CESifo WORKING PAPERS

7732
2019

July 2019

Correct Me If You Can – Optimal Non-Linear Taxation of Internalities

Andreas Gerster, Michael Kramm

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7732*
Category 1: Public Finance

Correct Me If You Can – Optimal Non-Linear Taxation of Internalities

Abstract

A growing literature has shown that behavioral biases influence consumer choices. Such so-called internalities are ubiquitous in many settings, including energy efficiency investments and the consumption of sin goods, such as cigarettes and sugar. In this paper, we use a mechanism design approach to characterize the optimal non-linear tax (or subsidy) for correcting behaviorally biased consumers. We demonstrate that market choices are informative about consumers' bias, which can be exploited for benevolent price discrimination via a non-linear tax schedule. We derive that such "internality revelation" depends on two sufficient statistics: the correlation between valuations and biases, as well as the signal-to-noise ratio of the bias. Furthermore, we find that there must be a minimum alignment of preferences among the designer and the consumer to ensure internality tax implementability. We contrast our results with the insights from standard non-linear income taxation and discuss that the optimal corrective tax schedule is typically convex. In addition, we apply our findings to the light bulb market and determine the optimal non-linear subsidy for energy efficiency.

JEL-Codes: H210, D820, D040, Q580.

Keywords: optimal commodity taxation, non-linear taxation, behavioral economics, public economics, internalities, environmental economics.

Andreas Gerster
University of Mannheim
L7 3-5
Germany / Mannheim
gerster@uni-mannheim.de

Michael Kramm
Technical University of Dortmund
Vogelpothsweg 87
Germany / Dortmund
michael.kramm@tu-dortmund.de

June 26, 2019
We are grateful for valuable comments and suggestions by Jason Abaluck, Sebastian Fuchs, Lorenz Götte, Andreas Haufler, Eckhard Janeba, Duk Gyoo Kim, Wolfgang Leininger, Jörg Peter Lingens, Lars Metzger, Adam Sanjurjo, Arthur Seibold, and Thomas Tröger. This paper also benefited from discussions with participants of the European Winter Meeting of the Econometric Society 2018 (Naples), the CESifo Area Conference on Public Sector Economics 2019, and the Maastricht Behavioral Economic Policy Symposium, as well as a seminar at University of Mannheim.

1 Introduction

A growing literature has demonstrated that consumers make decisions that are not in their best interest. For example, present biased consumers undervalue future cost from consuming "sin goods", such as cigarettes or sugar (Laibson, 1997; Loewenstein and Prelec, 1992; O'Donoghue and Rabin, 1999, 2006). Consumers are inattentive to opaque product attributes such as the energy efficiency of an appliance (Allcott and Taubinsky, 2015) or the expected out-of-pocket costs of a health insurance plan (Abaluck and Gruber, 2011). Furthermore, consumers hold biased beliefs in fields as diverse as schooling returns, the caloric content of nutrition, and energy efficiency (Attari et al., 2010; Bollinger et al., 2011; Jensen, 2010). Misoptimizing consumers inflict a so-called internality upon themselves, which provides a justification for corrective taxation beyond the classical case of externalities. Internalities are inherently heterogeneous as consumers differ in the degree of their bias and not all consumers misoptimize. As a consequence, targeting corrective taxes towards behaviorally biased consumers is crucial for inducing behavioral change and improving social welfare.

In this paper, we explore the potential of non-linear commodity taxation to target behaviorally biased consumers and thereby add to the rapidly growing literature in behavioral public economics (see, e.g., Bernheim and Taubinsky 2018). Following Farhi and Gabaix (2017) and Mullainathan et al. (2012), we employ a general model of biases that encompasses a broad class of behavioral failures driving a wedge between "experienced" and "decision utility" (Kahneman et al., 1997), such as present bias, limited attention and biased beliefs. Based on this generic specification of a behavioral bias, we derive the optimal non-linear commodity tax employing a mechanism design approach. Using data by Allcott and Taubinsky (2015), we apply our results to the light bulb market and determine the optimal non-linear subsidy for energy efficiency.

A novelty of our approach is that we analyze how a mechanism designer can employ benevolent price discrimination, i.e., optimally differentiate taxes to correct choices of behaviorally biased consumers. In particular, we investigate settings where consumer decisions reflect potentially biased perceived valuations at the time of purchase, such as the immediate gratification for a cigarette or a sugary product. We examine the information structure that is implied by internalities and analytically characterize the conditions under which market choices reveal consumers' bias – a phenomenon that we denote as "internality revelation". We also investigate how internalities induce disagreement on valuations between consumers and the designer, and we characterize the conditions that guarantee the implementability of the optimal corrective tax. Furthermore, we describe the properties of optimal non-linear com-

modity taxes and contrast them with findings on optimal non-linear income taxation, commonly referred to as the "ABC formula" (Diamond, 1998).

Our paper combines several strands of the literature. From a methodological angle, we built on Mirrlees's (1971) pathbreaking work on optimal non-linear income taxation and combine it with the presence of behavioral biases. In that regard, our paper is close to recent work on behavioral income taxation by Lockwood and Taubinsky (2016) and Gerritsen (2016), for example. By contrast, the focus of our paper is not the trade-off between efficiency and equity, as inherent in optimal income taxation, but on the potential of non-linear taxes to correct consumer choices.

Optimal corrective taxation of behaviorally biased consumers has recently been studied by a growing literature (O'Donoghue and Rabin, 2006; Allcott et al., 2014, 2015; Farhi and Gabaix, 2017; Mullainathan et al., 2012). Our paper closely relates to O'Donoghue and Rabin (2006) who show that linear corrective taxes on sin goods are welfare-improving when some consumers are present biased, whiles others are unbiased. More recently, Allcott and Taubinsky (2015) derive the optimal corrective tax for a binary choice and show that it corresponds to the average bias of consumers that are indifferent between both options. Farhi and Gabaix (2017) generalize that result and derive the optimal linear corrective tax when choices are continuous. Furthermore, Allcott et al. (2014) investigate the optimal combination of a linear tax and product subsidies when internalities and externalities are present, while Allcott et al. (2018) combine optimal linear taxation of sin goods with an optimal income taxation framework that considers redistributive motives.

We complement this strand of the literature by investigating the potential to target behaviorally biased consumers through non-linear taxation. While linear corrective taxes impose distortions on fully rational consumers (see, e.g., O'Donoghue and Rabin, 2006), non-linear taxes allow to minimize such distortions by targeting biased consumers. Traditionally, public tax schemes have targeted consumers through "tagging" (Akerlof, 1978), i.e., a conditioning of taxes on observable characteristics. While appealing in theory, it is difficult to avoid strategic behavior and to find immutable characteristics that are socially acceptable. More recently, Griffith et al. (2019) investigate optimal taxation in the alcohol market and find that a linear ethanol tax that varies among alcohol products can successfully target heavy drinkers and thus reduce the large externalities caused by them. While our paper is in principle also applicable to externalities, we show that internalities generally imply a distinct correlation pattern between consumers' choices and biases that offers particularly large potential for targeting.

Our paper is also related to a large literature in industrial organization that has explored how firms can use price discrimination to maximize profits (see, e.g., Mussa

and Rosen 1978, or Wilson 1997 for a summary). In industrial organization, the rationale of non-linear pricing is to separate consumers with different valuations, which allows firms to extract consumer surplus. We built on the same basic intuition, but explore the potential of benevolent price discrimination to correct choices of behaviorally biased consumers. Furthermore, our approach bears similarity with multidimensional mechanism design problems (McAfee and McMillan, 1988; Laffont et al., 1987) as consumers are characterized by perceived valuations and biases. An important distinction of our setting is that only one dimension, namely perceived valuations, influence consumers' choice behavior. We contribute to that strand of the literature by characterizing the conditions for the implementability of internality taxes.

We solve for the optimal non-linear tax by allowing for arbitrary consumer heterogeneity in preferences and biases.[1] In addition, we follow the reduced-form approach to behavioral public finance (Mullainathan et al., 2012), which allows us to determine the optimal tax for a variety of underlying behavioral mechanisms. Our results show that the optimal tax rate has two components. The first term corresponds to the average bias of consumers that make a particular consumption choice, while the second term captures redistribution aspects that only become relevant with non-utilitarian social preferences. Investigating the information structure implied by internalities, we show that consumers reveal their perceived type even when the aggregate information available to a policy maker is uninformative, e.g., when the bias has zero mean and is uncorrelated with valuations. We explore such internality revelation in more detail and show that it crucially depends on two sufficient statistics: the correlation between biases and valuations and the ratio of their standard deviations, which we denote as the signal-to-noise ratio of the bias. We derive that the slope of the expected conditional bias is a function of the informativeness of the choice setting, characterized by the above mentioned two sufficient statistics, and the informativeness of a consumers' report, given by the distance of her report to the average report. We also demonstrate that – as a consequence of internality revelation – the optimal non-linear corrective tax is typically convex and analytically characterize the exceptions when this finding does not hold true.

Furthermore, we investigate the conditions that ensure implementability of internality taxes and find that they are restrictive in a double sense. First, we show that the designer must rely on a one-dimensional mechanism based on perceived valuations and infer biases indirectly. Second, we find that corrective taxes can be implemented only as long as there is a minimum alignment between consumers' perceived preferences and the normative stance of the mechanism designer. In particular, we find that

[1] Our approach to model internalities is more general than O'Donoghue and Rabin (2006), for example, who focus on sin goods, where some consumers overvalue consumption, while others do not.

there exists a trade-off between the magnitude of internality revelation and implementability of the mechanism and characterize this trade-off in terms of the sufficient statistics mentioned above. When the designer learns that a consumer heavily overvalues consumption and imposes excessive corrective taxes, implementability of the mechanism breaks down and the designer cannot fully use the information that is revealed by consumers' reports. For the light bulb market, we illustrate that the optimal non-linear tax for energy efficiency increases welfare substantially beyond the optimal linear tax and show that the optimal non-linear tax can be approximated based on the few sufficient statistics that we have identified.

From a theoretical perspective, our approach is applicable to non-linear commodity taxation of quantities, as well as qualities, i.e., product attributes. In practice, non-linear taxation can be readily applied to product attributes such as energy efficiency, which is important as internalities often distort choices between product varieties. When applied to quantities, a prerequisite for non-linear taxation is that total consumption can be tracked, which may not be feasible for all everyday commodities. Yet, a number or important applications lend themselves to non-linear taxation of quantities. In recent years, many states have legalized the sale of cannabinoids in officially recognized pharmacies. In Uruguay, for example, sellers verify the identity of buyers via fingerprints and track total quantities to enforce a maximum consumption limit of 40g of cannabis per month. Similarly, many governments operate central registries on the private ownership of cars, guns, and houses (see, e.g., Cremer and Gahvari 1998). As digitization progresses, tracking costs will decrease (Goldfarb and Tucker, 2019) and comparable settings that allow for non-linear corrective taxation are likely to become increasingly prevalent.

As a policy alternative to corrective taxes, the recent literature has highlighted non-price interventions, so-called "nudges" (Thaler and Sunstein, 2008), to influence behaviorally biased consumers. Clearly, low-cost educational interventions that successfully remove consumers' bias would be first best from a welfare perspective. Yet, in practice, such "pure nudges" (e.g., Allcott and Taubinsky 2015) are rare as they, first, often face high or prohibitive cost. Second, they are only effective when biases are triggered by "mental frictions", i.e., high cost of acquiring and processing information, rather than by "mental gaps", i.e., psychological distortions such as self-control problems (Handel and Schwartzstein, 2018). By contrast, optimal taxation is applicable to all policies that aim at correcting allocations regardless of the specific behavioral bias at work. In particular, an optimal taxation framework can also be employed to investigate nudges that induce feelings of guilt or self-esteem and thus work as a "moral tax" or "moral subsidy".

The paper is structured as follows. In Section 2, we introduce our model. Section 3 contains the analytical characterization of the optimal tax scheme and the conditions that ensure internality tax implementability, while Section 4 illustrates optimal non-linear taxation based on a simple example. In Section 5, we explore internality revelation and analyze the information structure implied by internalities. Section 6 determines the optimal non-linear subsidy for energy efficiency in the light bulb market. Section 7 discusses our findings and concludes.

2 Model Setup

In this section, we present a model that allows us to analyze how the social planner (*she*), i.e., the mechanism designer, implements a welfare maximizing tax scheme in an economy with a behaviorally biased consumer (*he*). We model the interaction between the mechanism designer and the consumer as a dynamic Bayesian game with two stages. In period one, the designer commits to a (possibly non-linear) tax regime $t : X \to \mathbb{R}$, where $X \subseteq [0, \infty)$ is the consumer's choice set. In period two, the consumer chooses his consumption $x \in X$. The game is solved by backward induction. We begin by presenting the characteristics of the consumer side, and then describe the problem of the mechanism designer.

