to be

$$\boldsymbol{\tau}^b(\mathbf{q},w) = B_{\mathbf{c}}\big(\mathbf{c}(\mathbf{q},w),\mathbf{q}\big) - \frac{u_{\mathbf{c}}\big(\mathbf{c}(\mathbf{q},w)\big)}{v_w(\mathbf{q},w)}. \tag{31}$$

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

PROOF:

Relations (28) imply: $\boldsymbol{\tau}^b\cdot\mathbf{c}_w = \big(B_{\mathbf{c}} - \big(u_{\mathbf{c}}/v_w\big)\big)\cdot\mathbf{c}_w = 1-1 = 0$. Next, we differentiate $v(\mathbf{q},w) = u\big(\mathbf{c}(\mathbf{q},w)\big)$:

$$\begin{aligned}\frac{v_{q_i}}{v_w} = \frac{u_{\mathbf{c}}\cdot\mathbf{c}_{q_i}}{v_w} &= \frac{\big(u_{\mathbf{c}} - v_w B_{\mathbf{c}} + v_w B_{\mathbf{c}}\big)\cdot\mathbf{c}_{q_i}}{v_w}\\ &= \frac{\big(u_{\mathbf{c}} - v_w B_{\mathbf{c}}\big)\cdot\mathbf{c}_{q_i}}{v_w} - B_{q_i} \qquad \text{as } B_{\mathbf{c}}\cdot\mathbf{c}_{q_i} = -B_{q_i}, \text{ from (28)}\\ &= -\boldsymbol{\tau}^b\cdot\mathbf{c}_{q_i} - B_{q_i}.\end{aligned} \tag{32}$$

Next,

$$\begin{aligned}D_j = -\boldsymbol{\tau}^b\cdot\mathbf{c}_{q_j} &= -\boldsymbol{\tau}^b\cdot\left(\mathbf{S}_j^H + \mathbf{c}_w(\mathbf{p},w)\frac{v_{q_j}}{v_w}\right) \qquad \text{by (27)}\\ &= -\boldsymbol{\tau}^b\cdot\mathbf{S}_j^H \qquad \text{as } \boldsymbol{\tau}^b\cdot\mathbf{c}_w = 0.\end{aligned} \tag{33}$$

Likewise, (26) gives, using again $\boldsymbol{\tau}^b\cdot\mathbf{c}_w = 0$: $D_j = -\boldsymbol{\tau}^b\cdot\mathbf{c}_{q_j} = -\boldsymbol{\tau}^b\cdot\left(\mathbf{S}_j^C - \mathbf{c}_w B_{q_j}\right) = -\boldsymbol{\tau}^b\cdot\mathbf{S}_j^C$. ∎

PROPOSITION 14 (Slutsky Relation Modified): *With* $\mathbf{c}(\mathbf{q},w)$ *we have*

$$\begin{aligned}\mathbf{c}_{q_j}(\mathbf{q},w) &= -\mathbf{c}_w B_{q_j} + \mathbf{S}_j^H + \mathbf{c}_w D_j\\ &= -\mathbf{c}_w B_{q_j} - \mathbf{c}_w\big(\boldsymbol{\tau}^b\cdot\mathbf{S}_j^H\big) + \mathbf{S}_j^H = -\mathbf{c}_w B_{q_j} + \mathbf{S}_j^C,\end{aligned} \tag{34}$$

$$\mathbf{S}_j^C - \mathbf{S}_j^H = \mathbf{c}_w D_j = -\mathbf{c}_w\big(\boldsymbol{\tau}^b \cdot \mathbf{S}_j^H\big). \tag{35}$$

PROOF:

Note,

$$\begin{aligned}\mathbf{c}_{q_j} &= \mathbf{c}_w \frac{v_{q_j}}{v_w} + \mathbf{S}_j^H, \qquad \text{by (27)}\\ &= \mathbf{c}_w\big(-B_{q_j} + D_j\big) + \mathbf{S}_j^H, \qquad \text{by Proposition 13.}\end{aligned}$$