2.1 Consumer Side

For notational simplicity, we model one representative consumer, whose choice variable is the consumption of a good x, denoted by $x \in X \subseteq [0, \infty)$. The consumer's experienced per-unit valuation of the benefits of consuming good x is captured by the random variable v, which is distributed according to the cumulative distribution function F with support $supp(F) := [\underline{v}, \overline{v}] \subseteq (-\infty, \infty)$, density function f, expected value μ_v with $-\infty < \mu_v < \infty$, and variance σ_v. The bias b, reflecting a misperception of the true valuation of the consumption choice x, is distributed according to the cumulative distribution function G with support $supp(G) := \left[\underline{b}, \overline{b}\right] \subseteq (-\infty, \infty)$, density function g, expected value μ_b with $-\infty < \mu_b < \infty$, and variance σ_b. We allow for arbitrary correlation patterns between valuations and biases, given by the correlation coefficient ρ.

The consumer's perceived per-unit valuation of the benefits of consuming good x is given by $\hat{v} : [\underline{v}, \overline{v}] \times \left[\underline{b}, \overline{b}\right] \to \mathbb{R}$, $(v, b) \mapsto \hat{v}(v, b)$, which depends on the true valuation v and the bias b. We assume that the perceived valuation increases in the true valuation and in the bias. Furthermore, we specify the bias such that $\hat{v}(v, 0) = v$, i.e., a consumer with bias $b = 0$ is unbiased, while $b < 0$ ($b > 0$) imply underesti-

mation (overestimation) of the value of consumption. The perceived valuation $\hat{v}$ is distributed according to the cumulative distribution function P, which is induced by the distributions of v and b. The density is given by p, which is determined by the joint distribution $h(v, b)$ of v and b according to $p(\hat{v}) = \int_{-\infty}^{\infty} h(v, \hat{v} - v)\mathrm{d}v$. The support $supp(P) := \left[\underline{\hat{v}}, \overline{\hat{v}}\right]$ is determined by the support of the distributions F and G.

Let $z \in \mathbb{R}$ denote the money (numeraire good) a consumer spends for the consumption of other goods. The consumer's objective function is given by his decision utility $u^d : X \times \mathbb{R} \times \left[\underline{\hat{v}}, \overline{\hat{v}}\right] \to \mathbb{R}$ with $(x, z, \hat{v}) \mapsto u^d(x, z, \hat{v})$. The consumer's experienced utility is given by $u^e : X \times \mathbb{R} \times [\underline{v}, \overline{v}] \to \mathbb{R}$ with $(x, z, v) \mapsto u^e(x, z, v)$. Experienced utility increases in the true valuation and decision utility increases in the perceived valuation. Experienced and decision utility increase in z and x.

To separate corrective from redistributive taxation motives, we assume quasilinear utility that abstracts from income effects. Utility from consuming x is given by the increasing and weakly concave function $w : X \to \mathbb{R}$. Thus, we can write decision utility as $u^d(x, z, \hat{v}) = z + \hat{v} \cdot w(x)$ and experienced utility as $u^e(x, z, v) = z + v \cdot w(x)$. The increasing and weakly convex cost function of consuming x is given by $c : X \to \mathbb{R}$. We assume that good x is produced on competitive markets so that the cost function corresponds to the price of consuming x net of taxes. The exogenous, real-valued scalar $m > 0$ denotes the initial endowment with the numeraire. Therefore, the budget constraint is given by $z \leq m - c(x) - t(x)$.

We make the standard assumption that the utility function is quasiconcave in x, which in our case implies quasiconcavity of $w(x) - c(x)$. In the following, without loss of generality, we assume that the utility function is linear and that the cost function is strictly convex.[2] This implies that the decision utility can be written as

$$u^d(x, t, \hat{v}) = m + \hat{v}x - t(x) - c(x),$$

and the experienced utility as

$$u^e(x, t, v) = m + vx - t(x) - c(x).$$

Following Farhi and Gabaix (2017) and Mullainathan et al. (2012), we assume an additive bias $\hat{v} := v + b$, so that $u^e(x, t, v) = u^d(x, t, \hat{v}) - bx$.

The behavior of a biased consumer is captured by

$$x^d(\hat{v}, t) := \arg\max_x u^d(x, t, \hat{v}),$$

[2]This modeling choice is not restrictive as we can always redefine x in terms of an alternative quality measure $\tilde{x} = w(x)$ to yield a linear utility function $\tilde{w}(\tilde{x}) = w(w^{-1}(\tilde{x})) = \tilde{x}$ and a strictly convex cost function $\tilde{c}(\tilde{x}) = c(w^{-1}(\tilde{x}))$.

and that of an unbiased consumer by

$$x^e(v,t) := \arg\max_x u^e(x,t,v).$$

To simplify notation, we write x^d and x^e as shorthand notations for $x^d(\hat{v},t)$ and $x^e(v,t)$, respectively.

2.2 Mechanism Designer

The designer's objective is to elaborate a tax scheme $t : X \to \mathbb{R}$, based on information about the distributions P, F, and G. She observes the consumer's choice x, but not any realization of the random variables v, b, and $\hat{v}$. The designer's objective function consists of the increasing and weakly concave social welfare function $W : \mathbb{R} \to \mathbb{R}$ with $u^e \mapsto W(u^e)$. For simplicity, we assume that the welfare function depends on non-negative welfare weights $\alpha(\hat{v})$, so that $u^e \mapsto \alpha(\hat{v})u^e$ with $\alpha : \left[\underline{\hat{v}}, \overline{\hat{v}}\right] \to \mathbb{R}_+$. To isolate the corrective nature of taxation in our setting, we also assume that the designer returns tax revenue to the consumer via lump-sum taxes.

Let $\mathbb{T} := \{f | f : X \to \mathbb{R}\}$ denote the function space containing all functions with domain X and codomain $\mathbb{R}$. The designer's objective is then given by

$$\max_{t\in\mathbb{T}} \int_{\hat{v}} \alpha(\hat{v}) \cdot E\left[u^e\left(x^d, t, v\right)|\hat{v}\right] p(\hat{v}) d\hat{v} + \int_{\hat{v}} t(x^d) p(\hat{v}) d\hat{v}.$$

In the above objective, we can rewrite $E\left[u^e\left(x^d,t,v\right)|\hat{v}\right] = u^d\left(x^d,t,\hat{v}\right) - E\left[b|\hat{v}\right]x^d$. A crucial term in this expression is the conditional expectation $E\left[b|\hat{v}\right]$, which captures the information a designer can learn about individual biases from observing a report $\hat{v}$. Such internality revelation plays an important role in the characterization of the optimal non-linear tax scheme and we define it as follows.

Definition 1 (Internality Revelation)**.** *Internality revelation occurs when the designer can extract information about the magnitude of a consumer's internality after observing her perceived valuation $\hat{v}$. Formally, internality revelation occurs if and only if*

$$\exists \hat{v}_1, \hat{v}_2 \in \left[\underline{\hat{v}}, \overline{\hat{v}}\right],\ \hat{v}_1 \neq \hat{v}_2 : E\left[b|\hat{v}_1\right] \neq E\left[b|\hat{v}_2\right]. \tag{1}$$

Under internality revelation, the designer can apply Bayesian learning to extract information about a consumer's bias from observing her perceived valuation. Before we solve for the optimal non-linear tax and discuss the implications of internality revelation, we briefly discuss our model setup in the next subsection.

2.3 Discussion of the Model

While we allow the bias to have any (finite) magnitude in expectation, empirical applications have typically found that consumers undervalue certain attributes, for instance due to limited attention (Allcott et al., 2015). To fix ideas, assume that a consumer chooses one of several varieties of a horizontally differentiated product, such as an electricity-using durable of varying energy efficiency levels. In this example, x can be interpreted as the energy efficiency of the durable, measured relative to the worst variety on the market, and v corresponds to the actual reduction in operating cost through better energy efficiency, given actual usage and prices. The random variable b can be interpreted as a misperception of the value of energy efficiency at the time of purchase, i.e., a wedge between perceived valuation $\hat{v}$ and the experienced valuation v of energy efficiency.

In many applications of interest, misoptimizing consumers also exacerbate problems from externalities. For example, consumers who buy energy-inefficient durables can be worse off individually, but also contribute to climate change as their energy consumption causes greenhouse gas emissions. For conceptual clarity, we focus on internalities in the main text, but solve a model with internalities and externalities in Appendix B. In particular, we find that the presence of global externalities merely shifts the marginal tax rate by a constant term that corresponds to the global per-unit damage of consumption.

Our model implicitly assumes that consumer biases exist even though aggregate information on their presence is available. When aggregate information is available to both the designer and the consumers, consumers might update beliefs about their type in exactly the same way as the designer. When biases can purely be attributed to lack of information, such updating might remove the bias. Yet, the literature on behavioral economics has demonstrated that consumers often fail to correctly update their beliefs. For example, the literature has found evidence of biased beliefs regarding energy efficiency, the calory content of nutrition, car driving abilities, and schooling returns (Attari et al., 2010; Bollinger et al., 2011; Kruger and Dunning, 1999; Jensen, 2010). In addition, when behavioral biases are due to "mental gaps" (Handel and Schwartzstein 2018), such as self-control problems, Bayesian updating does not necessarily help consumers to overcome their bias. For example, a wide literature has documented that present biases induce consumers to undervalue products that pay out only in the future, such as healthy foods and energy efficient durables (Laibson, 1997; Loewenstein and Prelec, 1992; O'Donoghue and Rabin, 1999, 2006).

For the ease of exposition and without loss of generality, we define welfare weights $\alpha(\cdot)$ as a function of $\hat{v}$. As our setup is characterized by a quasi-linear and money-

metric utility function, the welfare weights $\alpha(\hat{v})$ represent the increase in money-metric experienced utility when increasing numeraire consumption by one unit for a consumer with perceived valuation $\hat{v}$. This specification of welfare weights allows us to consider distributional implications of corrective taxation, as for example highlighted by Allcott et al. (2018). Our welfare weights can be generalized to consider a wide range of societal preferences for redistribution that may also take into account notions of fairness (Saez and Stantcheva, 2016).[3]

A central distinction to the standard mechanism design problems is that a designer evaluates *experienced utility* at x^d, the solution to the maximization problem involving the (possibly biased) *decision utility*. In Section 3, we analyze how the designer can possibly correct the internality to increase social welfare using non-linear taxation.

3 Optimal Tax Scheme

In this section, we apply the concept of a perfect Bayesian Nash equilbirium to derive the optimal tax scheme. We proceed in two steps. First, we solve for the choice of the behaviorally biased consumer that maximizes decision utility and determine the conditions under which a truthful mechanism can be implemented. Second, we derive the optimal tax schedule and discuss its properties.

3.1 Internality Tax Implementability

Employing the Revelation Principle for dominant-strategy implementation (Gibbard, 1971), we solve for a direct mechanism, where consumers truthfully reveal information about their perceived valuations. We start by discussing the dimensionality of mechanisms for internality taxation.

In our model, a consumer decides based on his perceived valuation $\hat{v}(v, b)$ rather than v, which can occur when he is either naive or sophisticated. While naive consumers are unaware of their bias, sophisticated consumers know the bias, but disregard it when when making decisions, for example owing to self-control problems (e.g., O'Donoghue and Rabin 1999).[4]

Corollary 1. *To determine optimal internality taxes, we can without loss of generality restrict the analysis to one-dimensional mechanisms in perceived valuations $\hat{v}$.*

[3]Furthermore, as long as such generalizied welfare weights are non-negative, the resulting optimal tax system is second-best Pareto efficient, so that no feasible tax reform can improve the welfare of everybody (Saez and Stantcheva, 2016).

[4]More formally, a naive biased consumer i only holds information about his perceived valuation $\hat{v}_i$. A sophisticated biased consumer i only holds information on his perceived valuation $\hat{v}_i$ and his bias b_i, but acts according to his (biased) decision utility u_i^d.

Proof. See Appendix A.A □

Even though a consumer can be characterized by his perceived valuation and his bias, Corollary 1 states that we can restrict our analysis to one-dimensional mechanisms, where reports correspond to the consumers' perceived valuations $\hat{v}$. The intuition for Corollary 1 is as follows. Naive consumers cannot truthfully report their bias, so that a mechanism designer can only employ a one-dimensional mechanism in $\hat{v}$ to correct them. By contrast, sophisticated consumers know their bias and can in principle report it, as in multi-dimensional screening problems (e.g., Basov 2005). Yet, as biases do not influence decision utility, truthtelling in a two-dimensional mechanism is not incentive-compatible. For example, a smoker who is about to purchase cigarettes would always pretend to be unbiased to avoid paying corrective taxes. Accordingly, no matter whether consumers are naive or sophisticated, we can without loss of generality focus on mechanisms that are one-dimensional in perceived valuations.

Corollary 1 allows for a straightforward application of the Revelation Principle to our setting, where the space of reports for a consumer is given by the space of his perceived valuation $\hat{v}$. For simplicity, we refer to $\hat{v}$ as a consumer's type in the following. The designer confines herself to designing a direct mechanism $(\xi, \tau) : \left[\underline{\hat{v}}, \overline{\hat{v}}\right] \to X \times \mathbb{R}$ under truthtelling to implement the welfare maximizing outcome. Based on the consumer's strategical report $\tilde{v}$, the direct mechanism assigns the consumed quantity, $\xi(\tilde{v}) \in X$, and the amount of taxes to be paid, $\tau(\tilde{v}) \in \mathbb{R}$. Without loss of generality, we assume that participation contraints are fulfilled for any consumer.[5]

Under the direct mechanism, the decision utility for report $\tilde{v}$ given the perceived valuation $\hat{v}$ is

$$u^d(\xi(\tilde{v}), \tau(\tilde{v})|\hat{v}) = m + \hat{v} \cdot \xi(\tilde{v}) - \tau(\tilde{v}) - c(\xi(\tilde{v})).$$

Since the consumer may strategically misreport his perceived valuation, truthtelling can be induced by the designer by implementing an incentive compatible mechanism. This implies that the tax scheme must satisfy

$$u^d(\xi(\hat{v}), \tau(\hat{v})|\hat{v}) \geq u^d(\xi(\tilde{v}), \tau(\tilde{v})|\hat{v}) \; \forall \hat{v}, \tilde{v} \in \left[\underline{\hat{v}}, \overline{\hat{v}}\right]. \tag{IC}$$

[5] If consumers with low perceived valuations preferred the outside option to participation, the designer could uniformly increase the tax while preserving tax rates to ensure that participation constraints are not binding. In particular, we can specify the model such that $\hat{u}^d\left(\underline{\hat{v}}\right) = \underline{u} > m$ and $\hat{u}^d\left(\overline{\hat{v}}\right) \geq \underline{u}$, where m is the utility of the outside option, i.e., not consuming.

Optimal strategic reporting of a consumer implies that the solution v^* to the problem $\max_{\tilde{v}} u^d(\xi(\tilde{v}), \tau(\tilde{v})|\hat{v})$ has to satisfy

$$\hat{v}\xi'(v^*) - \tau'(v^*) - \xi'(v^*)c'(\xi(v^*)) \stackrel{!}{=} 0. \tag{2}$$

As incentive compatibility implies that $v^* = \hat{v}$, equilibrium decision utility in an incentive-compatible direct mechanism is given by $\hat{u}^d(\hat{v}) := u^d(\xi(\hat{v}), \tau(\hat{v})|\hat{v})$, while equilibrium experienced utility is given by $\hat{u}^e(\hat{v}, b) := u^e(\xi(\hat{v}), \tau(\hat{v})|v) = \hat{u}^d(\hat{v}) - b\xi(\hat{v})$. Put differently, incentive compatibility implies that for all $\hat{v} \in \left[\underline{\hat{v}}, \overline{\hat{v}}\right]$ it has to hold that

$$\frac{\partial \hat{u}^d(\hat{v})}{\partial \hat{v}} = \xi(\hat{v}) + \hat{v}\xi'(\hat{v}) - \tau'(\hat{v}) - \xi'(\hat{v})c'(\xi(\hat{v})) \stackrel{(2)}{=} \xi(\hat{v}). \tag{3}$$

Next, we investigate the conditions that guarantee the implementability of internality taxes. In standard mechanism design theory, implementability of an incentive compatible mechanism hinges on two necessary conditions. First, the consumer's utility function must satisfy the single-crossing condition, which is met in our case since $\partial\left(\frac{\partial u^d/\partial x}{\partial u^d/\partial t}\right)/\partial\hat{v} < 0$. Second, ξ must be increasing in $\hat{v}$. A standard sufficient condition for this requirement is that P has an increasing hazard rate, that is, $\partial[p(\hat{v})/(1 - P(\hat{v}))]/\partial\hat{v} > 0$. In Proposition 1, we show that an increasing hazard rate is not sufficient any more in our setting.[6]

Proposition 1 (Internality Tax Implementability)**.** *The implementability of an incentive compatible mechanism implies that ξ is increasing in $\hat{v}$. A necessary and sufficient condition for this requirement is satisfied if*

$$\underbrace{1 - \alpha(\hat{v})\frac{\partial E[b|\hat{v}]}{\partial \hat{v}}}_{>0 \text{ if "no excessive bias correction"}} + \underbrace{\{\alpha(\hat{v}) - 1\}\frac{\partial[(1-P(\hat{v}))/p(\hat{v})]}{\partial \hat{v}} + \frac{\partial \alpha(\hat{v})}{\partial \hat{v}}\left(\frac{1-P(\hat{v})}{p(\hat{v})} - E[b|\hat{v}]\right)}_{\text{function of the hazard rate } (=0 \text{ for normalized utilitarian welfare weights})} \geq 0. \tag{4}$$

With an utilitarian welfare function and welfare weights normalized to one, the expression simplifies to a "no excessive bias correction" condition, i.e., $\partial E[b|\hat{v}]/\partial\hat{v} \leq 1$.

Proof. See Appendix A.B □

To give an intuition for Proposition 1, we consider its first term ("no excessive bias correction"), which can be rewritten as $\partial E[v|\hat{v}]/\partial\hat{v} > 0$, using that $\hat{v} = v + b$. It requires that there exists a minimum alignment between the normative stance of a designer and consumers perceived valuations. In particular, this condition is violated

[6]Note that if the implementability conditions stated in Proposition 1 are not satisfied, one can – as in the standard mechanism design setting – resort to the so-called Myerson-ironing, as described in Myerson (1981).

if a designer associates higher perceived valuations with lower experienced valuations ($\partial E[v|\hat{v}]/\partial\hat{v} < 0$), and thus fundamentally disagrees with the preferences as perceived by consumers. This condition is novel to our setting and directly follows from the distinction between perceived and experienced valuations. In addition, with non-utilitarian welfare weights $\alpha(\hat{v})$, internality tax implementability involves two further terms that are a function of the hazard rate, the expected bias and the welfare weights.

3.2 Optimal Non-Linear Internality Tax

To determine the optimal tax schedule, the designer solves a dynamic optimization problem which can be analyzed using an optimal control approach. Note that determining the equilibrium values of $\xi(\hat{v})$ and $u^d(\hat{v})$ for all $\hat{v}$ pins down the equilibrium value of $\tau(\hat{v})$ for all $\hat{v}$. Hence, the mechanism design problem of the designer is given by

$$\max_{\xi\in\mathbb{X}} \int_{\hat{v}} \alpha(\hat{v}) \cdot E\left[\hat{u}^e\left(\hat{v},b\right)|\hat{v}\right] dP(\hat{v}) + \int_{\hat{v}} \tau\left(\hat{v}\right) dP(\hat{v}), \tag{5}$$

subject to the condition from Equation (3), where $\mathbb{X} := \{f|f : \left[\underline{\hat{v}},\overline{\hat{v}}\right] \to X\}$ is the function space containing all functions with domain $\left[\underline{\hat{v}},\overline{\hat{v}}\right]$ and codomain X. The boundary conditions of the problem are given by $\hat{u}^d\left(\underline{\hat{v}}\right) = \underline{u}$ and $\hat{u}^d\left(\overline{\hat{v}}\right) \geq \underline{u}$. The control variable is ξ and the law of motion of the state variable $\hat{u}^d$ is determined by incentive compatibility and optimal strategic reporting, as defined in Equation (3). Using the definition of decision utility to replace the tax and rewriting equilibrium experienced utility in terms of equilibrium decision utility, the Hamiltonian for the above problem for all $\hat{v} \in \left[\underline{\hat{v}},\overline{\hat{v}}\right]$ is given by

$$H\left(\hat{v},\xi,\hat{u}^d\right) = \left[\alpha(\hat{v})\cdot\underbrace{\left(\hat{u}^d(\hat{v}) - E[b|\hat{v}]\xi(\hat{v})\right)}_{=E[\hat{u}^e(\hat{v},b)|\hat{v}]} + \underbrace{\left(m+\hat{v}\xi(\hat{v}) - \hat{u}^d(\hat{v}) - c(\xi(\hat{v}))\right)}_{=\tau(\hat{v})}\right] p(\hat{v}) + \mu(\hat{v})\xi(\hat{v}).$$

Following the standard solution procedure for such mechanism design problems, we employ Pontryagin's Maximum Principle, which yields the following necessary conditions for the optimal tax.[7]

[7] In addition, sufficiency is given if the control region is convex and the Hamiltonian is concave in $(\xi,\hat{u}^d)$ for every $\hat{v}$. Both conditions are satisfied in our setup, as discussed in Appendix A.C.

$$\text{FOC on control: } \frac{\partial H}{\partial \xi} = \left[-E[b|\hat{v}] \cdot \alpha(\hat{v}) + \hat{v} - c'(\cdot)\right] p(\hat{v}) + \mu(\hat{v}) \stackrel{!}{=} 0, \qquad (\text{FOC}_\text{x})$$

$$\text{FOC on state: } \frac{\partial H}{\partial \hat{u}^d} = [\alpha(\hat{v}) - 1]\, p(\hat{v}) \stackrel{!}{=} -\mu'(\hat{v}), \qquad (\text{FOC}_\text{u})$$

$$\text{transversality cond.: } \mu(\underline{\hat{v}}) \cdot \hat{u}^d(\underline{\hat{v}}) = \mu\left(\overline{\hat{v}}\right) \cdot \hat{u}^d\left(\overline{\hat{v}}\right) = 0. \qquad (\text{TVC})$$

We characterize the optimal non-linear tax scheme in Proposition 2.

Proposition 2 (Optimal Non-linear Commodity Tax)**.** *If internality tax implementability is satisfied (see Proposition 1), the optimal non-linear commodity tax in our model is implicity given by*

$$t'(x) = \frac{\int_{\hat{v}_x}^{\overline{\hat{v}}} [1 - \alpha(m)]\, p(m) dm}{p(\hat{v}_x)} + E[b|\hat{v}_x] \cdot \alpha(\hat{v}_x) \quad \forall x \in X, \tag{6}$$

where $\hat{v}_x$ is the report of a consumer that yields the allocation x under the optimal tax scheme.

Proof. See Appendix A.C □

To convey the intuition of the optimal marginal tax, we first examine the second summand of Equation (6). This term embodies the behavioral aspect of our model and does not appear in the standard literature on non-linear taxation. It corresponds to the expected bias of a consumer conditional on his reported perceived valuation, which represents the information a designer can infer about internalities from observing consumer choices. The designer uses her potential to correct consumers' choices to the extent that she can infer their internality – "she corrects them, if she can".

If the designer has a utilitarian social welfare function, the marginal tax rate from Equation (6) simplifies to the expected bias conditional on the report, $t'(x) = E[b|\hat{v}_x]$. This result contrasts with the findings by Allcott and Taubinsky (2015) who show that in a binary investment setting the optimal tax is equal to the average bias of the consumers who are indifferent between both goods at market prices. Importantly, the optimal non-linear tax differs from that result by its dependence on a report $\hat{v}$ rather than a fixed market price. This finding reflects that non-linear taxation improves upon a constant per-unit tax by exploiting additional information conveyed by consumers' reports that allows to target behaviorally biased consumers.[8]

Next, we contrast the optimal non-linear tax for behaviorally biased consumers with the famous ABC formula from the theory of optimal non-linear income taxation, derived by Diamond (1998). The ABC formula contains three factors: efficiency considerations (A), redistribution issues (B), and the dependence of the incentive compatibility constraint on the density functions via the hazard rate (C). In our model,

[8]For comparison, assuming that $c'''(\cdot) = 0$, the optimal linear tax is given by $t^* = E[b]$, that is, the unconditional expected bias (see Appendix A.D for a derivation).

efficiency considerations show up in the optimal tax formula, albeit in a novel form. Typically, the aim of optimal taxation is to reduce the distortive effects of taxation. In our model, efficiency considerations (A) reflect the motive of a designer to correct for internality biases, as captured by the second term of the optimal tax formula, which does not appear in the standard ABC formula.

Redistribution issues (B) and the density of $\hat{v}$ (C) are contained in the first summand of Equation (6), although in modified form. The intuition of this summand is as follows. When the designer changes the marginal tax at, say, x, she extracts money from all consumers with $\hat{v} \geq \hat{v}_x$. The change of her objective is captured by the term $\int_{\underline{\hat{v}}}^{\overline{\hat{v}}_x} [1 - \alpha(m)]\, p(m)dm$, since the marginal value of an additional unit of tax money is one and welfare decreases by $\alpha(\hat{v}_x)$ for a consumer with type $\hat{v}_x$. Intuitively, if the average welfare weights for these consumers exceed unity (and thus the marginal value of the tax income to the designer), the designer's objective function decreases and she should reduce the tax. The term is weighted more strongly in the optimal tax formula if the density of the type $\hat{v}_x$ is low, i.e., if it is unlikely that the consumer is marginal to the tax change at x and thus has an incentive to change his behavior.