Also, (26) gives $\mathbf{c}_{q_j} = -\mathbf{c}_w B_{q_j} + \mathbf{S}_j^C$. ∎

LEMMA 1: *We have*

$$B_{\mathbf{c}} \cdot \mathbf{S}_j^C = 0, \quad B_{\mathbf{c}} \cdot \mathbf{S}_j^H = -D_j. \tag{36}$$

PROOF:

Relations (28) imply $B_{\mathbf{c}} \cdot \mathbf{S}_j^C = B_{\mathbf{c}} \cdot \big(\mathbf{c}_{q_j} + \mathbf{c}_w B_{q_j}\big) = -B_{q_j} + B_{q_j} = 0$. Also, $B_{\mathbf{c}} \cdot \mathbf{S}_j^H = B_{\mathbf{c}} \cdot \big(\mathbf{S}_j^C - \mathbf{c}_w D_j\big) = -D_j$. ∎

B. *Application in Specific Behavioral Models*

For clarity, we consider the two models of misperceptions separately.

Misperceived Utility Model.—In the decision-utility model there is an experience utility function $u(\mathbf{c})$, and a perceived utility function $u^s(\mathbf{c})$. Demand is $\mathbf{c}(\mathbf{q}, w) = \arg\max_{\mathbf{c}} u^s(\mathbf{c})$ subject to $B(\mathbf{q}, \mathbf{c}) \leq w$.

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

PROPOSITION 15: *In the misperceived utility model,* $\mathbf{S}_j^C = \mathbf{S}_j^{s,r}$ *is the Slutsky matrix of a rational agent with utility* $u^s(\mathbf{c})$. *The behavioral wedge is*

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

Misperceived Prices Model.—To illustrate this framework, we take the misperceived prices model (Gabaix 2014). It comprises a perception function $\mathbf{q}^s(\mathbf{q}, w)$ (which itself can be endogenized, something we consider later). The demand satisfies

$$\mathbf{c}(\mathbf{q}, w) = \mathbf{h}^r\big(\mathbf{q}^s(\mathbf{q}, w), v(\mathbf{q}, w)\big),$$

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

PROPOSITION 16: *Take the misperceived prices model. Then, with* $\mathbf{S}^r(\mathbf{q}, w) = \mathbf{h}^r_{\mathbf{q}^s}\big(\mathbf{q}^s(\mathbf{q}, w), v(\mathbf{q}, w)\big)$ *the Slutsky matrix of the underlying rational agent, we have*

$$\mathbf{S}^H_j(\mathbf{q}, w) = \mathbf{S}^r(\mathbf{q}, w)\Big(\mathbf{q}^s_{q_j}(\mathbf{q}, w) - \mathbf{q}^s_w(\mathbf{q}, w)\frac{v_{q_j}}{v_w}\Big), \tag{37}$$

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

$$\boldsymbol{\tau}^b = B_{\mathbf{c}}(\mathbf{c}, \mathbf{q}) - \frac{B_{\mathbf{c}}(\mathbf{c}, \mathbf{q}^s)}{B_{\mathbf{c}}(\mathbf{c}, \mathbf{q}^s) \cdot \mathbf{c}_w(\mathbf{q}, w)}. \tag{38}$$

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

$$D_j = -\big(B_{\mathbf{c}}(\mathbf{q}, \mathbf{c}) - B_{\mathbf{c}}(\mathbf{q}^s, \mathbf{c})\big) \cdot \mathbf{S}^H_j = -B_{\mathbf{c}}(\mathbf{q}, \mathbf{c}) \cdot \mathbf{S}^H_j, \tag{39}$$

so that

$$D_j = -\bar{\boldsymbol{\tau}}^b \cdot \mathbf{S}^H_j \quad \textit{with} \quad \bar{\boldsymbol{\tau}}^b = B_{\mathbf{c}}(\mathbf{q}, \mathbf{c}) - B_{\mathbf{c}}(\mathbf{q}^s, \mathbf{c}). \tag{40}$$