4 Illustrating Example

We now use a simple example to illustrate the rationale of internality revelation and the optimal non-linear internality tax. Let consumers have either low or high valuations for x, $v_l = 1$ and $v_h = 2$, and either undervalue x or not, which we denote by subscripts b and n, respectively, where $b_b = 1$ and $b_n = 0$. Furthermore, we assume that biases and valuations are independently distributed and that all realizations of v and b occur with equal probability.

In that setting, consumers have three distinct perceived valuations $\hat{v} \in \{\hat{v}_1 = 1, \hat{v}_2 = 2, \hat{v}_3 = 3\}$, which occur with probability $P(\hat{v} = \hat{v}_1) = 0.25$, $P(\hat{v} = \hat{v}_2) = 0.5$, $P(\hat{v} = \hat{v}_3) = 0.25$, respectively. It is then straightforward to show that:

$$E(b|\hat{v}) = \begin{cases} 0 & \text{if } \hat{v} = 1 \\ 0.5 & \text{if } \hat{v} = 2 \\ 1 & \text{if } \hat{v} = 3. \end{cases}$$

This example shows how a consumer's perceived valuation partly reveals his internality. As higher perceived valuations translate into higher consumption levels of x, they can be exploited for benevolent price discrimination. Importantly, internality revelation occurs even in settings that are uninformative a priori, when valuations and biases are independently distributed, for example. Our example also illustrates that

extreme perceived valuations have higher informational value than less extreme ones. While $\hat{v}_1$ and $\hat{v}_3$ allow to perfectly infer consumers' bias, $\hat{v}_2$ does not.

How can internality revelation be exploited for optimal taxation? For simplicity, assume that the cost function is given by $c(x) = 0.5x^2$, so that consumers should optimally choose $x^e = v$ rather than $x^d = \hat{v}$, which corresponds to their choice based on perceived valuations $\hat{v}$. Now, consider the following mechanism that directly assigns an allocation x to a consumer, based on his report $\tilde{v}$:

$$\xi(\tilde{v}) = \begin{cases} \tilde{v} & \text{if } \tilde{v} = 1 \\ \tilde{v} - 0.5 & \text{if } \tilde{v} = 2 \\ \tilde{v} - 1 & \text{if } \tilde{v} = 3. \end{cases}$$

This direct mechanism determines allocations in a way that optimally corrects for the average bias of a consumer type $\hat{v}$, which a designer can infer through internality revelation.

Can the mechanism designer set up a tax schedule that would induce consumers to truthfully report their actual perceived valuations? Consider the following direct mechanism that assigns taxes as follows:

$$\tau(\tilde{v}) = \begin{cases} 0 & \text{if } \tilde{v} = 1 \\ 0.375 & \text{if } \tilde{v} = 2 \\ 1 & \text{if } \tilde{v} = 3. \end{cases}$$

In our discrete example, the tax schedule is a step function and represents an approximation to the optimal tax we have derived for the continuous case in Proposition 2. It is straightforward to show that the tax scheme would induce consumers to truthfully report their actual perceived valuations, i.e., that $u^d(\xi(\hat{v}), \tau(\hat{v})|\hat{v}) \geq u^d(\xi(\tilde{v}), \tau(\tilde{v})|\hat{v}) \; \forall \hat{v}, \tilde{v}$.[9] Note also that the tax schedule is convex: incremental tax increases are larger for consumers who report high perceived valuations and thus target consumers with larger biases.

[9]For example, when a consumer with perceived valuation $\hat{v}_3$ pretends to be of type $\hat{v}_2$, he realizes decision utility $u^d(\xi(\hat{v}_2), \tau(\hat{v}_2)|\hat{v} = \hat{v}_3) = 3 \cdot 1.5 - 0.375 - 0.5(1.5)^2 = 3$, while pretending to be of type $\hat{v}_1$ yields $u^d(\xi(\hat{v}_1), \tau(\hat{v}_1)|\hat{v} = \hat{v}_3) = 3 \cdot 1 - 0.5(1)^2 = 2.5$. Both values are not larger than decision utility under truthtelling $u^d(\xi(\hat{v}_3), \tau(\hat{v}_3)|\hat{v} = \hat{v}_3) = 3 \cdot 2 - 1 - 0.5(2)^2 = 3$. It can easily be shown that the same holds true for the other perceived valuation types.

5 Implications of Internality Revelation

We now investigate internality revelation in detail and explore its consequences for internality tax implementabililty and the properties of the optimal non-linear tax schedule. For the ease of exposition, we assume a utilitarian social welfare function with equal weights normalized to one for each consumer. In that case, the optimal marginal tax rate equals the expected bias conditional on the report, i.e., $t'(x) = E[b|\hat{v}_x]$ (Proposition 2).

Corollary 2 (Internality Revelation)**.** *The Minimum Mean Squared Error (MMSE) linear approximation of the conditional expectation $E[b|\hat{v}]$ is given by:*

$$\hat{E}[b|\hat{v}] = E[b|\mu_{\hat{v}}] + \underbrace{\frac{(\sigma_b/\sigma_v) + \rho}{(\sigma_b/\sigma_v) + (\sigma_v/\sigma_b) + 2\rho}}_{=:A} \cdot (\hat{v} - \mu_{\hat{v}}). \tag{7}$$

Internality revelation occurs whenever $A \neq 0$ and breaks down only in non-generic situations when $\rho = -(\sigma_b/\sigma_v)$. If v and b are independent or follow a bivariate normal distribution, we have $E[b|\mu_{\hat{v}}] = \mu_b$.

Proof. See Appendix A.E and Appendix C. □

Corollary 2 shows that internality revelation occurs whenever a designer can improve upon $E[b|\mu_{\hat{v}}]$ by using the information contained in a report $\hat{v}$. It illustrates that the magnitude of updating depends crucially on two factors. First, updating increases in the (absolute value of the) term A, which can be interpreted as the information value of a report $\hat{v}$ in a particular choice setting. The term A only depends on two sufficient statistics that can be readily estimated without individual-level data: the correlation coefficient ρ and the signal-to-noise ratio for b, σ_b/σ_v.[10] Second, the magnitude of updating increases in the distance between a report $\hat{v}$ and its expectation $\mu_{\hat{v}}$, which reflects that extreme reports are particularly informative. One implication of this finding is that biases are particularly large for consumers with extreme reports, who will thus be targeted by non-linear corrective taxes. Another implication is that the optimal non-linear tax scheme implies "no distortion at the top *and* at the bottom" when the distribution functions F and G both have a bounded support.[11]

[10] The comparative statics are as follows. For $\rho \geq 0$, the value of the information A increases in the signal-to-noise ratio σ_b/σ_v. For $\rho < 0$, it decreases in σ_b/σ_v for low or high values of σ_b/σ_v, while it increases in σ_b/σ_v for intermediate values of σ_b/σ_v. Furthermore, it increases in ρ if $\sigma_b/\sigma_v < 1$, and it decreases in ρ if $\sigma_b/\sigma_v \geq 1$.

[11] With bounded support, internality revelation resolves all uncertainty about the bias for the 'extreme' reports $\tilde{v} = \underline{\hat{v}}$ and $\tilde{v} = \overline{\hat{v}}$ from the boundaries of the support of P. As a result, consumers with the reports $\tilde{v} = \underline{\hat{v}}$ or $\tilde{v} = \overline{\hat{v}}$ obtain their optimal allocation under the optimal non-linear tax scheme.

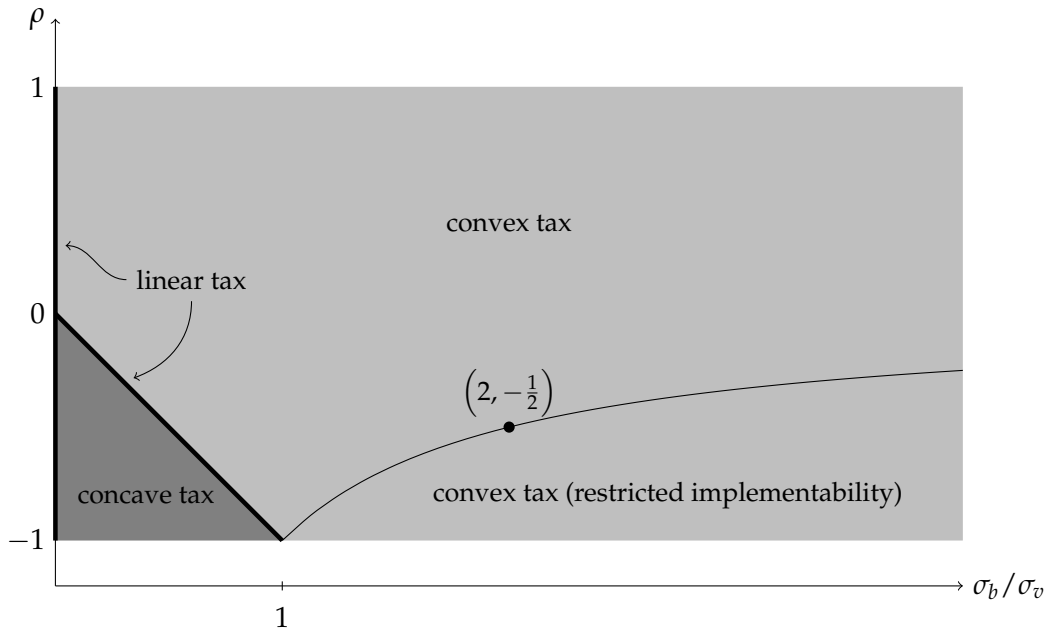


Figure 1: Shape of the optimal internality tax as a function of the sufficient statistics ρ and σ_b/σ_v.

We now link our findings on internality revelation to the optimal non-linear tax from Equation (6) and analyze the implications of non-linear taxation on welfare.

Proposition 3. *Optimal non-linear taxes (weakly) increase welfare beyond the optimal linear tax whenever internality revelation occurs, i.e., whenever $E[b|\hat{v}]$ is not constant in $\hat{v}$, or, equivalently, whenever term A in Equation* (7) *is not equal to zero.*

Proposition 3 reflects that giving the designer the freedom to construct a non-linear rather than a linear tax schedule cannot reduce welfare. Beyond that, it shows that knowledge about the statistics ρ and σ_b/σ_v is sufficient to judge whether non-linear taxation can improve upon linear taxation. The welfare gains of optimal non-linear taxation over linear taxation increase in the information value A of a report, as we demonstrate in Appendix A.F. This finding reflects that in settings with high information value, consumption choices reveal much about a consumer's bias, which can be exploited by non-linear taxation.

Next, we explore the implications of internality revelation for the properties of optimal non-linear tax schemes. As shown in Corollary 2, internality revelation depends crucially on the correlation between experienced valuations and biases, ρ, and the signal-to-noise ratio σ_b/σ_v. In Proposition 4, we characterize the optimal tax schedule in terms of these statistics and illustrate our findings in Figure 1.

Proposition 4. *The shape of the optimal tax schedule is determined by the correlation between experienced valuations and biases, ρ, and the signal-to-noise ratio σ_b/σ_v. Using Corollary 2 we obtain the following:*

1. *The optimal tax scheme is convex if and only if $A > 0 \Leftrightarrow \rho > -(\sigma_b/\sigma_v)$.*
   If $A > 1 \Leftrightarrow \rho < -(\sigma_v/\sigma_b)$, internality tax implementability is restricted.
2. *The optimal tax scheme is concave if and only if $A < 0 \Leftrightarrow \rho < -(\sigma_b/\sigma_v)$.*
3. *The optimal tax scheme is linear if and only if $A = 0$, i.e., if either $(\sigma_b/\sigma_v) = 0$ or $\rho = -(\sigma_b/\sigma_v)$.*

Proposition 4 has immediate practical relevance as knowledge on consumers' behavioral bias allows to infer the sign of ρ, even in the absence of empirical elicitations of that statistic. For example, exogenous inattention implies $\rho = 0$, while rational inattention implies that consumers with larger valuations have smaller biases, i.e., $\rho < 0$. In contrast, rational addiction models presume that consumers with larger valuations for an addictive good consume it and finally develop addictions, which would be captured by $\rho > 0$ in our setting. Corollary 3 links leading behavioral biases to the shape of the optimal non-linear tax schedule and shows that the optimal non-linear tax is typically convex.

Corollary 3. *For the following three leading behavioral biases, internality revelation induces a convex optimal non-linear tax scheme:*

1. $\rho > 0$ *(e.g., rational addiction),*
2. $\rho = 0$ *(e.g., exogeneous inattention),*
3. $-(\sigma_b/\sigma_v) < \rho < 0$ *(e.g., rational inattention with a large signal-to-noise ratio).*

With a convex tax schedule, the marginal tax rate increases in the consumption level, which implies that taxes target high types $\hat{v}$, while subsidies target low types. Convexity reflects the core finding from internality revelation that individuals with higher reports (who consume more of good x) have higher biases. We conclude this section by investigating how internality tax implementability and internality revelation interact.

Corollary 4 (Trade-off Between Internality Revelation and Implementability)**.** *The necessary and sufficient condition ensuring internality tax implementability, $\rho > -\sigma_v/\sigma_b$, becomes more restrictive, as the signal to noise ratio σ_b/σ_v increases.*

Corollary 4, which is also visualized in Figure 1, states that implementability is guaranteed only if internality revelation is not too pronounced. In particular, the trade-off between implementability and internality revelation in our setting reflects that consumers' perceived preferences and the normative stance of the designer, given by $E[v|\hat{v}]$, may not diverge excessively. For instance, at the point $\left(2, -\frac{1}{2}\right)$ in Figure 1, an increase in σ_b/σ_v would increase the information value A. However, the designer cannot exploit the additional information as implementability would fail if she did.

6 Optimal Non-Linear Taxation in the Light Bulb Market

We now apply our results to determine optimal subsidies for energy efficiency in the light bulb market. We take the demand-side data from Allcott and Taubinsky (2015), who investigate consumer choices between an incandescent and compact florescent light (CFL) bulbs, and elicit comprehensive information on consumers' energy efficiency valuations and time preferences. For the supply side, we retrieve product data and prices for a wide range of light bulbs that give us a comprehensive picture of the light bulb varieties that were on the market at the time of the study by Allcott and Taubinsky (2015). We start by discussing both data sources and detail how we approximate the cost function for energy efficiency, consumers' experienced valuations v, perceived valuations $\hat{v}$, and bias b. Based on our theoretical results, we then determine the optimal non-linear tax on energy efficiency and investigate its welfare implications in the light bulb market.

Our supply data stems from the price comparison website *geizhals.de*. This website reports the cheapest price of a product for all months since it is offered on a website in the internet. We focus on light bulbs that are typically purchased by households. In particular, we consider bulbs with an energy intensity of 25 to 75 Watt-equivalents and a warm light color of around 2700 Kelvin. To reduce the impact of branding effects, we focus on bulbs produced by one of the large manufacturers *Osram* and *Philips* that offer bulbs both in the EU and the US. As in Allcott and Taubinsky (2015), we express all prices in 2012 US dollars (USD) and extract product prices during that year. Some LED and CFL bulbs enter the market after 2012 – in these cases, we extrapolate their 2012 price based on aggregate annual price trends that imply a 20% and 10% price decrease per annum for LED and CFL bulbs, respectively. For every bulb, we determine the operating and replacement cost to consume 8.000 hours of light over eight years, which corresponds to three hours per day, assuming electricity prices of 0.1 USD per kWh (Allcott and Taubinsky, 2015).

For every bulb, we determine the purchase price premium and the operating and replacement cost (ORC) saving relative to the most electricity intensive bulb. In the fol-

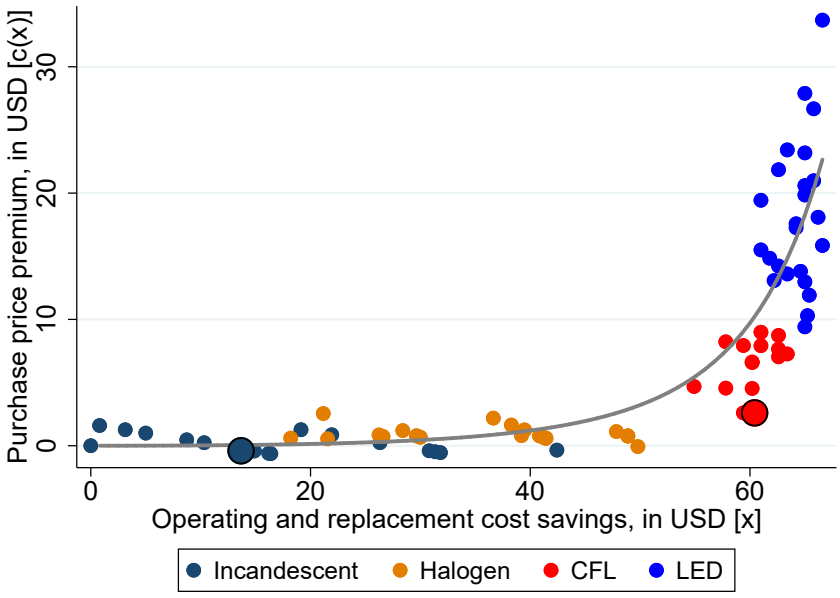


Figure 2: Energy Efficiency Cost Function in the Light Bulb Market. *Note:* Price premiums, as well as operating and replacement cost savings are determined relative to the most electricity intensive bulb. Operating and replacement cost assume eight years of total usage (8.000 hours) and an electricity price of 0.1 USD per kWh, as in Allcott and Taubinsky (2015).

lowing, we use ORC savings as the measure of quality x, i.e., energy efficiency. Figure 2 draws price premiums against ORC savings, which corresponds to the cost function $c(x)$ in our model. The least energy inefficient, yet cheapest, bulbs are incandescents, followed by halogen, CFL and LED bulbs. The cost curve is convex, which illustrates that a one unit reduction in ORC savings becomes increasingly more expensive as the level of energy efficiency increases. In 2012, the most energy efficient LED bulbs sold at a price premium of around 30 USD and yielded cost savings of some 60 USD over the course of 8 years, compared to the most energy intensive incandescent bulb.

For the demand side, we determine consumers' valuation and bias for energy efficiency based on the experimental data from Allcott and Taubinsky (2015). To identify the distribution of the bias, we rely on the identification strategy by Allcott and Taubinsky (2015). In their experiment, consumers choose between two bulbs, an incandescent and a CFL bulb, which are highlighted in Figure 2 by large dots. After a baseline elicitation of relative WTP for the more energy efficient CFL bulb, Allcott and Taubinsky (2015) present lifetime cost information to consumers and then measure relative WTP again. Under the assumption that providing lifetime cost information eliminates all bias without distorting choices otherwise, the difference in relative WTP

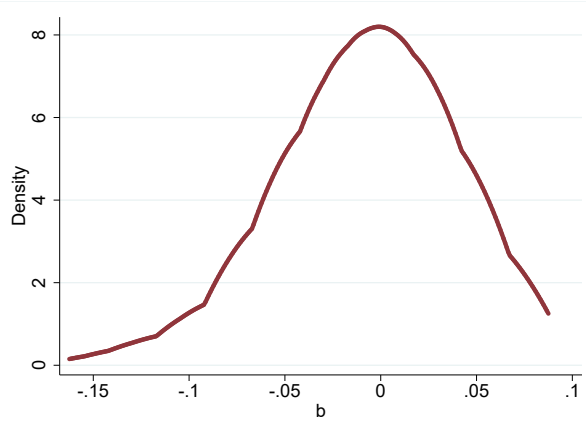
identifies the distribution of the bias in the experimental population. We divide this bias measure by the operating and replacement cost difference between both bulbs, which yields a per unit bias (in terms of x), whose distribution is presented in Figure 3a.

Data on individual valuations v are unavailable, but we exploit that differences in time preferences naturally induce heterogeneity in consumers' experienced valuation of ORC savings. In particular, the net present value (NPV) of a 1 USD increase in total ORC savings is 1 USD only if consumers' discount rate were exactly equal to zero. In contrast, if consumers discount future savings by a discount rate δ, the NPV of 1 USD amounts to a discount factor $D(\delta)$. For illustration, assume a consumer with a high discount rate $\delta = 20\%$ and annual operating cost of $1/8$ USD for 8 years, which results in a total cost of 1 USD, the unit that we use for x. Her NPV of this 1 USD cost increase is then $D(\delta) = (1 + 1/(1+\delta) + \cdots + 1/(1+\delta)^7) \cdot (1/8) = 0.58$ USD.

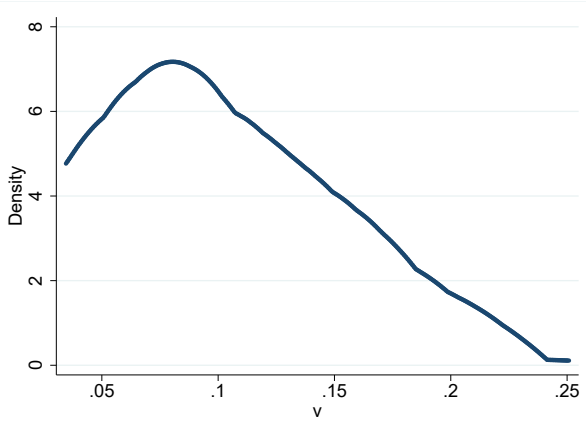
We use the elicitation of time preferences by Allcott and Taubinsky (2015) to determine individual-specific discount factors. We assume that all other factors that influence experienced valuations do not vary by participant and thus constitute merely a scaling factor. This assumption allows us to calibrate valuations to match actual consumer choices in Allcott and Taubinsky (2015). In particular, we set valuations to $v = s \cdot D(\delta)$, where s is a scaling factor that ensures that the same percentage of consumers would choose the more energy efficient CFL light bulb when confronted with the two bulbs from the endline elicitation in Allcott and Taubinsky (2015). While this approach is clearly imperfect, it serves as a useful approximation that allows us to illustrate how our results can be implemented.[12]

Based on the joint distribution of v and b, we can determine the distribution of $\hat{v}$ and the conditional expectation of the bias $E[b|\hat{v}]$, which is crucial for determining the optimal non-linear tax. Figure 3c shows that the $E[b|\hat{v}]$ is increasing in $\hat{v}$. It also illustrates that our analytical solution for the minimum mean square error (MMSE) approximation closely fits the actual conditional expectation $E[b|\hat{v}]$ in our setting. Accordingly, we can approximate the shape of the optimal non-linear tax based on the sufficient statistics ρ and σ_v/σ_b. In our data, the correlation between v and b is $\rho = -0.21$, and the standard deviation of b and v are $\sigma_b = 0.0396$ and $\sigma_v = 0.0469$, respectively. These statistics imply that the conditional expectation $E[b|\hat{v}]$ increases in $\hat{v}$, as $\rho > -\sigma_b/\sigma_v$ (Corollary 2).

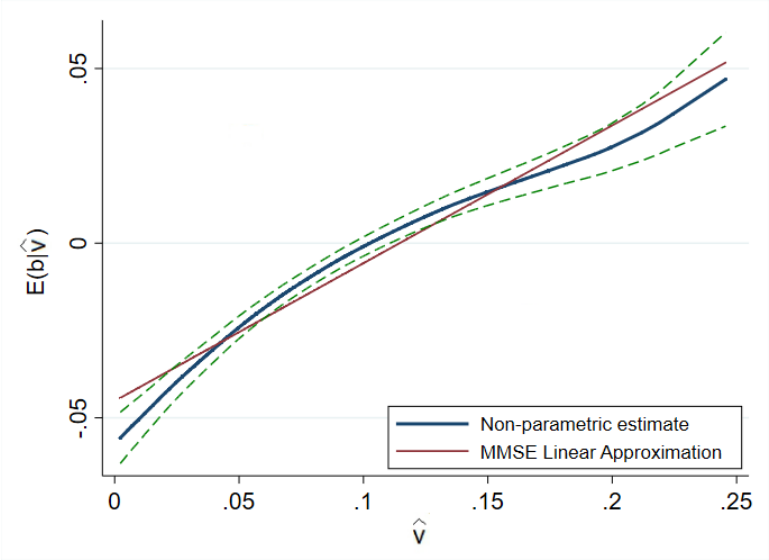
[12] We need to impose some consistency restrictions on the data that go beyond Allcott and Taubinsky (2015). For example, we disregard observations that have missing values for discount rates and biases. In addition, to avoid negative perceived valuations, we omit extreme values for v and b and disregard the top and bottom 2% of observations. We also drop observations if – after rescaling v – perceived valuations would become negative, which leaves us with 668 observations. Our sample restrictions tend drop observations with larger negative biases, so that we underestimate the welfare gains of non-linear taxation.



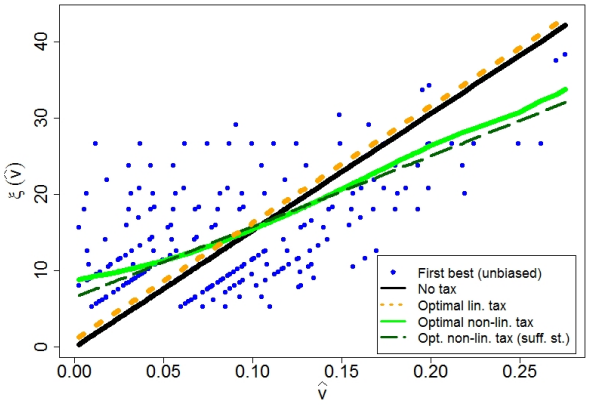
(a) Distribution of bias b, estimated via kernel density estimation (Epanechikov kernel, bandwidth: 0.03).