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

PROOF:

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

$$\mathbf{S}^H_j = \mathbf{c}_{q_j}(\mathbf{q}, w) - \mathbf{c}_w(\mathbf{q}, w)\frac{v_{q_j}(\mathbf{q}, w)}{v_w(\mathbf{q}, w)} = \mathbf{h}^r_{\mathbf{q}^s}\mathbf{q}^s_{q_j}(\mathbf{q}, w) + \mathbf{h}^r_u v_{q_j} - \mathbf{c}_w\frac{v_{q_j}}{v_w}$$

$$= \mathbf{h}^r_{\mathbf{q}^s}\mathbf{q}^s_{q_j}(\mathbf{q}, w) + \mathbf{h}^r_u v_{q_j} - \big(\mathbf{h}^r_u v_w + \mathbf{h}^r_{\mathbf{q}^s}\mathbf{q}^s_w(\mathbf{q}, w)\big)\frac{v_{q_j}}{v_w} \quad \text{as } \mathbf{c}_w = \mathbf{h}^r_u v_w + \mathbf{h}^r_{\mathbf{q}^s}\mathbf{q}^s_w$$

$$= \mathbf{S}^r\Big(\mathbf{q}^s_{q_j}(\mathbf{q}, w) - \mathbf{q}^s_w(\mathbf{q}, w)\frac{v_{q_j}}{v_w}\Big).$$

Next, observe that the demand satisfies $u_{\mathbf{c}}\big(\mathbf{c}(\mathbf{q}, w)\big) = \Lambda B_{\mathbf{c}}(\mathbf{q}^s, \mathbf{c})$ for some Lagrange multiplier Λ, and that $B_{\mathbf{c}}(\mathbf{q}^s, \mathbf{c})\mathbf{S}^r = \mathbf{0}$ for a rational agent (see equation (36) applied to that agent). So, $B_{\mathbf{c}}(\mathbf{q}^s, \mathbf{c})\mathbf{S}^H = \mathbf{0}$. Next,

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

Given (28), $\frac{u_{\mathbf{c}}(\mathbf{c}(\mathbf{q},w))}{v_w(v,w)} = \frac{u_{\mathbf{c}}}{u_{\mathbf{c}}\cdot\mathbf{c}_w} = \frac{B_{\mathbf{c}}(\mathbf{c},\mathbf{q}^s)}{B_{\mathbf{c}}(\mathbf{c},\mathbf{q}^s)\cdot\mathbf{c}_w}$. Finally, (4) comes from (35):[45]

$$\mathbf{S}_j^C = \mathbf{S}_j^H-\mathbf{c}_w\big(\boldsymbol{\tau}^b\cdot\mathbf{S}_j^H\big) = \Big(I-\mathbf{c}_w\big(\boldsymbol{\tau}^b\big)'\Big)\mathbf{S}_j^H.\ \blacksquare$$

Appendix C. Additional Proofs

PROOF OF PROPOSITION 1:

We have

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

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

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

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

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

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

PROOF OF PROPOSITION 3:

We have

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

[45] Another useful relation is that $u_{\mathbf{c}}\mathbf{S}^H = \mathbf{0}$ in the (static) misperceived prices model (this is because $u_{\mathbf{c}} = \Lambda B_{\mathbf{c}}(\mathbf{c},\mathbf{q}^s)$ for some scalar Λ, and $B_{\mathbf{c}}(\mathbf{c},\mathbf{q}^s)\mathbf{S}^H = 0$ from equation (36)). This is not true in the misperceived utility model.

so $d\xi/d\chi = \left(\sum_h \xi_{\mathbf{c}^h} \cdot \mathbf{c}^h_\chi\right)/\left(1 - \sum_h \xi_{\mathbf{c}^h} \cdot \mathbf{c}^h_\xi\right)$. Thus, the additional term in $\partial L/\partial\chi$ arising due to externality is