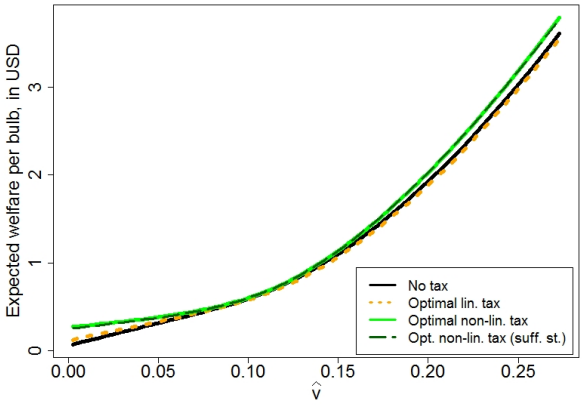
(b) Distribution of experienced valuation v, estimated via kernel density estimation (Epanechikov kernel, bandwidth: 0.03).



(c) Conditional expectation of the bias $E[b|\hat{v}]$, estimated via local linear regression (Epanechnikov kernel, rule-of-thumb bandwidth).



(d) Choices of x in the absence of a tax ("no tax"), the optimal linear tax, and the optimal non-linear tax. Blue dots are optimal outcomes for every individual, given v.



(e) Expected welfare per bulb and perceived valuation $\hat{v}$.

Figure 3: Optimal Non-Linear Taxation in the Light Bulb Market

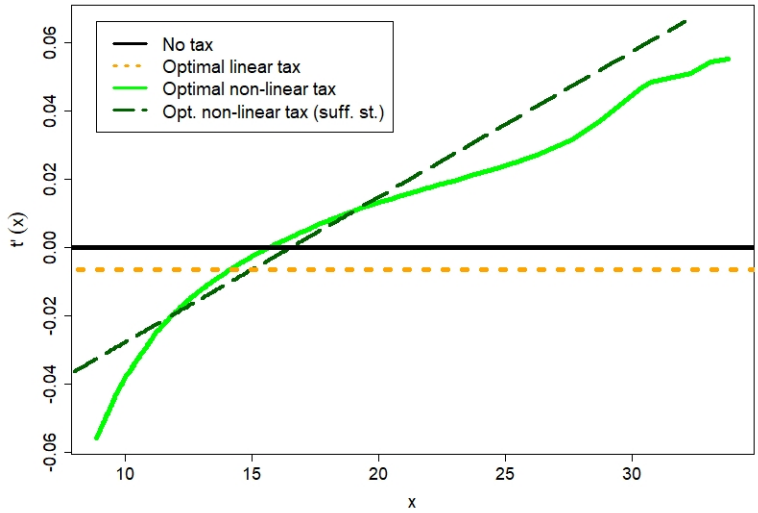


Figure 4: Optimal tax rates as a function of x

In the following, we assume a quadratic cost function and estimate it based on the data from Figure 2. We derive consumers' choices in four scenarios: the absence of a corrective tax, the presence of the optimal linear tax, the optimal non-linear tax, and the MMSE approximation of the optimal non-linear tax based on sufficient statistics (Corollary 2), which we depict in Figure 3c.

Figure 3d shows how consumers of type $\hat{v}$ choose quantities in these scenarios and contrasts these choices with the first-best outcome that materialized if consumers decided based on experienced valuations v rather than perceived valuations $\hat{v}$, depicted as blue dots. The optimal linear tax induces a parallel shift in demand, which improves choices for consumers with low perceived valuations $\hat{v}$, but distorts choices for consumers with large perceived valuations. In contrast, the optimal non-linear tax corrects consumers' choices in a flexible manner. In addition, the difference to the MMSE approximation of the non-linear tax is very small. Accordingly, knowing a few sufficient statistics is enough to implement a good approximation to the optimal non-linear tax. In Figure 3e, we plot the expected welfare by type and show that the non-linear tax induces a welfare improvement for all consumer types. In contrast, the optimal linear tax increases welfare only for consumers with low perceived valuations, but at the cost of reducing welfare for consumers with high perceived valuations.

Figure 4 visualizes the optimal tax rate as a function of x, which corresponds to the indirect mechanism in our model.[13] We find that the optimal tax rate increases

[13]We can also determine marginal tax rates in the direct mechanism, i.e., as a function of $\hat{v}$, which we show in Figure 5 in Appendix D

in $\hat{v}$ (and x), which implies that the optimal non-linear tax for energy efficiency in the light bulb market is convex. Intuitively, this result reflects that individuals with a large positive bias b tend to choose better qualities x, while the opposite holds true for individuals with large negative biases. The convex tax schedule exploits this information and imposes marginal taxes on consumers with large perceived valuations that intend to buy an energy efficient product in the first place, while giving marginal subsidies to consumers that intend to buy energy inefficient products.

7 Discussion and Conclusion

In this paper, we have derived the optimal non-linear tax to correct behaviorally biased consumers. Using a mechanism design approach, we show that consumers' consumption choices contain information that can be employed to improve upon a linear tax. We investigate the information structure implied by internalities and show that consumers reveal their type even when the aggregate information available to a policy maker is uninformative, e.g., when the bias has zero mean and is uncorrelated with valuations.

We explore such internality revelation in more detail and show that it crucially depends on two sufficient statistics: the correlation between biases and valuations and the ratio of their standard deviations, which corresponds to the "signal-to-noise" ratio of the bias. We derive that the slope of the expected conditional bias is a function of the informativeness of the choice setting, characterized by the two above mentioned sufficient statistics, and the informativeness of a consumer's report, as given by the distance of his report to the average report. We also demonstrate that – as a consequence of internality revelation – the optimal non-linear corrective tax is typically convex, and analytically characterize the exceptions when this finding does not hold true.

More broadly, we contribute to the literature on behavioral mechansim design by finding a novel trade-off between internality revelation and internality tax implementability, which we characterize in terms of the sufficient statistics mentioned above. Even when standard conditions such as an increasing hazard rate are satisfied, implementability of the behavioral mechanisms is not guaranteed, as internalities can induce disagreement among the designer and consumers on what is best for the latter. Applying our results to the light bulb market, we demonstrate that optimal non-linear taxation can increase welfare beyond the optimal linear tax. Furthermore, we illustrate that the informational requirements for implementing an approximation of the optimal non-linear tax are low and require only knowledge about few sufficient statistics.

Our finding that the expected conditional bias is revealed by consumers' choices is supported by empirical evidence. For example, Allcott et al. (2015) show for energy efficiency investments and hybrid car purchases that perceived valuations are positively correlated with the bias. As we have shown, in such cases, the optimal non-linear tax scheme is convex, giving the largest marginal subsidies to participants with low perceived valuations. The insight that internality taxes are typically convex also informs the design of policies more generally. Many subsidy schemes employed in practice are effectively antipodal to the optimal non-linear tax derived in this paper. For example, the German government grants subsidies for energy efficiency in housing only if a newly built (or retrofitted) house meets predefined minimum efficiencies, so-called "KfW-Effizienzhaus" standards. In other words, marginal subsidies are essentially zero for consumers with low perceived valuations. As a consequence, the most heavily biased consumers receive no subsidies and have thus no incentive for behavioral change, which can substantially hamper the effectiveness of a policy. By specifically targeting those consumers, implementing a non-linear subsidy promises substantial welfare improvements.

A Proofs

A.A Proof of Corollary 1: Dimensionality of Internality Tax

We want to show that two-dimensional mechanisms, where sophisticated consumers report both their perceived valuation and their bias, induce a violation of truthtelling and thus cannot be implemented. Without loss of generality, we assume that there exists at least one realization of perceived valuations $\hat{v}_1 \in \left[\underline{\hat{v}}, \overline{\hat{v}}\right]$ for which biases differ, so that some consumers are characterized by ($\hat{v}_1$, b_1) and others by ($\hat{v}_1$, b_0), where $b_1 \neq b_0$.

The mechanism designer wants to implement a direct two-dimensional mechanism where the allocation $\xi\left(\tilde{v}, \tilde{b}\right)$ and the tax $\tau\left(\tilde{v}, \tilde{b}\right)$ depend on reported valuations $\tilde{v}$ and biases $\tilde{b}$. Consumers choose their reports to maximize decision utility:

$$(\tilde{v}^*(\hat{v}), \tilde{b}^*(\hat{v})) = \arg\max_{\tilde{v},\tilde{b}} u^d\left(\xi\left(\tilde{v}, \tilde{b}\right), \tau\left(\tilde{v}, \tilde{b}\right) | \hat{v}\right).$$

Importantly, this maximization problem depends only on reported biases $\tilde{b}$ and not on actual biases b. As a consequence, every sophisticated consumer with perceived valuation $\hat{v}$ will report the same bias $\tilde{b}^*(\hat{v})$. As biases differ for $\hat{v}_1$, truthtelling is violated.

A.B Proof of Proposition 1: Internality Tax Implementability

We first establish some helpful lemmata.

Lemma 1 (Monotonic Allocation Rule)**.** *ξ is non-decreasing in $\hat{v}$.*

Proof. Let $\hat{v}_1 > \hat{v}_2$. Incentive compatibility implies

$$\begin{aligned}
&\hat{v}_1\xi(\hat{v}_1) - c[\xi(\hat{v}_1)] - \tau(\hat{v}_1) \geq \hat{v}_1\xi(\hat{v}_2) - c[\xi(\hat{v}_2)] - \tau(\hat{v}_2), \text{ and} \\
&\hat{v}_2\xi(\hat{v}_1) - c[\xi(\hat{v}_1)] - \tau(\hat{v}_1) \leq \hat{v}_2\xi(\hat{v}_2) - c[\xi(\hat{v}_2)] - \tau(\hat{v}_2).
\end{aligned}$$

Subtracting the second line from the first yields $(\hat{v}_1 - \hat{v}_2)\xi(\hat{v}_1) \geq (\hat{v}_1 - \hat{v}_2)\xi(\hat{v}_2) \Leftrightarrow \xi(\hat{v}_1) \geq \xi(\hat{v}_2)$. Since this holds for all $\hat{v}_1, \hat{v}_2$ with $\hat{v}_1 > \hat{v}_2$, we know $\xi'(\hat{v}) \geq 0$. □

Lemma 2 (Properties of Equilibrium Decision Utility)**.** *$\hat{u}^d$ increases and is convex in $\hat{v}$ with $\partial\hat{u}^d/\partial\hat{v} = \xi(\hat{v})$.*

Proof. Equation (3) shows that $\partial\hat{u}^d/\partial\hat{v} = \xi(\hat{v})$. By Lemma 1, we then know that $\hat{u}^d$ is convex in $\hat{v}$. Note that the previous step assumes differentiability; it can also be shown without this assumption by using, for instance, Theorem 18 on page 132 in

Royden and Fitzpatrick (2010). Define $\tilde{u}^d(\hat{v}) := \hat{v}\xi(\tilde{v}) - c[\xi(\tilde{v})] - \tau(\tilde{v})$. By definition, $\hat{u}^d(\hat{v}) = \max_{\tilde{v}} \tilde{u}^d(\hat{v})$. Obviously, $\tilde{u}^d(\hat{v})$ increases in $\hat{v}$. Since $\hat{u}^d(\hat{v})$ is the maximum of increasing functions, it is increasing in $\hat{v}$. □

Lemma 3 (Consumer Utility Pinned Down by $\hat{u}^d(\underline{\hat{v}})$ and ξ)**.** $\hat{u}^d(\hat{v}) = \hat{u}^d(\underline{\hat{v}}) + \int_{\underline{\hat{v}}}^{\hat{v}} \xi(\theta)d\theta$.

Proof. By Lemma 2, $\hat{u}^d(\hat{v})$ is convex, so it is also absolutely continuous, see, for instance, Corollary 17 on page 132 in Royden and Fitzpatrick (2010). This in turn implies $\hat{u}^d(\hat{v}) = \hat{u}^d(\underline{\hat{v}}) + \int_{\underline{\hat{v}}}^{\hat{v}} \frac{\partial \hat{u}^d(\theta)}{\partial \theta} d\theta \overset{L1}{=} \hat{u}^d(\underline{\hat{v}}) + \int_{\underline{\hat{v}}}^{\hat{v}} \xi(\theta)d\theta$, see, for instance, Theorem 10 on page 124 and Proof of Theorem 11 on page 125 in Royden and Fitzpatrick (2010). □

Lemma 4 (Characterization of τ in terms of ξ and Designer's Revenue)**.** $\tau(\hat{v}) = -\hat{u}^d(\underline{\hat{v}}) + \hat{v}\xi(\hat{v}) - \underbrace{\left\{ c[\xi(\hat{v})] + \int_{\underline{\hat{v}}}^{\hat{v}} \xi(\theta)d\theta \right\}}_{\textit{information rent}}$.

Proof. Plugging in the definition of $\hat{u}^d(\hat{v})$ in the formula of Lemma 3 and rearranging gives the result. □

Lemma 5. $\int_{\hat{v}} \int_{\underline{\hat{v}}}^{\theta} \xi(t)dt p(\theta)d\theta = \int_{\hat{v}} \xi(t)[1 - P(t)]dt$.

Proof. We can rewrite

$$\int_{\hat{v}} \int_{\underline{\hat{v}}}^{\theta} \xi(t)dt p(\theta)d\theta = \int_{\hat{v}} \int_{\underline{\hat{v}}}^{\hat{v}} \xi(t)p(\theta)dtd\theta \overset{(*)}{=} \int_{\hat{v}} \int_{t}^{\overline{\hat{v}}} \xi(t)p(\theta)d\theta dt = \int_{\hat{v}} \xi(t) \int_{t}^{\overline{\hat{v}}} p(\theta)d\theta dt$$
$$= \int_{\hat{v}} \xi(t)[1 - P(t)]dt,$$

using Fubini's Theorem in $(*)$. □

Plugging the result from Lemma 4 into the designer's objective yields the following function which only depends on the variable ξ

$$
\begin{aligned}
&\int_{\hat{v}} \alpha(\theta) \Bigg\{ \underbrace{m+\theta\xi(\theta)-c[\xi(\theta)]-\overbrace{\Bigg[-\hat{u}^d(\underline{\hat{v}})+\theta\xi(\theta)-c[\xi(\theta)]-\int_{\underline{\hat{v}}}^{\theta}\xi(t)dt\Bigg]}^{\tau(\theta)}}_{u^d(\hat{v})} - E[b|\theta]\xi(\theta) \Bigg\} p(\theta)d\theta \\
&+\int_{\hat{v}} \underbrace{\Bigg[-\hat{u}^d(\underline{\hat{v}})+\theta\xi(\theta)-c[\xi(\theta)]-\int_{\underline{\hat{v}}}^{\theta}\xi(t)dt\Bigg]}_{\tau(\theta)} p(\theta)d\theta \\
=&\int_{\hat{v}} \alpha(\theta)\{m-E[b|\theta]\xi(\theta)\}p(\theta)d\theta+\int_{\hat{v}}\{1-\alpha(\theta)\}\Bigg\{-\hat{u}^d(\underline{\hat{v}})-\int_{\underline{\hat{v}}}^{\hat{v}}\xi(t)dt\Bigg\}p(\theta)d\theta \\
&+\int_{\hat{v}}\{\theta\xi(\theta)-c[\xi(\theta)]\}p(\theta)d\theta \\
\overset{L5}{=}&\int_{\hat{v}} \alpha(\theta)\{m-E[b|\theta]\xi(\theta)\}p(\theta)d\theta+\int_{\hat{v}}\{1-\alpha(\theta)\}\Bigg\{-\hat{u}^d(\underline{\hat{v}})-\xi(\theta)\tfrac{1-P(\theta)}{p(\theta)}\Bigg\}p(\theta)d\theta \\
&+\int_{\hat{v}}\{\theta\xi(\theta)-c[\xi(\theta)]\}p(\theta)d\theta
\end{aligned}
$$

Now take the first-order condition with respect to ξ for one specific $\hat{v}$ (we can reintroduce the "dynamic nature" of the optimization problem by finding a condition later on which guarantees that incentive compatibility across the types is satisfied). Rearrange the first-order condition to obtain an implicit characterization of the optimal allocation rule

$$
c'[\xi(\hat{v})] \overset{!}{=} \hat{v} - \alpha(\hat{v}) \cdot E[b|\hat{v}] - (1-\alpha(\hat{v}))\,\tfrac{1-P(\hat{v})}{p(\hat{v})}. \tag{8}
$$

The left-hand side increases in $\hat{v}$ if and only if ξ increases in $\hat{v}$, since c is convex. Remember, that we need to guarantee that ξ increases in $\hat{v}$ to obtain an incentive compatible mechanism. Thus, the right-hand side needs to be increasing in $\hat{v}$, which implies

$$
1-\tfrac{\partial\alpha(\hat{v})}{\partial\hat{v}}E[b|\hat{v}]-\tfrac{\partial E[b|\hat{v}]}{\partial\hat{v}}\alpha(\hat{v})+\tfrac{\partial\alpha(\hat{v})}{\partial\hat{v}}\cdot\tfrac{1-P(\hat{v})}{p(\hat{v})}-\{1-\alpha(\hat{v})\}\tfrac{\partial[1-P(\hat{v})/p(\hat{v})]}{\partial\hat{v}} \geq 0.
$$

A.C Proof of Proposition 2: Optimal Non-Linear Tax

The consumer's first-order condition characterizing optimal consumption x^d is given by

$$
\left.\frac{\partial u^d(x,t,\hat{v})}{\partial x}\right|_{x=x^d} = \hat{v} - c'(x^d) - t'(x^d) \overset{!}{=} 0 \Leftrightarrow c'(x^d) = \hat{v} - t'(x^d). \tag{9}
$$

The second order condition is satisfied if $-c''(x) - t''(x) \leq 0$ for all $x \in X$. Since the costs are convex in x by assumption, this condition is satisfied, if the optimal tax

schedule is convex in x as well.[14] Generally, an interior solution exists, if "c is convex enough compared to t", i.e., $c''(x) \geq -t''(x)$ for all $x \in X$.

As discussed in the text we can always guarantee that $\hat{u}\left(\underline{\hat{v}}\right) = \underline{u} > 0$ and $\hat{u}\left(\overline{\hat{v}}\right) \geq \underline{u}$, so that the transversality condition immediately implies $\mu\left(\underline{\hat{v}}\right) = 0$ and $\mu\left(\overline{\hat{v}}\right) = 0$. We now use Equation ($\text{FOC}_\text{u}$). By integrating and using $\mu\left(\overline{\hat{v}}\right) = 0$ we obtain

$$\int_{\hat{v}}^{\overline{\hat{v}}} -\mu'(n)dn = -\mu\left(\overline{\hat{v}}\right) - \left[-\mu\left(\hat{v}\right)\right] = \mu\left(\hat{v}\right) \stackrel{!}{=} \int_{\hat{v}}^{\overline{\hat{v}}} \left[\alpha(m) - 1\right] p(m)dm. \tag{10}$$

Using the above equations we rearrange Equation (FOC_x), to obtain the result:

$$\begin{aligned}
&\left(\hat{v} - c'(\cdot)\right) \stackrel{!}{=} -\frac{\mu\left(\hat{v}\right)}{p(\hat{v})} + E[b|\hat{v}] \cdot \alpha(\hat{v}) \\
\stackrel{(9)}{\Leftrightarrow} \quad &t'(x) = -\frac{\mu\left(\hat{v}_x\right)}{p(\hat{v}_x)} + E[b|\hat{v}_x] \cdot \alpha(\hat{v}_x) \\
\stackrel{(10)}{\Leftrightarrow} \quad &t'(x) = \frac{\int_{\hat{v}_x}^{\overline{\hat{v}}} \left[1 - \alpha(m)\right] p(m)dm}{p(\hat{v}_x)} + E[b|\hat{v}_x] \cdot \alpha(\hat{v}_x).
\end{aligned}$$

A.D Derivation of the Optimal Linear Tax

In this proof we additionally assume that $c'''(x) = 0$, which simplifies the calculation of the optimal linear tax, but does not change our results on the optimal non-linear tax scheme. Anticipating the behavior on the consumer side, the problem of the designer can be written as $\max_{t \in \mathbb{R}} \int_v \int_b u^e(x^d, t, v)dG(b|v)dF(v) + \int_v \int_b t \cdot x^d dG(b|v)dF(v) =: V(t)$. We evaluate the derivative with respect to the linear tax t:

$$\begin{aligned}
\frac{\partial V(t)}{\partial t} &= \int_v \int_b \left[-x^d + \left(v - t - c'(\cdot)\right)\frac{\partial x^d}{\partial t}\right] dG(b|v)dF(v) + \int_v \int_b \left[x^d + t \cdot \frac{\partial x^d}{\partial t}\right] dG(b|v)dF(v) \\
&= \int_v \int_b \left[\left(v - c'(\cdot)\right)\frac{\partial x^d}{\partial t}\right] dG(b|v)dF(v).
\end{aligned}$$

The individually optimal consumption is again characterized by Equation (9), i.e., $c'(\cdot) = \hat{v} - t'(x) = (v + b) - t$, where the last equality holds since t is linear. Thus, $\frac{\partial V}{\partial t} = \int_v \int_b \left[(t - b)\frac{\partial x^d}{\partial t}\right] dG(b|v)dF(v)$. Using that t is constant, we can rewrite the equation as follows: $\frac{\partial V}{\partial t} = t\frac{\partial \bar{x}^d}{\partial t} - \int_v \int_b \left[b\frac{\partial x^d}{\partial t}\right] dG(b|v)dF(v)$, where the change in total demand $\bar{x}^d$ in response to a tax increase is given by $\frac{\partial \bar{x}^d}{\partial t} = \int_v \int_b \left[\frac{\partial x^d}{\partial t}\right] dG(b|v)dF(v)$. The optimal tax t^* is given by $\frac{\partial V}{\partial t}|_{t=t^*} \stackrel{!}{=} 0 \Leftrightarrow t^* = \int_v \int_b b\left[\frac{\partial x^d}{\partial t} / \frac{\partial \bar{x}^d}{\partial t}\right] dG(b|v)dF(v)$, where $\frac{\partial x^d}{\partial t} / \frac{\partial \bar{x}^d}{\partial t}$ denotes the relative responsiveness of a consumer type (v, b), i.e., the

[14] Corollary 3 illustrates that this is the case in many leading examples.

change in demand for that consumer type in repsonse to a tax increase, relative to change in total demand.

Assuming that $c'''(\cdot) = 0$ and that c is convex, we can further evaluate $\frac{\partial x^d}{\partial t}$. Differentiating Equation (9) with respect to t yields $\frac{\partial x^d}{\partial t} = -\frac{1}{c''(x^d)} = a$, for some real-valued constant $a < 0$. Accordingly, the optimal tax t^* is given by $t^* = E[b]$.

A.E Proof of Corollary 2: Minimum Mean Squared Error Approximation

As a consequence of the Regression Conditional Expectation Function Theorem (Angrist and Pischke, 2009), the Minimum Mean Squared Error (MMSE) linear approximation of the conditional expectation $E\left[b|\hat{v}\right]$ is given by:

$$\hat{E}\left[b|\hat{v}\right] = E(b|\hat{v} = \mu_{\hat{v}}) + \frac{cov(b,\hat{v})}{\sigma^2_{\hat{v}}} \cdot \left[\hat{v} - \mu_{\hat{v}}\right]. \tag{11}$$

Using that $cov(b,\hat{v}) = cov(b,v) + \sigma^2_b$, $\sigma^2_{\hat{v}} = \sigma^2_v + \sigma^2_b + 2\,cov(b,v)$, and $cov(b,v) = \rho\sigma_v\sigma_b$, we obtain after rearranging:

$$\hat{E}\left[b|\hat{v}\right] = E\left[b|\mu_{\hat{v}}\right] + \frac{\rho + (\sigma_b/\sigma_v)}{(\sigma_b/\sigma_v) + (\sigma_v/\sigma_b) + 2\rho} \cdot \left[\hat{v} - \mu_{\hat{v}}\right]. \tag{12}$$

More broadly, the existence of the conditional expectation as a function of the $\hat{v}$ is guaranteed by the Factorization Lemma and the Radon-Nikodym theorem.

A.F Details on the Welfare Effects of an Increase in A

In the following we want to show that, for every $\hat{v}$, the relative welfare increase through non-linear taxation, i.e., the difference in experienced utility between implementing the optimal non-linear and the optimal linear tax, increases in A. We assume that the support of P is unchanged due to the increase in A, i.e., there are no types $\hat{v}$ due to the increase in A that do not exist without the increase.

Because of quasi-linear utility and welfare weights being normalized to one, we can evaluate welfare implications of an increase in A ignoring the impacts of this change on the collected tax money. Let $E[v|\hat{v}](A) = \hat{v} - E[b|\hat{v}](A)$ denote the expected valuation, which is a function of A. To evaluate welfare implications of a change in A we need to evaluate the derivative with respect to A of the expected equilibrium experienced utility (net of taxes) for consumers with $\hat{v}$, which we denote as $u^e(A)$ (see also the Hamiltonian of the designer's problem):

$$\hat{u^e}(A) = E[v|\hat{v}](A)x^{alloc}(A) - c\left(x^{alloc}(A)\right). \tag{13}$$

We need to compare that derivative for two allocation rules: the allocation rule allocation rule implied by non-linear taxation, $x^{alloc} = \xi(A)$, and that implied by linear taxation, where $x^{alloc} = x^L$ is independent of A. Using the implicit characterization of the optimal allocation rule in Equation (8) with $\alpha(\hat{v}) = 1$, we know by Equation (7) that the RHS of Equation Equation (8) is decreasing in A whenever $\hat{v} > \mu_{\hat{v}}$, which implies that in this case $\xi'(A) < 0$, since purchase cost c is increasing in ξ.