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

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

$$\begin{aligned}
\frac{\partial L}{\partial \chi} &= \sum_h \left\{ W_{v^h} v^h_w \frac{u^h_{\mathbf{c}}}{v^h_w} \mathbf{c}^h_\chi + W_{v^h} v^h_w \frac{u^h_\chi}{v^h_w} + \lambda \boldsymbol{\tau} \cdot \mathbf{c}^h_\chi + \Xi \xi_{\mathbf{c}^h} \cdot \mathbf{c}^h_\chi \right\} \\
&= \sum_h \left\{ \left[W_{v^h} v^h_w \frac{u^h_{\mathbf{c}}}{v^h_w} + \lambda\left(\boldsymbol{\tau} - \boldsymbol{\tau}^{\xi,h}\right) \right] \mathbf{c}^h_\chi + W_{v^h} v^h_w \frac{u^h_\chi}{v^h_w} \right\} \\
&= \sum_h \left\{ \left[\beta^h \left(\frac{u^h_{\mathbf{c}}}{v^h_w} - \mathbf{q} + \mathbf{q} \right) + \lambda\left(\boldsymbol{\tau} - \boldsymbol{\tau}^{\xi,h}\right) \right] \mathbf{c}^h_\chi + \beta^h \frac{u^h_\chi}{v^h_w} \right\} \\
&= \sum_h \left\{ \left[-\lambda \tilde{\boldsymbol{\tau}}^{b,h} + \lambda\left(\boldsymbol{\tau} - \boldsymbol{\tau}^{\xi,h}\right) \right] \mathbf{c}^h_\chi + \beta^h \frac{u^h_\chi}{v^h_w} \right\}. \blacksquare
\end{aligned}$$

Tax Formula in the Limit of Small Taxes without Quasilinear Utility—We can obtain a formula similar to (13) for the optimal tax, without assuming quasilinear utility (for simplicity, we assume no Pigouvian externality). We assume that for small taxes agents consume $\mathbf{c}^h(\mathbf{p} + \boldsymbol{\tau}) = \mathbf{c}^{r,h}(\mathbf{p} + \boldsymbol{\tau}) + \hat{\mathbf{c}}^{u,h}(\mathbf{p}, w) + \hat{\mathbf{c}}^{M,h}(\mathbf{p}, w)\boldsymbol{\tau} + O\left(\|\boldsymbol{\tau}\|^2\right) + O\left(\|\hat{\mathbf{c}}^{u,h}(\mathbf{p}, w)\|^2\right)$. This formulation captures two forces. First, even if taxes are 0, consumers may misoptimize, as captured by the term $\hat{\mathbf{c}}^{u,h}(\mathbf{p}, w)$, which we take to be small in our limit of small taxes. Second, they may misreact to taxes, as captured by the term $\hat{\mathbf{c}}^{M,h}(\mathbf{p}, w)\boldsymbol{\tau}$. This general formulation gives an attention $\mathbf{M}^h = \mathbf{I} + \left(\mathbf{c}^{r,h}_{\mathbf{p}}\right)^{-1} \hat{\mathbf{c}}^{M,h}$. Then, (as detailed in online Appendix Section X.A), the optimal tax is, up to the second order in $\tilde{\eta}$:

$$\boldsymbol{\tau} = -\left[\sum_h \left(\mathbf{S}^{r,h} + \hat{\mathbf{c}}^{M,h}\right)' \left(\mathbf{I} - \Omega^h \hat{\mathbf{c}}^{M,h}\right) + \frac{v_{ww}}{v_w} \mathbf{c}^h \mathbf{c}^{h\prime}\right]^{-1}_{>0} \left[\sum_h \left(1 - \frac{b^h}{\lambda}\right) \mathbf{c}^h - \left(\mathbf{S}^{r,h} + \hat{\mathbf{c}}^{M,h}\right)' \Omega^h \hat{\mathbf{c}}^{u,h}\right]_{>0}, \tag{41}$$

where $\Omega^h = -u^h_{\mathbf{cc}}(\mathbf{p}, w)/v^h_w(\mathbf{p}, \mathrm{w})$, $\mathbf{S}^{r,h} = \mathbf{c}^h_{\mathbf{p}} + \mathbf{c}^h_w \mathbf{c}^{h\prime}$, $\tilde{\eta} = \sum_h |b^h - \lambda| + \left\|\hat{\mathbf{c}}^{u,h}(\mathbf{p}, w)\right\|$. All the variables are evaluated at $(\mathbf{p}, w)$ and subscript >0 indicates the selection of the $(n-1) \times (n-1)$ submatrix corresponding to all goods except good 0.

The numerator of (41) features $\left(1 - \left(b^h/\lambda\right)\right)\mathbf{c}^h$, which is the revenue-raising/redistributive motive; $\hat{\mathbf{c}}^{u,h}$, which captures the consumption

mistakes made by the agents before any taxes; and $\left(\mathbf{S}^{r,h}+\hat{\mathbf{c}}^{M,h}\right)'\Omega^h$, which captures the Slutsky matrix of the agent, corrected by their misperception to taxes $\hat{\mathbf{c}}^{M,h}$. The denominator is a matrix version of the inverse elasticity, adjusted for income effects. This is the expression that shows up in more user-friendly terms throughout Section II and in (13).

REFERENCES

Abaluck, Jason, and Jonathan Gruber. 2011. "Heterogeneity in Choice Inconsistencies among the Elderly: Evidence from Prescription Drug Plan Choice." *American Economic Review* 101 (3): 377–81.

Aguiar, Victor H., and Roberto Serrano. 2017. "Slutsky Matrix Norms: The Size, Classification, and Comparative Statics of Bounded Rationality." *Journal of Economic Theory* 172: 163–201.

Alesina, Alberto, Stefanie Stantcheva, and Edoardo Teso. 2018. "Intergenerational Mobility and Preferences for Redistribution." *American Economic Review* 108 (2): 521–54.

Allcott, Hunt, Christopher Knittel, and Dmitry Taubinsky. 2015. "Tagging and Targeting of Energy Efficiency Subsidies." *American Economic Review* 105 (5): 187–91.

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

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

Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market." *American Economic Review* 105 (8): 2501–38.

Anagol, Santosh, and Hugh Hoikwang Kim. 2012. "The Impact of Shrouded Fees: Evidence from a Natural Experiment in the Indian Mutual Funds Market." *American Economic Review* 102 (1): 576–93.

Atkinson, A. 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): 1623–67.

Becker, Gary S., and Kevin M. Murphy. 1993. "A Simple Theory of Advertising as a Good or Bad." *Quarterly Journal of Economics* 108 (4): 941–64.

Bénabou, Roland, and Jean Tirole. 2006a. "Belief in a Just World and Redistributive Politics." *Quarterly Journal of Economics* 121 (2): 699–746.

Bénabou, Roland, and Jean Tirole. 2006b. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–78.

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.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2013. "Salience and Consumer Choice." *Journal of Political Economy* 121 (5): 803–43.

Brown, Jennifer, Tanjim Hossain, and John Morgan. 2010. "Shrouded Attributes and Information Suppression: Evidence from the Field." *Quarterly Journal of Economics* 125 (2): 859–76.

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." *American Economic Review* 105 (7): 2183–2203.

Chetty, Raj. 2009. "The Simple Economics of Salience and Taxation." NBER Working Paper 15246.

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

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review* 99 (4): 1145–77.

Choné, Philippe, and Guy Laroque. 2005. "Optimal Incentives for Labor Force Participation." *Journal of Public Economics* 89 (2–3): 395–425.

Cremer, Helmuth, and Pierre Pestieau. 2011. "Myopia, Redistribution and Pensions." *European Economic Review* 55 (2): 165–75.

Dávila, Eduardo. 2017. "Optimal Financial Transaction Taxes." http://www.eduardodavila.com/research/davilaoptimalftt.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): 315–72.