Let us proceed without loss of generality with the assumption that $\hat{v} > \mu_{\hat{v}}$ (the reasoning is analogous for the reverse inequality). If the designer were to use the information contained in the consumer's report, i.e., $x^{alloc} = \xi(A)$, then

$$\frac{\partial \hat{u}^e(A)}{\partial A} = \underbrace{\frac{\partial E[v|\hat{v}]}{\partial A}\xi}_{<0} + \underbrace{\frac{\partial \xi}{\partial A}E[v|\hat{v}]}_{<0} - \underbrace{c'(\xi)\frac{\partial \xi}{\partial A}}_{<0}.$$

The first summand captures the fact that for a change in A, the designer learns that the consumer's misoptimization is more severe. The second and the third summand capture the fact that for a change in A the optimal allocation rule will prescribe a different consumption level to the consumer, which in the case of $\hat{v} > \mu_{\hat{v}}$ is a decrease in consumption. The second summand stands for the negative impact this has on the expected experienced consumption utility. The third summand stands for the negative impact this has on the consumer's expended purchase costs.

If the designer were not to use the information contained in the consumer's report, $x^{alloc} = x^L$, then

$$\frac{\partial \hat{u}^e(A)}{\partial A} = \frac{\partial E[v|\hat{v}]}{\partial A}x^L < 0.$$

The change in welfare gains is described by the difference of the two above expressions,

$$\underbrace{\overbrace{\frac{\partial E[v|\hat{v}]}{\partial A}}^{<0}\overbrace{\left[\xi - x^L\right]}^{<0}}_{\geq 0} + \underbrace{\overbrace{\frac{\partial \xi}{\partial A}}^{<0}\left[E[v|\hat{v}] - c'(\xi)\right]}_{=0},$$

where fact that the second summand is zero follows from optimal consumer behavior as described in Equation (9) and optimal non-linear taxtion, i.e., $t'(x^d) = E[b|v_{x^d}]$.

B A Model with Internalities and Externalities

Let us assume the following model that takes into account both internalities and externalities. There are n different consumers, indexed i, and i's decision utility can be written as

$$u_i^d(x_i, t, v_i, b_i) = m + (v_i + b_i)x_i - t(x_i) - c(x_i),$$

where, as in the main part of the paper, the random variable b_i reflects the bias of consumer i caused by an internality. Consumption of x also causes an externality β_i. Accordingly, the "normative" utility, an analogue to the experienced utility in the main part of the paper, which internalizes the costs consumer i inflicts on the other consumers, can be written as

$$u_i^n(x_i, t, v_i, \beta_i) = m + v_i x_i - t(x_i) - c(x_i) - \beta_i x_i.$$

To allow for individual heterogeneity in consumption externalities, we index β by i. For local externalities such as air pollution, β_i would reflect the damage on i's neighbors, while β_i would be constant in the case of global externalities such as greenhouse gas emissions. The consumer maximizes his decision utility and thus ignores both externalities and internalities. Note, that β_i does not have an impact on the optimal decision $x_i^d(v_i, b_i, t)$ as perceived by the consumer. We can rewrite

$$u_i^n(x_i, t, v_i, \mathfrak{b}_i) = u_i^d(x_i, t, v_i, b_i) - \mathfrak{b}_i x_i, \text{ with } \mathfrak{b}_i := b_i + \beta_i,$$

which clarifies that solving this model is analogous to solving the model in the main part of the paper by substituting $\mathfrak{b}_i = b_i + \beta_i$ for b_i. Specifically, the optimal tax described in Proposition 2 with utilitaristic social preferences becomes

$$t'(x) = E[b_i | v_i + b_i] + E[\beta_i | v_i + b_i],$$

where the first and second summand capture internalities and externalities, respectively.

Importantly, a report $\hat{v}_i = v_i + b_i$ contains no specific information about β_i, which strongly reduces the information content that can be exploited to correct consumers. While the term $E[b_i | v_i + b_i]$ induces internality revelation and a distinct pattern between reports and biases that we investigate in Corollary 3, for example, the term $E[\beta_i | v_i + b_i]$ shows no such regularities and varies arbitrarily by context. In particular, in the case of global externalities, the second term reduces to a constant $\beta_i = \beta$ that is added to the optimal marginal tax rate $t'(x)$. As the purpose of our paper is to inves-

tigate how non-linear taxation allows to target consumers, this case is only of limited interest to us.

C Internality Revelation: Bivariate Normal Case

C.A Density $p(\hat{v})$ of the Perceived Valuation

Let (v,b) be jointly normal distributed with:

$$(v, b) \sim N\left(\left[\begin{array}{c} \mu_v \\ \mu_b \end{array}\right], \left[\begin{array}{cc} \sigma_v^2 & \rho\sigma_v\sigma_b \\ \rho\sigma_v\sigma_b & \sigma_b^2 \end{array}\right]\right). \tag{14}$$

The distribution of the sum of v and b is given by

$$\hat{v} \sim N\left(\mu_v + \mu_b, \sigma_v^2 + \sigma_b^2 + 2\rho\sigma_v\sigma_b\right) =: \left(\mu_{\hat{v}}, \sigma_{\hat{v}}^2\right). \tag{15}$$

C.B Conditional Expectation $E[b|\hat{v}]$ of the Bias

It can be shown that

$$E\left[b|\hat{v}\right] = \frac{cov(b,\hat{v})}{\sigma_{\hat{v}}^2} \cdot \hat{v} + \left[\mu_b - \frac{cov(b,\hat{v})}{\sigma_{\hat{v}}^2} \cdot \mu_{\hat{v}}\right] = \mu_b + \frac{cov(b,\hat{v})}{\sigma_{\hat{v}}^2} \cdot \left[\hat{v} - \mu_{\hat{v}}\right]. \tag{16}$$

with $cov(b,\hat{v}) = cov(b,v) + \sigma_b^2$. Rearranging gives that

$$E\left[b|\hat{v}\right] = \mu_b + \frac{\rho + (\sigma_b/\sigma_v)}{(\sigma_b/\sigma_v) + (\sigma_v/\sigma_b) + 2\rho} \cdot \left[\hat{v} - \mu_{\hat{v}}\right]. \tag{17}$$

This implies that

$$\frac{\partial E\left[b|\hat{v}\right]}{\partial \hat{v}} > 0 \Leftrightarrow cov(b,\hat{v}) > 0 \Leftrightarrow cov(b,v) > -\sigma_b^2 \Leftrightarrow \rho > -\frac{\sigma_b}{\sigma_v}. \tag{18}$$

Analogously, it can be shown that

$$\frac{\partial E\left[v|\hat{v}\right]}{\partial \hat{v}} > 0 \Leftrightarrow \rho > -\frac{\sigma_v}{\sigma_b}. \tag{19}$$

D Optimal tax in direct mechanism

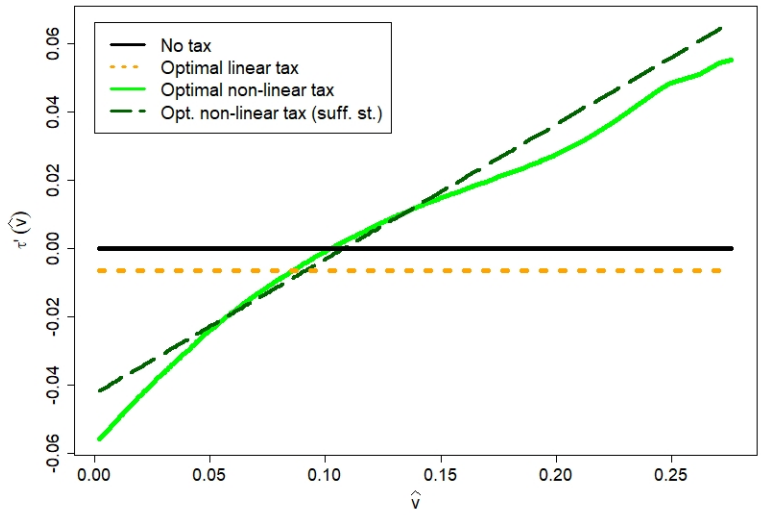


Figure 5: Optimal tax rates as a function of $\hat{v}$.

References

Abaluck, Jason and Jonathan Gruber, Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program, *American Economic Review*, 2011, *101*, 1180–1210.

Akerlof, George A., The Economics of "Tagging" as Applied to the Optimal Income Tax, Welfare Programs, and Manpower Planning, *American Economic Review*, 1978, *68*, 8–19.

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

Allcott, Hunt, Benjamin B. Lockwood, and Dmitry Taubinsky, Regressive Sin Taxes, with an Application to the Optimal Soda Tax, 2018. NBER Working Paper No. 25841.

Allcott, Hunt, Christopher Knittel, and Dmitry Taubinsky, Tagging and Targeting of Energy Efficiency Subsidies, *American Economic Review*, 2015, *105*, 187–191.

Allcott, Hunt, Sendhil Mullainathan, and Dmitry Taubinsky, Energy Policy with Externalites and Internalities, *Journal of Public Economics*, 2014, pp. 72–88.

Angrist, Joshua David and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton, NJ: Princeton Univ. Press, 2009.

Attari, Shahzeen Z., Michael L. DeKay, Cliff I. Davidson, and Wändi Bruine de Bruin, Public Perceptions of Energy Consumption and Savings, *Proceedings of the National Academy of Sciences of the United States of America*, 2010, *107*, 16054–16059.

Basov, Suren, *Multidimensional Screening*, New York: Springer, 2005.

Bernheim, B. Douglas and Dmitry Taubinsky, Behavioral Public Economics, in B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, eds., *Handbook of Behavioral Economics - Foundations and Applications*, Vol. I, Elsevier, 2018, pp. 381–516.

Bollinger, Bryan, Phillip Leslie, and Alan Sorensen, Calorie Posting in Chain Restaurants, *American Economic Journal: Economic Policy*, 2011, *3*, 91–128.

Cremer, Helmuth and Firouz Gahvari, On Optimal Taxation of Housing, *Journal of Urban Economics*, 1998, *43*, 315–335.

Diamond, Peter A., Optimal Income Taxation: An Example with a U-Shaped Pattern of Optimal Marginal Tax Rates, *American Economic Review*, 1998, *88*, 83–95.

Farhi, Emmanuel and Xavier Gabaix, Optimal Taxation with Behavioral Agents, 2017. NBER Working Paper No. 21524.

Gerritsen, Aart, Optimal Taxation When People Do Not Maximize Well-being, *Journal of Public Economics*, 2016, *144*, 122–139.

Gibbard, Allan, Manipulation of Voting Schemes: A General Result, *Econometrica*, 1971, *41*, 587–601.

Goldfarb, Avi and Catherine Tucker, Digitial Economics, *Journal of Economic Literature*, 2019, *57* (1), 3–43.

Griffith, R., M. O'Connell, and K. Smith, Tax Design in the Alcohol Market, *Journal of Public Economics*, 2019, *172*, 308–343.

Handel, Benjamin and Joshua Schwartzstein, Frictions or Mental Gaps: What's Behind the Information We (Don't) Use and When Do We Care?, *Journal of Economic Perspectives*, 2018, *32* (1), 155–78.

Jensen, Robert, The (Perceived) Returns to Education and the Demand for Schooling, *Quarterly Journal of Economics*, 2010, *125*, 515–548.

Kahneman, Daniel, Peter P. Wakker, and Rakesh Sarin, Back to Bentham? Explorations of Experienced Utility, *Quarterly Journal of Economics*, 1997, *112*, 375–405.

Kruger, Justin and David Dunning, Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments, *Journal of Personality and Social Psychology*, 1999, *77*, 1121–1134.

Laffont, Jean-Jacques, Eric Maskin, and Jean-Charles Rochet, Optimal Nonlinear Pricing with Two-Dimensional Characteristics, *Information, Incentives and Economic Mechanisms*, 1987, pp. 256–266.

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

Lockwood, Benjamin B. and Dmitry Taubinsky, Optimal Income Taxation with Present Bias, 2016. Working Paper.

Loewenstein, George and Drazen Prelec, Anomalies in Intertemporal Choice: Evidence and an Interpretation, *Quarterly Journal of Economics*, 1992, *107*, 573–597.

McAfee, R Preston and John McMillan, Multidimensional Incentive Compatibility and Mechanism Design, *Journal of Economic Theory*, 1988, *46* (2), 335–354.

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

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

Mussa, Michael and Sherwin Rosen, Monopoly and Product Quality, *Journal of Economic Theory*, 1978, *18*, 301–317.

Myerson, Roger B., Optimal auction design, *Mathematics of Operation Research*, 1981, *6*, 58–73.

O'Donoghue, Ted and Matthew Rabin, Doing It Now or Later, *American Economic Review*, 1999, *89*, 103–124.

O'Donoghue, Ted and Matthew Rabin, Optimal Sin Taxes, *Journal of Public Economics*, 2006, *90*, 1825–1849.

Royden, Halsey and Patrick M. Fitzpatrick, *Real Analysis*, Harlow: Pearson Education, 2010.

Saez, Emmanuel and Stefanie Stantcheva, Generalized Social Marginal Welfare Weights for Optimal Tax Theory, *American Economic Review*, 2016, *106*, 24–45.

Thaler, Richard H. and Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, New Haven: Yale University Press, 2008.

Wilson, Robert B., *Nonlinear Pricing*, Oxford: Oxford University Press, 1997.