Diamond, Peter A. 1975. "A Many-Person Ramsey Tax Rule." *Journal of Public Economics* 4 (4): 335–42.

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

Farhi, Emmanuel, and Xavier Gabaix. 2019. "Optimal Taxation with Behavioral Agents." NBER Working Paper 21524.

Feldman, Naomi E., Peter Katuscak, and Laura Kawano. 2016. "Taxpayer Confusion: Evidence from the Child Tax Credit." *American Economic Review* 106 (3): 807–35.

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. 2019. "Behavioral Inattention." In *Handbook of Behavioral Economics,* Vol. 2, edited by B. Douglas Bernheim, Stefano Dellavigna, and David Laibson, 261–343. Amsterdam: North-Holland.

Gabaix, Xavier, and David Laibson. 2006. "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets." *Quarterly Journal of Economics* 121 (2): 505–40.

Galle, Brian. 2014. "Tax, Command… or Nudge? Evaluating the New Regulation." *Texas Law Review* 92 (4): 837.

Gerritsen, Aart. 2016. "Optimal Taxation When People Do Not Maximize Well-Being." *Journal of Public Economics* 144: 122–39.

Glaeser, Edward L. 2006. "Paternalism and Psychology." *University of Chicago Law Review* 73 (1): 133–56.

Green, Jerry, and Eytan Sheshinski. 1976. "Direct versus Indirect Remedies for Externalities." *Journal of Political Economy* 84 (4): 797–808.

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

Gruber, Jonathan, and Botond Koszegi. 2004. "Tax Incidence When Individuals Are Time-Inconsistent: The Case of Cigarette Excise Taxes." *Journal of Public Economics* 88 (9–10): 1959–87.

Hastings, Justine, and Jesse M. Shapiro. 2018. "How Are SNAP Benefits Spent? Evidence from a Retail Panel." *American Economic Review* 108 (12): 3493–3540.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* London: Macmillan.

Kamenica, Emir. 2008. "Contextual Inference in Markets: On the Informational Content of Product Lines." *American Economic Review* 98 (5): 2127–49.

Khaw, Mel Win, Ziang Li, and Michael Woodford. 2017. "Risk Aversion as a Perceptual Bias." NBER Working Paper 23294.

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): 443–77.

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

Lockwood, Benjamin. 2018. "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 (114): 175–208.

Mullainathan, Sendhil, Joshua Schwartzstein, and William J. Congdon. 2012. "A Reduced-Form Approach to Behavioral Public Finance." *Annual Review of Economics* 4: 511–40.

O'Donoghue, Ted, and Matthew Rabin. 2006. "Optimal Sin Taxes." *Journal of Public Economics* 90 (10–11): 1825–49.

Pigou, A. 1920. *The Economics of Welfare.* London: Macmillan.
Ramsey, F. P. 1927. "A Contribution to the Theory of Taxation." *Economic Journal* 37: 47–61.
Rees-Jones, Alex, and Dmitry Taubinsky. Forthcoming. "Measuring 'Schmeduling.'" *Review of Economic Studies***.**
Saez, Emmanuel. 2001. "Using Elasticities to Derive Optimal Income Tax Rates." *Review of Economic Studies* 68 (1): 205–29.
Saez, Emmanuel. 2002. "Optimal Income Transfer Programs: Intensive versus Extensive Labor Supply Responses." *Quarterly Journal of Economics* 117 (3): 1039–73.
Salanié, Bernard. 2011. *The Economics of Taxation*. Cambridge, MA: 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): 1423–52.
Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50 (3): 665–90.
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): 130–67.
Taubinsky, Dmitry, and Alex Rees-Jones. 2018. "Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment." *Review of Economic Studies* 85 (4): 2462–96.
Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* New Haven, CT: Yale University Press.
Weitzman, Martin L. 1974. "Prices vs. Quantities." *Review of Economic Studies* 41 (4): 477–91.